Contribution à la modélisation numérique du transport de polluants en rivière

Laure Simon

Mémoire présenté pour l'obtention du titre de
Docteur de l'Ecole Nationale des Ponts et Chaussées

Spécialité Sciences et Techniques de l'Environnement

Contribution à la modélisation numérique du transport de polluants en rivière.

par

Laure SIMON

Thèse soutenue le 13 Janvier 1995

Jury

M. Bruce BECK	Directeur de Thèse
M. Kim DAN N'GUYEN	Rapporteur
M. László SOMLYÓDY	Rapporteur
M. Jean CUNGE	Examinateur
M. Ghislain de MARSILY	Examinateur
M. Rémy POCHAT	Examinateur

Remerciements

A l'issue de ce travail, je voudrais remercier tous ceux qui, directement ou indirectement, ont participé à son élaboration.

Ma reconnaissance va tout d'abord à M. Rémy POCHAT, directeur du département Équipements pour l'Eau et l'Environnement au CEMAGREF. Son accueil, quand il était directeur du CERGRENE où j'effectuais un DEA, a pesé lourd dans ma décision de "commettre" une thèse. Merci pour les conseils et les encouragements prodigués, d'avoir accepté le rôle ingrat de correcteur "en temps réel" et enfin de participer à ce jury !

Le professeur B. BECK a bien voulu, d'Angleterre ou des USA, accepter d'assurer pendant de longues années le pilotage de cette thèse, même si je l'entraînais quelquefois sur des sujets éloignés de ses domaines de prédilection. Qu'il trouve ici l'expression de ma gratitude.

Toujours souriant et disponible, ouvert aux questions, aux suggestions comme à la critique, le professeur Dan N'GUYEN est certainement celui qui a le plus fait pour m'initier aux arcanes de la modélisation numérique, et m'aider à m'y retrouver. Je lui suis extrêmement reconnaissante d'avoir, de plus, été rapporteur de ce travail.

Mes plus sincères remerciements vont également au professeur László SOMLYÓDY qui, malgré ses doubles responsabilités à l'IIASA comme à l'université de Budapest, a bien voulu accepter de décrypter ce "pavé" expédié de France et, qui plus est, de rédiger un rapport à son sujet.

M. Jean CUNGE, directeur scientifique du LHF, est avec sa vaste expérience d'hydraulicien, des développements aux applications les plus variées, une référence pour les apprentis modélisateurs. Je suis donc particulièrement honorée qu'il ait bien voulu examiner cette thèse.

L'application sur la Seine n'aurait pas été possible en dehors du cadre du PIREN-SEINE, dont l'infatigable directeur, le professeur G. de MARSILY, me fait l'honneur de participer à ce jury. Qu'il en soit ici grandement remercié, ainsi que les divers membres et financeurs du PIREN-SEINE. Je souhaiterais tout particulièrement exprimer ma reconnaissance à l'égard :

- du Service de Navigation de la Seine, qui a gracieusement mis à notre disposition l'ensemble de ses données bathymétriques;
- de l'équipe de MM. Grange et Rollin au LROP, sans l'assistance de laquelle nous n'aurions pu réaliser les campagnes de traçage (nb : j'espère qu'ils gardent un aussi bon souvenir que nous du barbecue organisé lors du premier traçage !);
- de S. Even et M. Poulin (CIG - Ecole des Mines de Paris), pour m'avoir permis, grâce à

la mise à disposition du modèle PROSE, d'analyser le bief étudié avant d'y appliquer un outil bidimensionnel, ... ainsi que pour d'autres collaborations fructueuses sur la Seine, en dehors de la thèse;

- de mes camarades du groupe "Orages", et d'autres courageux volontaires du CERGRENE, qui se sont retrouvés sur les bords de la Seine à des heures indues pour prélever une eau d'une couleur plus que douteuse.

Au cours de ces années, j'ai eu également le plaisir de travailler avec plusieurs étudiants, soit sur certains problèmes de modélisation soit pour le recueil de données (bathymétriques notamment). Je voudrais ainsi remercier C. Costel, D. François, H. Lin, E. Lucas-Aiguier, E. Tellier et L. Théry.

Enfin, ils se préoccupent fort peu de recherche, d'hydraulique en général et de la Seine en particulier mais ils ont su être là aux moments où l'on désespère un peu de voir le bout de la route : merci à ce petit groupe d'amis et parents dont le soutien m'a été si essentiel.

A ma mère.
et en mémoire de mon père.

Résumé

Pour traiter des problèmes d'environnement se posant dans les cours d'eau, il est en général indispensable d'appréhender la dynamique des écoulements, à un degré divers selon les applications envisagées. La modélisation mathématique est un des outils les plus appropriés pour ce faire.

Le présent travail retrace le développement d'un modèle bidimensionnel plan d'écoulement et de transport des composés dissous, destiné à permettre des analyses locales mais détaillées du devenir de rejets polluants.

Avant de se lancer dans son élaboration, on a cherché à acquérir une vision globale des phénomènes physiques en jeu et de leur formalisation usuelle ainsi que de l'état de l'art en matière d'analyse numérique appliquée à l'hydraulique à surface libre. Quelques options-clefs étant choisies (recours aux différences finies, utilisation d'une approche à pas fractionnaire), on s'est intéressé plus particulièrement :

- au traitement des termes advectifs quand ceux-ci sont prépondérants par rapport aux phénomènes dispersifs, ce qui est souvent le cas lors des premiers stades de la dilution d'un polluant;
- à la résolution des termes de propagation des ondes et frottement sur le fond et les berges, cruciaux en hydraulique fluviale.

La méthodologie utilisée repose sur le recours systématique à des cas-tests documentés qui permettent un classement objectif des différents algorithmes en fonction tout à la fois de leur précision, leur robustesse et leur coût informatique.

L'applicabilité du modèle proposé à des problèmes grandeur nature est illustrée par l'interprétation d'un traçage en Seine. Cet exemple démontre le potentiel prédictif de ce type d'outil vis à vis de formulations plus simples mais plus éloignées de la physique des phénomènes.

Abstract

When dealing with environmental problems in natural water bodies, some knowledge of the flow dynamics is usually required, more or less precise according to the problem at hand. Mathematical models are useful tools for providing this insight.

The present work deals with the development of a two-dimensional, depth-averaged, model of surface flow and transport, designed to perform local but detailed analysis of pollutants dilution.

Prior to undertaking its elaboration, we tried to achieve a global understanding of the various physical phenomena controlling flow and transport, of their usual representations and of the scope of techniques available in the field of computational hydraulics. Key options were chosen based on this review (choice of a finite differences method and of a fractional step approach). Then, we have been focussing :

- on the computation of advective terms when advection is predominant over dispersion, which occurs often in the first stages of pollutant dilution;
- on the solution of propagation and friction terms, particularly important in fluvial hydraulics.

The followed methodology consists of the systematic application of benchmark tests which allow an objective ranking of different numerical algorithms, considering at the same time their accuracy, robustness and computational efficiency.

The ability of the proposed model to deal with full scale situations is illustrated by the interpretation of a dye-tracing experiment in the Seine River. This example demonstrates the power of prediction of this kind of tool with respect to simpler models less faithful to the physics.

Mots clés

Hydraulique fluviale
Transfert de polluants
Analyse numérique
rivière Seine

Key words

Fluvial Hydraulics
Pollutant transport
Numerical Analysis
Seine River

Contents

List of Figures

APPENDIX

List of Tables

Chapter 1

Outline of the dissertation

When dealing with environmental problems concerning water bodies, some knowledge of the flow dynamics is generally needed. Indeed, transport by the flow is the first process which governs the fate of chemical species, whether they are dissolved or linked to solid (suspended) matters. That it is the first process does not mean necessarily it's the most important : that depends on the features (notably the typical time and space scales) of the studied phenomenon. For instance, an accurate description of transport and dilution is obviously of overwhelming importance when studying the short-term, near-field fate of sudden pollutant inputs. It is much less relevant when tackling such aspects as seasonal algal development within a river system : there, an assessment of flow and mass inputs through the river tributaries and catchments and of their downstream displacement is required, but most often on a daily basis.

Consequently, in order to fit these different kinds of applications, a lot of hydraulic models have been developed, from the one-dimensional Hayami or Muskingum-Cunge simplified models to fully three-dimensional models eventually including detailed turbulence modelling. Similarly, dissolved species transport has been described by models ranging from the plug-flow or cells in series models to three-dimensional turbulent models, without forgetting to mention models relying on the one- or multi-dimensional advection-dispersion equation.

The present work concerns the development of a two-dimensional mechanistic depth-averaged surface flow and transport model. This tool can be said to belong to the category of most sophisticated flow and transport models, although it is far from being the more complete and complex member of this group.
This model development has been undertaken in the frame of the Piren-Seine research program. The Seine is a relatively small river heavily impacted by the intense anthropogenic activitities in its catchment (notably it flows through Paris and its region, which represents roughly 10 million inhabitants). The Piren-Seine investigates the functioning of the river ecosystem and its

reaction to inputs from its rural and urban catchments, both in the short term (e.g. impact of combined sewer overflows) and in the long term (ex : eutrophication, sensitivity to changes in sewage treatment). It involves notably the building of a chain of models. The two-dimensional one should be applied to perform local (i.e. on reaches a few kilometers long) detailed analysis of pollutant inputs dilution and assess the errors linked to simplifications used in one-dimensional models (which should ultimately be applied as planning and management tools).

Given the planned applications, the model should estimate accurately flow and transport. This explains why our work deals mainly with the quest for reliable numerical algorithms.

The following report is divided into three parts. Part I just sets the background for our work, while parts II and III report our contribution :

- Chapter 2 reviews the different phenomena (notably turbulence and dispersion) which govern flow patterns and the usual formulations applied to model them. It explains the derivation of the equations used in our model, namely the shallow water St-Venant equations and the accompanying depth-averaged advection-dispersion equation.
- Chapter 3 provides a panorama of the great families of numerical methods available to solve flow and transport equations. It exposes briefly their principles, while introducing some definitions and concepts essential in numerical analysis. It also mentions approaches currently applied to reduce the problem complexity (e.g. splitting) as well as possible sources of troubles (e.g. system linearization, treatment of boundary conditions ...).

Most numerical approaches to the solution of the flow and transport equations are fractional step methods. This means that the equation to be solved is split into several pieces, each one corresponding to a specific differential operator. These operators are then dealt with successively, with algorithms best suited to their mathematical properties. The "first generation" of fractional step methods was merely using directional splitting (e.g. separating differentiation along the different coordinate directions) for achieving computational efficiency. The most clever methods now apply "process splitting" : this means that an equation is splitted according to the different physical phenomena it involves. We shall see that in two-dimensional situations, transport results from the combination of advection and dispersion processes while in the flow equations (conservation of momentum), we find advection, dispersion, propagation and friction processes (cf chapter 2). Solving advection and dispersion is thus the part common to flow and transport models. This is the reason why we dealt with it first, throughout part II :

- Computing advection accurately has long been known as one of the most challenging tasks in the field of computational fluid mecanics. It is particularly critical when solving the transport equation of a dissolved compound (indeed, in the flow equations, advection can be balanced, or even outweighted, by other physical forces).

The literature review summarized in chapter 4 guides us in the selection of a handful of schemes for further testing.

- Recommended test cases, dealing with one-dimensional (chapter 6) and two-dimensional (chapter 7) situations, have been systematically applied to the chosen schemes. They allow us to grade objectively the algorithms according to their accuracy, robustness and computational efficiency and lead us to define a strategy for solving advection in our forthcoming applications (summarized in section 7.5).

Part III deals with the development of the two-dimensional St-Venant flow model :

- In chapter 8, we propose different models for the flow computation. These models have several features in common. They all rely on process splitting and use similar algorithms to compute advection and dispersion terms. Their approach to the solution of propagation and friction terms is the same : it consists first of uncoupling the different flow equations after linearizing each one and performing appropriate substitutions, secondly of factorizing the resulting two-dimensional differential operators into the product of one-dimensional ones. However, the models differ in the choice of dependent variables (velocity components or unit-width discharges) and in the amount of simplification applied to the flow equations.

- As in the classification of advection algorithms the ranking of these flow models is based on hypothetical test cases. These concern both steady-state (chapter 9) and unsteady (chapter 10) flows. Model forecasts can be compared, either to a reference solution or measurements, either to some qualitative information about the plausible flow patterns.

- Part III concludes with an application of the most adequate model (cf section 10.4) to a full scale situation, involving the interpretation of a dye-tracing experiment on the Seine river (chapter 11).

Lastly, the experience collected throughout the development of these models is summed up, together with the further research issues it suggests, within a general conclusion, which forms final chapter 12.

Quand on s'intéresse aux problèmes d'environnement se posant dans les cours d'eau ou les plans d'eau, il est en général indispensable de posséder quelque information quant à la dynamique des écoulements. En effet, transport et dilution par l'écoulement sont les premiers phénomènes controlant le devenir dans le milieu naturel de substances chimiques dissoutes ou en suspension ...sans être nécessairement les phénomènes majeurs. En fait, le degré de détail avec lequel l'hydraulique doit être appréhendée dépend du problème considéré, notamment de ses échelles de temps et d'espace. Ainsi tandis que pour évaluer correctement le devenir à court terme de rejets ponctuels une description précise de l'hydraulique est requise, il sera généralement suffisant, pour étudier des problèmes d'eutrophisation à l'échelle saisonnière, de disposer d'une évaluation quotidienne des débits de la rivière et de ses affluents.

Afin de répondre de façon adéquate aux besoins liés à différentes applications, une large gamme de modèles hydrauliques a été développée, allant des modèles monodimensionnels simplifiés de type Hayami ou Muskingum-Cunge aux modèles tridimensionnels incluant une représentation détaillée de la turbulence. Parallélement, de nombreux modèles de transport dissous ont vu le jour, du type "réservoirs parfaitement mélangés en série" (cells-in-series models) inspirés des approches utilisées en génie chimique aux modèles tridimensionnels encore une fois. Notons néammoins que la plupart de ces modèles, quelque soit leur dimension, sont basés sur l'équation d'advection-dispersion.

Le travail présenté ci-dessous se rapporte au développement d'un modèle déterministe bidimensionnel plan d'ecoulement et transport. On peut considérer que cet outil relève de la catégorie des modèles hydrauliques et de transport les plus sophistiqués, même s'il est loin d'en constituer le membre le plus complet ni le plus complexe.

Ce développement a été entrepris dans le cadre du programme de recherches CNRS PIREN-SEINE. La Seine est une rivière de taille plutôt modeste en regard des diverses perturbations d'origine anthropique qu'elle subit, notamment lors de sa traversée de l'agglomération parisienne. Le PIREN-SEINE se consacre à l'étude du fonctionnement de divers compartiments de l'hydrosystème Seine, aux influences sur sa qualité des apports des bassins versants urbains comme ruraux, à court (impact des déversements d'orage) ou long terme (eutrophisation, accumulation des métaux lourds). Le PIREN-SEINE a notamment pour objectif le développement d'une "chaîne" de modèles. L'outil bidimensionnel devrait être appliqué pour réaliser des analyses détaillées (sur des biefs de quelques kilomètres) des modalités de dilution de certains rejets, et pour estimer les erreurs liées aux simplifications inhérentes aux modèles mono-dimensionnels, lesquels devraient, à terme, évoluer du stade d'outils de compréhension et recherche à celui d'outils de prévision et gestion.

Compte tenu des applications envisagées notre modèle devait être capable de simuler de façon précise hydraulique et transport. Ceci explique pourquoi cette thèse a été consacrée pour

l'essentiel à la recherche d'algorithmes numériques fiables.

Le présent rapport est divisé en trois parties. La partie I sert essentiellement à dresser le cadre dans lequel ce travail s'inscrit tandis que les parties II et III présentent notre contribution :

- Dans le chapitre 2 on introduit les différents phénomènes (notamment turbulence et dispersion) qui gouvernent hydraulique et transport, ainsi que leurs modélisations usuelles. On aboutit ainsi aux équations bidimensionnelles de St-Venant et d'advection-dispersion qui sont les fondements physiques de notre modèle.
- Au chapitre 3 on présente un panorama des "grandes familles" de méthodes numériques disponibles pour résoudre les équations de l'hydraulique. On expose brièvement leurs principes, tout en introduisant quelques concepts et définitions incontournables en analyse numérique. On signale également les approches couramment appliquées pour réduire la complexité des problèmes (e.g. l'éclatement des opérateurs) ainsi que certaines sources potentielles d'erreurs, voire pire (mode de linéarisation des équations, traitement des conditions aux limites, etc ...).

La plupart des méthodes appliquées au calcul de l'hydraulique et du transport sont des méthodes "à pas fractionnaires". Ceci signifie que la (les) équation(s) à résoudre est décomposée (éclatée) en différents morceaux, chacun correspondant à un opérateur différentiel spécifique. Chaque opérateur est ensuite traité séparément et successivement, par la méthode la plus appropriée à ses caractéristiques mathématiques. La "première génération" des méthodes à pas fractionnaires reposait essentiellement sur une décomposition selon les différentes directions de l'espace, ceci afin de rendre les algorithmes plus efficaces en se ramenant à la résolution de systèmes monodimensionnels plutôt que multidimensionnels. Les méthodes les plus poussées de nos jours combine à l'éclatement directionnel une séparation en fonction du phénomène physique que représente chaque opérateur. Nous verrons par la suite (chapitre 2) que, dans un cadre bidimensionel plan (i.e. grandeurs moyennées suivant la verticale), le transport résulte de la combinaison de phénomènes d'advection et de dispersion tandis qu'on voit apparaître en outre dans les équations hydrauliques des termes de propagation et friction. Calculer advection et dispersion est donc une étape commune aux modèles de transport et d'écoulement. C'est pourquoi nous avons traité ce problème en premier lieu, dans la partie II :

- Prendre en compte précisément les termes advectifs, voilà qui constitue depuis longtemps une des questions les plus délicates dans le domaine de la mécanique des fluides. Ce problème est particulièrement crucial au niveau des équations de transport (en effet, dans les équations de l'écoulement, l'advection peut être contrebalancée, voire dominée, par d'autres forces).

Le chapitre 4 est une revue bibliographique de différentes méthodes proposées pour la résolution de l'advection. C'est sur sa base que nous avons a priori retenu un groupe de schémas afin de les tester de façon approfondie.

- Les schémas sélectionnés ont été appliqués à la résolution de cas-tests de nature mono- (chapitre 6) ou bidimensionnelle (chapitre 7). Ces tests permettent d'établir un classement objectif des algorithmes, en fonction tout à la fois de leur précision, leur robustesse et leur coût en terme de temps calcul. Ils nous conduisent à definir une stratégie de résolution des termes advectifs dans le cadre de nos futures applications (cf section 7.5).

La partie III est consacrée au développement du modèle hydraulique basé sur les équations bidimensionnelles de St-Venant :

- Au chapitre 8, nous proposons différents modèles d'ecoulement. Ces modèles ont plusieurs importantes caractéristiques en commun :
  - Ils sont tous basés sur l'éclatement des opérateurs en fonction des phénomènes physiques représentés.
  - Ils utilisent les mêmes algorithmes pour traiter advection et dispersion.
  - L'approche adoptée pour la résolution des termes de propagation et friction est similaire : elle consiste en linéariser chaque équation puis les découpler en pratiquant des substitutions de variables appropriées. On se ramène ainsi à une seule équation bidimensionnelle dont l'inconnue est la cote de la surface libre, qui est, pour sa résolution, factorisée en opérateurs monodimensionnels.

  Cependant ces modèles diffèrent dans le choix des variables de travail (composantes de la vitesse ou des débits par unité de largeur) et dans le degré de simplification apporté aux équations de l'hydraulique.

- L'évaluation de ces modèles est basée, comme pour les algorithmes de transport, sur le traitement de cas-tests, concernant aussi bien des écoulements stationnaires (chapitre 9) qu'instationnaires (chapitre 10). Les simulations sont comparées, soit à des solutions analytiques de référence ou des mesures, soit à des informations qualitatives quant aux caractéristiques plausibles des champs de vitesses et profils de surface libre.

- La partie III se conclut par une application "grandeur nature" du modèle le plus adéquat (cf section 10.4) à l'interprétation d'un traçage sur la Seine (chapitre 11).

En conclusion, l'expérience acquise lors du développent de ces divers modèles nous a inspiré quelques commentaires et suggéré quelques pistes de recherche. Le tout fait l'objet du chapitre 12.

Part I

Introduction to the modelling of surface water flow and transport phenomena

Chapter 2

Surface water flow and transport phenomena

For the past thirty years, civil engineers and researchers have been developing and using computer models of surface water flow and transport. These have been applied to a variety of problems, which may be roughly classified into the following sub-classes :

- **Engineering hydraulics** concern flows in man-made geometries, over, near or around structures. Examples of these problems are flows around jetties, in diversions or settling tanks, near-field calculation of heat or sewage effluent plumes from power or wastewater treatment plant, etc ...

- **Environmental hydraulics** deal with flow and transport in naturally existing water bodies. Examples are flows in rivers, estuaries, coastal regions, lakes or reservoirs.

There exist some obvious differences between the two groups of applications. First, environmental hydraulics problems generally concern much larger water bodies than in engineering hydraulics applications. The time scale of the transport phenomena and of their relevant fluctuations may also be significantly different. Secondly, in environmental applications, the transport process represents only one of the factors which govern the fate of chemical species. Thus, the level of precision achieved in describing and computing the transport process must be consistent with the detail with which biogeochemical processes are accounted for. Lastly, in environmental applications, we may often expect that some information will be available with more uncertainty than in engineering problems (e.g. the bathymetry of an estuary is generally less known, more difficult to monitor than the exact shape of a settling tank).
Inside each group of problems, there also exists a great variability among time scales, space

scales, relative importance of different forcing variables. Thus, some flows truly require a three-dimensional description while others are satisfactorily characterized by depth-averaged, width-averaged or even cross-section averaged variables.
Considering these points, it is no surprise that so many models, more or less tailored to specific applications, have sprung from basically the same equations.

This chapter covers the following topics :

1. First, we introduce briefly both the equations and fundamental phenomena which govern surface flow and transport (section 2.1).

2. Secondly, we give an overview of the different methods currently in use to model these basic phenomena and reduce the flow and transport equations to a workable form (sections 2.2 and 2.3). This review is limited to the approaches most frequent in surface flow problems.

3. Lastly, we discuss the relevance of different levels of modelling with respect to such problems as computational efficiency, data availability and intended applications (section 2.4). As an example, we introduce the case of the Seine river, which should be the first full-scale application of the modelling tools developed in this work.
   Considering the conditions which generally prevail when undertaking hydraulic modelling, it is no wonder that the two-dimensional St-Venant shallow water equations and the companion advection-dispersion equation for dissolved compounds are widely used (section 2.4.5). They will notably be applied for the Seine River.

2.1 From Navier-Stokes to depth-averaged equations

2.1.1 Navier-Stokes equations

The only mathematical model which adequately describes the flow under all possible situations is made of the exact Navier-Stokes equations (which basically express the conservation of mass, momentum, thermal energy and/or species concentration). These well-known equations read, for incompressible flows (nb: we use tensor notation for the sake of simplicity) :

- *mass conservation (continuity equation)*

$$\frac{\partial \tilde{U}_i}{\partial x_i} = 0 \tag{2.1}$$

- *momentum conservation*

$$\frac{\partial \tilde{U}_i}{\partial t} + \tilde{U}_j \frac{\partial \tilde{U}_i}{\partial x_j} = \frac{-1}{\rho_r} \frac{\partial \tilde{P}}{\partial x_i} + \nu \frac{\partial^2 \tilde{U}_i}{\partial x_j^2} + g_i \frac{\rho - \rho_r}{\rho_r} \quad i = 1, 2, 3 \tag{2.2}$$

- *thermal energy/species concentration conservation*

$$\frac{\partial \tilde{\phi}}{\partial t} + \tilde{U}_j \frac{\partial \tilde{\phi}}{\partial x_j} = \lambda \frac{\partial^2 \tilde{\phi}}{\partial x_j^2} + S_\phi \tag{2.3}$$

where $\tilde{U}_i$ denotes the instantaneous velocity component in the coordinate direction x_i, $\tilde{P}$ is the static pressure and $\tilde{\phi}$ a scalar quantity (either temperature or concentration). ν and λ are respectively the kinematic molecular viscosity and the molecular diffusivity. S_ϕ is a volumetric source or sink term related for instance to biogeochemical reactions.

The last term of the momentum equations is a buoyancy term accounting for the effect of variable density according to the Boussinesq approximation. It involves a reference density ρ_r and the component of the gravitational acceleration g in the direction x_i, namely g_i .

Together with an adequate equation of state for the variable density ρ, these equations form a closed set. The buoyancy term may introduce some coupling between the momentum (eq. 2.2) and concentration (eq. 2.3) equations. Indeed, the density ρ is affected by the temperature and eventually by some species (e.g. salt) concentration. However, this term is relevant only in the case of strong salt concentration or temperature gradients (nb : the relative difference between density at 4° C and 25° C is only 0.25 % (Tassin, 1986)).

2.1.2 Three-dimensional Reynolds equations

In hydraulics, as in other areas of fluid mechanics, the flows of practical relevance are almost always turbulent. As stated in (Hug, 1975; Rodi, 1980; ASCE Task Committee on Turbulence

Models in Hydraulic Computations, 1988), turbulence is an eddying, unsteady, three dimensional motion with a wide spectrum of eddy sizes and a corresponding spectrum of fluctuation frequencies. Due to this eddying motion which agitates the fluid, flow features (velocity, pressure, ...) are in fact varying quickly, in all directions.

The largest eddies, associated with the low frequencies, are of the size of the flow domain, while the smallest eddies, associated with the high frequencies, are typically several orders of magnitude smaller. Describing these small scale motions would require an extremely fine spatial and temporal discretization. Thus, despite the continuous advances in computers, solving the complete Navier-Stokes equations for most relevant applications is still out of the question. Besides, from a practical point of view, we are generally not interested in all the details of turbulent motion.

Therefore, a statistical approach, as suggested first by Reynolds, is adopted. Mean quantities are defined by :

$$F = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \tilde{F} dt \tag{2.4}$$

where $\tilde{F}$ stands for the instantaneous value of either velocity components, pressure, temperature or species concentration. The averaging time $t_2 - t_1$ is long compared to the time scale of turbulent motion but should be small with respect to that of the mean flow in transient problems. The instantaneous value $\tilde{F}$ is considered as the sum of the mean value F and a fluctuating component f .

Neglecting viscous stresses (in eq. 2.2) and molecular heat or mass flux (in eq. 2.3), the process of time averaging according to 2.4 yields the following equations :

$$\frac{\partial U_i}{\partial x_i} = 0 \tag{2.5}$$

$$\frac{\partial U_i}{\partial t} + U_j \frac{\partial U_i}{\partial x_j} = \frac{-1}{\rho} \frac{\partial P}{\partial x_i} - \frac{\partial \overline{u_i u_j}}{\partial x_j} + g_i \frac{\rho - \rho_r}{\rho_r} \tag{2.6}$$

$$\frac{\partial \phi}{\partial t} + U_j \frac{\partial \phi}{\partial x_j} = - \frac{\partial \overline{\varphi u_j}}{\partial x_j} \tag{2.7}$$

Time averaging has suitably removed the fluctuations from the governing equations. However, since the original Navier-Stokes and scalar transport equations are not linear (due to advective terms), it has introduced instead unknown correlations, terms such as $\overline{u_i u_j}$ and $\overline{\varphi u_j}$ (the time-averaged value of cross-products between fluctuating velocities or between velocities and scalar fluctuations).

Physically, these terms, when multiplied by the density ρ, represent the rate of transport of momentum, heat or mass, due to the fluctuating turbulent motion. The velocity correlations $-\rho \overline{u_i u_j}$ appear to act as stresses on the fluid and are therefore called turbulent or Reynold

stresses. From now on we shall use notations

$$\tau_{ij} = -\rho\,\overline{u_i u_j} \quad \text{and} \quad J_j = -\rho\,\overline{\varphi u_j}$$

or, considering that directions x_1, x_2, x_3 correspond respectively to $x-$, $y-$ and $z-$axis,

$$\tau_{xy} = -\rho\,\overline{u_1 u_2},\ J_x = -\rho\,\overline{\varphi u_1},\ \text{and so on ...}$$

Viscous stresses and molecular fluxes were discarded because they are generally several orders of magnitude smaller than their turbulent counterparts except in the viscous sublayer near walls. In addition, in eq. 2.6 we did not consider explicitly such forcing effects as Coriolis forces; this is for two reasons. First, Coriolis effects are perceptible only in very large water bodies. For instance, they are barely noticeable in a river reach a few kilometers long. Secondly, the Coriolis force is expressed as a *linear* function of the velocities, so that its temporal or depth-averaging does not introduce any troublesome unknown term.

Equations 2.5 to 2.7 can be solved for the mean flow variables U_i, P_i and ϕ only if terms such as $\overline{u_i u_j}$ and $\overline{\varphi u_j}$ can be determined in some way. This is basically the task of turbulence-closure models we shall introduce briefly in next section.

It is worth noting that the right-hand side of momentum equations 2.6 include other terms than the turbulent ones. In some flows, for instance in certain large and shallow water bodies, the inertial terms on the left-hand side of 2.6 are chiefly balanced by the pressure gradient and/or buoyancy terms, so that a detailed description and simulation of the turbulent stresses is useless. In contrast, the turbulent fluxes in eq. 2.7 are the only ones which balance the advective ones and should therefore always been accounted for.

2.1.3 Depth-averaged equations

In many free-surface, shallow water situations, the mean-flow characteristics vary but little in the vertical direction. Therefore, it is sufficient to describe their horizontal distribution. (A shallow water body is such that its characteristic scale in the vertical direction $\widehat{H}$ is much smaller than its horizontal scale $\widehat{L}$. Typically, ratio $\widehat{L}/\widehat{H}$ is larger than 15).

Provided that the vertical acceleration of a fluid particle and vertical viscous and turbulent stresses are negligible with respect to the gravity acceleration, the momentum equation in the vertical direction can be replaced by :

$$\frac{\partial P}{\partial z} = -\rho\, g \tag{2.8}$$

This amounts to assuming that *the pressure distribution is hydrostatic within the fluid.* For instance, according to (Mary, 1982), this hypothesis is fair enough far from the banks of a river

which bathymetry is smoothly varying, when current lines have but a moderate curvature and free surface is nearly horizontal.

Using relation 2.8, depth-averaged equations, obtained by integrating eq. 2.5 to 2.7 over the water column, read :

$$\frac{\partial h}{\partial t} + \frac{\partial h\overline{U}}{\partial x} + \frac{\partial h\overline{V}}{\partial y} = 0 \tag{2.9}$$

$$\begin{aligned}\frac{\partial \overline{U}}{\partial t} + \overline{U}\frac{\partial \overline{U}}{\partial x} + \overline{V}\frac{\partial \overline{U}}{\partial y} &= -g\frac{\partial \zeta}{\partial x} \\ &+ \underbrace{\frac{1}{\rho h}\frac{\partial (h\bar{\tau}_{xx})}{\partial x} + \frac{1}{\rho h}\frac{\partial (h\bar{\tau}_{xy})}{\partial y}}_{\text{turbulence}} + \frac{\tau_{sx} - \tau_{bx}}{\rho h} \\ &+ \underbrace{\frac{1}{\rho h}\frac{\partial}{\partial x}\int_{z_b}^{\zeta} \rho \left(U - \overline{U}\right)^2 dz + \frac{1}{\rho h}\frac{\partial}{\partial y}\int_{z_b}^{\zeta} \rho \left(U - \overline{U}\right)\left(V - \overline{V}\right) dz}_{\text{dispersion}}\end{aligned} \tag{2.10}$$

$$\begin{aligned}\frac{\partial \overline{V}}{\partial t} + \overline{U}\frac{\partial \overline{V}}{\partial x} + \overline{V}\frac{\partial \overline{V}}{\partial y} &= -g\frac{\partial \zeta}{\partial y} \\ &+ \frac{1}{\rho h}\frac{\partial (h\bar{\tau}_{xy})}{\partial x} + \frac{1}{\rho h}\frac{\partial (h\bar{\tau}_{yy})}{\partial y} + \frac{\tau_{sy} - \tau_{by}}{\rho h} \\ &+ \frac{1}{\rho h}\frac{\partial}{\partial x}\int_{z_b}^{\zeta} \rho \left(U - \overline{U}\right)\left(V - \overline{V}\right) dz + \frac{1}{\rho h}\frac{\partial}{\partial y}\int_{z_b}^{\zeta} \rho \left(V - \overline{V}\right)^2 dz\end{aligned} \tag{2.11}$$

$$\begin{aligned}\frac{\partial \overline{\phi}}{\partial t} + \overline{U}\frac{\partial \overline{\phi}}{\partial x} + \overline{V}\frac{\partial \overline{\phi}}{\partial y} &= \frac{q_s - q_b}{\rho h} + \frac{1}{\rho h}\frac{\partial (h\bar{J}_x)}{\partial x} + \frac{1}{\rho h}\frac{\partial (h\bar{J}_y)}{\partial y} \\ &+ \frac{1}{\rho h}\frac{\partial}{\partial x}\int_{z_b}^{\zeta} \rho \left(U - \overline{U}\right)\left(\phi - \overline{\phi}\right) dz + \frac{1}{\rho h}\frac{\partial}{\partial y}\int_{z_b}^{\zeta} \rho \left(V - \overline{V}\right)\left(\phi - \overline{\phi}\right) dz\end{aligned} \tag{2.12}$$

In the above equations. the overbars denote depth-averaged quantities,

$$\overline{f} = \frac{1}{h}\int_{z_b}^{\zeta} f(z)\, dz \quad \text{for } f = U, V, \tau \ldots$$

ζ and z_b are respectively the free surface and bed elevation and $h\,(= \zeta - z_b)$ is the water depth.

Buoyancy effects cannot be accounted for in depth-averaged equations, so that the hydrodynamic model (consisting of equations 2.9 to 2.11) is perfectly independent of the scalar transport model (eq. 2.12).

Interface stresses and fluxes Vertical turbulent transport has been eliminated by depth-averaging and appears only as bed (τ_b) and surface (τ_s) stresses (eq. 2.10 and 2.11), and as bed (q_b) and surface (q_s) heat or mass fluxes (eq. 2.12).

The bed resistance is defined in terms of the fluid velocity, depth, and such bed properties as small-scale roughness and large-scale features (dunes, flat bed, antidunes). It may eventually be

affected by bed and suspended load. The bed resistance is particularly relevant in fluvial flows.

Usually, the bed stress is expressed with the help of a quadratic power law, i.e. its modulus reads :

$$\tau_b = \rho U_*^2 = \rho C_f \left(\overline{U}^2 + \overline{V}^2\right) \tag{2.13}$$

where C_f is a friction (or bed resistance) coefficient and U_* denotes the friction (or shear) velocity. Formula 2.13 is eventually modified in order to account for the transverse bed slope (Rodi *et al.*, 1981) when it is important (for instance, in the immediate vicinity of banks).

Numerous empirical formula which relate C_f to the bed features have been proposed. The most popular involve either the Chezy (C_h) or the Strickler (K_s) rugosity coefficient, which order of magnitude for different kinds of rivers can be assessed (e.g. (Carlier, 1986)). However, the determination of their optimal values for specific sites usually requires an additional calibration exercise.

In brief, terms τ_{bx} and τ_{by} may be expressed respectively by the formula $\tau_{bx}/\rho h = f_b \overline{U}$ and $\tau_{by}/\rho h = f_b \overline{V}$ where the friction factor f_b (dimension s^{-1}) depends either on the Chezy or the Strickler coefficient :

$$f_b = g \frac{\sqrt{\overline{U}^2 + \overline{V}^2}}{C_h^2 h} \tag{2.14}$$

$$f_b = g \frac{\sqrt{\overline{U}^2 + \overline{V}^2}}{K_s^2 h^{4/3}} \tag{2.15}$$

The free surface resistance depends on the wind velocity and on the state of the free surface (waves and their features). In shallow water bodies (Masbernat *et al.*, 1987; Rosello-Tournoud, 1991), free surface stresses τ_s may be the determining factors in the generation of circulation patterns.

τ_s is usually expressed with a formula analogous to 2.13 except the fact that the fluid velocity $(\overline{U}, \overline{V})$ is replaced by the wind velocity. The empirical surface resistance coefficient (analogous to C_f) depends also on the altitude at which the wind velocity is measured.

The heat flux at the bottom of most rivers and ocean areas can generally be assumed nul, except in case of geothermal sources. The heat flux at the free surface is due to radiation and evaporation phenomena. It depends on several factors, including the atmospheric pressure and temperature, the humidity, the wind, etc ...(Tassin, 1986; Vinçon-Leite, 1991).

The relative importance of mass surface and bottom fluxes depends on the studied species. For instance, considering reaeration at the water-atmosphere interface is usually important for

simulating the evolution of oxygen within a water body. On the other hand, in rivers submitted to human activities, atmospheric inputs in heavy metals like lead or zinc are usually much smaller than direct inputs from combined sewer overflows during rainfall events (Mouchel *et al.*, 1993). Benthic reactions may also induce a significant flux at the water-sediment interface (concerning for instance nutrients). Understanding and modelling these processes is a whole field of research, for which an extensive body of literature exists. Consequently, we shall not discuss the detail of these processes any further.

Turbulence terms Terms involving on the one hand $\bar{\tau}_{xx}$, $\bar{\tau}_{xy}$ and $\bar{\tau}_{yy}$, on the other hand $\bar{J}_x$ and $\bar{J}_y$, account respectively for the horizontal momentum transport and for the horizontal mass or heat transport due to the turbulent motion.

Dispersion terms Equations (2.10) to (2.12) contain so-called dispersion terms accounting for vertical non-uniformities of the U, V and ϕ profiles. In a process quite analogous to the one which gives birth to the turbulent correlations, these terms arise from splitting local quantities into depth-averaged values and deviations from these values and then carrying out the depth-averaging of equations. They are also due to the non-linearity of convective terms (this time, of equations 2.6 and 2.7). Their physical meaning (and effect) is similar to that of the turbulent stresses and fluxes in that both represent contributions to the transport of momentum, heat or mass induced by deviations from the mean flow.

While the dispersion terms in eq. 2.12 may eventually vanish as ϕ becomes homogenous over the water column, there will always be some vertical non-uniformity in the velocity profiles because the velocity has to go to zero at the river bed.

Dispersion terms may be especially important when there exist secondary motions in the cross-sectional plane. In natural rivers, such secondary motions may arise from large scale irregularities in the river bed and, of course, from river bends.

Summary *When contemplating the resolution of an environmental hydraulics problem, we are faced with the problem of choosing the most suitable model, which is not necessarily the most universal. In particular, when dealing with shallow water bodies, the choice of a depth-averaged representation is tempting. However, we have to determine first if it is sufficient enough to grasp the main features of the flow and transport processes or at least be conscious of the kind of errors or uncertainties this choice can imply.*

The horizontal distribution of depth-averaged velocities, temperature and species concentration within a water body is governed by several phenomena (apart from biogeochemical reactions

for the latter variables) : advection by the mean flow, transport due to turbulent motion, dispersion effects due to vertical non-uniformities, surface and bottom stresses. When looking for working equations; the main problem lies with the modelling of the turbulent and dispersive terms. We shall give hereafter a brief overview about how this problem has been approached and discuss the merits of different methods in the frame of application to environmental problems in large and shallow water bodies.

2.2 Turbulence modelling of surface water flow and transport

2.2.1 Scope of the section

The problem of turbulence modelling is discussed first for two reasons :

1. it is probably the most challenging task;
2. it is the choice of the level of turbulence representation which chiefly influences the choice of working equations in hydraulic models.

The turbulence motion can be thought as a tangle of eddies which stretch each other. The largest eddies are influenced by the boundary conditions of the flow. They interact with the mean flow (as the scale of both are similar), extracting kinetic energy from the mean motion and feeding it into the large scale turbulent motion. Due to vortex stretching, the energy is passed on to smaller and smaller eddies until viscous forces become active and dissipate the energy. This process is called the energy cascade.

Because of its interaction with the mean flow, the large scale motion has often preferred directions : thus, both the intensity and length scale of turbulent fluctuations are direction dependent. During the cascade process, the direction sensitivity decreases and can eventually vanish : the small scale motion becomes isotropic (which is called the "local isotropy concept").

It is primarily the large-scale motion that transports momentum, heat and mass and thus, which effects need to be simulated as far as practical applications are concerned. The basic question is twofold :

1. When evaluating the effects of turbulence, which part of the turbulence motion shall we represent explicitly ? none, a part (the largest eddies), all (the whole spectrum) ?
2. Once this choice is made, how can we take into account the action of small scales - not explicitly simulated - onto large scales ?

The first generation of turbulence models was developed at a time computers power was rather limited, when it was already a challenge to solve three-dimensional equations on a relatively coarse grid. Besides, it was based on the idea that turbulence is a perfectly random

phenomena, typical of a system possessing a large number of degrees of freedom. Hence, a **statistical** approach was deemed logical for solving practical problems. Emphasis was put on modelling the evolution of *averaged* quantities of the flow.
Reynolds averaging is applied to Navier-Stokes equations : in this framework, the task of turbulence models is to relate the second-order moments such as $\overline{u_i u_j}$ and $\overline{\varphi u_j}$ (eq. 2.6 and 2.7) to the mean-flow quantities so that the Reynolds equations become a closed set. This vision and approach to turbulence has prevailed during fifty years (Ha Minh, 1993).
nb : Few models have been developed which compute directly the depth-averaged turbulent terms such as $\bar{\tau}_{xy}$ or $\bar{J}_x$ (cf eq. 2.10, 2.11 and 2.12), apart from Rodi and co-workers depth-averaged $k-\epsilon$ model (cf A.3.3).

The development in the last twenty years of another approach can be ascribed to the following reasons :

- progress in experimental methods; (video techniques, velocity measurements with laser, ...)
  This allows a closer look at turbulence, which no longer seems to be so perfectly random. Observation of shear flows or mixing layers, for instance, has demonstrated that turbulence motion may include coherent, "organized" structures occurring at a quasi-regular frequency.
- the spectacular increase in computers speed and capacity;
  - Direct resolution of Navier-Stokes equations is now possible, albeit limited to low Reynolds numbers.
  - More generally, our definition of "small scales" and "large scales" can be modified. It is feasible to compute explicitly some part of the turbulent spectrum (the largest eddies - which may correspond to the coherent structures identified experimentally) and apply a statistical treatment - adequate for random motions - only to the remaining part (the smallest eddies). This is termed Large Eddies Simulation (LES).
- theoretical advances in the understanding of dynamic systems behavior and chaos, which lead to the conclusion that the chaotic features of turbulence do not necessary stem from a large number of degrees of freedom : thus they might be more predictable than was believed.

This better understanding of the physical and theoretical aspects of turbulence has justified renewed attempts at a **deterministic** modelling of its features.

However, state-of-the-art reviews (Rodi, 1980; ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988) underline that **few developments in turbulence modelling have been achieved in the frame of surface water flow studies. The models in use** are generally borrowed from the field of mechanical and aeronautical engineering, once they have been thoroughly tested there. This explain why they **belong mostly to the category of statistical models.** Consequently, the summary review we provide hereafter will exclude advanced modelling strategies (an introduction to these methods can be found in such books as (Lesieur, 1987; Chassain, 1993; Lesieur, 1994)) and focus only on classical, statistical, one-point turbulence closure models.

2.2.2 Statistical one-point closure models

(nb : A companion appendix (A) provides supplementary informations on the derivation of these models and on their spplication to fluvial hydraulics problems)

Definition One important feature of turbulence is that it is not a local phenomenon. Fluctuations observed at a given time at some point of the flow are not independent from fluctuations observed in the past at the point itself and, moreover, in some vicinity of the point.

"One-point closure models" do not give a fair account of the existence of these spatio-temporal correlations as, in these models, unknown moments and mean flow quantities are related with the help of variables or parameters expressed at a same point, a same time.

An introduction to so called *two-points closure models*, which aim at taking into account more adequately the part of history in the development of turbulence, can be found for instance in (Chassain, 1993), chapters 4, 9 & 11. However, as pointed out in (Silveira Neto, 1991) (chap. 1), two-points methods are far more complex than one-point methods and consequently, can be handled only in the simplified case of homogeneous flows. Homogeneous flows are such that statistical properties of the turbulence are independent from location within the flow. Real flows found in practical applications are generally inhomogeneous ! Two-points methods might nevertheless prove useful for dealing with inhomogeneous flows in the frame of LES, applying them to the modelisation of small scales (Lesieur, 1987; Silveira Neto, 1991). Research on this topic is under way.

Classification Numerous one-point closure models have been developed. Yet, they can be classified into two categories (Ha Minh, 1993) :

1. models based on the *eddy viscosity concept*, which rely on some *parametrization* of the turbulence;
2. models which *compute directly the turbulent stresses and fluxes*, with the help of transport equations which express their conservation and their spatial and temporal evolution.

We find also an intermediate class, the Algebric Stresses and fluxes Models, in which the anisotropy of turbulent stresses is taken into account without using complete transport equations ...so that the model is somewhat simpler and easier to solve. Finally, no matter how complex and sophisticated a model can be : it always contains some amount of empirical information or experience.

Turbulence parameterization and related models In such models, the local state of turbulence, and thus the turbulent correlations, are assumed to be described only by a few

parameters. The task of modelling is to postulate some relationship between parameters and correlations and to determine the distribution of the parameters over the flow field. The parameters should be related to the large scale motion, as it is this motion which governs primarily the mass and momentum transfer.

The link between parameters and correlations is usually established with the help of the **eddy-viscosity concept**, which assumes, that, as viscous stresses in laminar flow, turbulent stresses (fluxes) are proportional to the mean velocity (transported scalar) gradients :

$$-\overline{u_i u_j} = \nu_t \left(\frac{\partial U_i}{\partial x_j} + \frac{\partial U_j}{\partial x_i} \right) - \frac{2}{3} k \, \delta_{ij} \tag{2.16}$$

$$-\overline{\varphi u_j} = \Gamma_t \frac{\partial \phi}{\partial x_j} \tag{2.17}$$

where δ_{ij} stands for the Kronecker symbol (zero when $i \neq j$, unity if $i = j$).

In the above equations, the symbol k denotes the kinetic energy per unit mass contained in the turbulent motion :

$$k = \frac{1}{2} \overline{u_i^2} \tag{2.18}$$

In contrast to their molecular counterparts, the turbulent eddy viscosity ν_t and eddy diffusivity Γ_t are not intrinsic fluid properties : they depend strongly on the state of turbulence, and can vary considerably over the flow field. The Reynolds analogy between heat or mass transport and momentum transport suggests that ν_t and Γ_t should be closely related. Indeed, experiments have shown that their ratio varies but little across the flow and from flow to flow (Rodi, 1980). Thus, in most models, Γ_t is evaluated by $\Gamma_t = \nu_t / \sigma_t$, with σ_t, the turbulence Prandtl or Schmidt number, as a constant.

On dimensional grounds, the eddy viscosity is assumed to be proportional to a velocity scale $\widehat{V}$ and a length scale L of turbulent motion :

$$\nu_t \propto \widehat{V} L \tag{2.19}$$

If relations 2.16 and 2.19 are accepted, the problem of turbulence modelling shifts to the problem of assessing the distribution of $\widehat{V}$ and L, or of some combinations of these variables.

- The simplest models, usually termed "zero-equation models", relate ν_t directly to the local mean velocity distribution. For instance, in Prandtl mixing length model (cf. A.3.2), $\widehat{V}$ appears as the product of the mean velocity gradient by a typical length scale l_m which is prescribed by empirical formulae adapted to different kind of flows (free shear flows, wall boundary layers), so that we obtain finally $\nu_t = l_m^2 \left| \frac{\partial U}{\partial z} \right|$.

As ν_t depends only on *local instantaneous* mean flow characteristics, it is not possible to account for any transport or history effect (namely, influence of turbulence generated elsewhere or previously). Zero-equation models assume implicitly that turbulence is dissipated where it is generated.

- A decisive step in the development of turbulence models was to give up the direct link between the $\widehat{V}$ scale and the mean flow gradients and to determine this scale from a transport equation.

  If a single scale has to characterize the velocity fluctuations, the natural and most meaningful candidate is the kinetic energy k (cf eq. 2.18) which provides a direct measure of the intensity of the fluctuating velocities and is indeed mainly contained in the large-scale eddies.

  Some models, termed "one-equation models", replace L by $\sqrt{k}$ in eq. 2.19 and solve only one transport equation for k (relation A.15) while still relying on empirical formulae (cf (Rodi, 1980)) for assessing L. Due to the lack of universality of these formulae, they have not met with much success in the field of industrial or fluvial flows, whereas, on the other hand, satisfying applications have been reported in the simulation of marine hydrodynamics (Nihoul & Djenidi, 1987; Nihoul *et al.*, 1989a; Beckers, 1991). In fact, the most common turbulence models resolve also one equation for the turbulent length scale ("two-equations models", cf A.3.3).

  The length-scale determining equation need not have L itself as a dependent variable : any combination of the form $Z = k^m L^n$ will suffice as k is known already from solving the $k-$equation. The most popular combination is undoubtedly $\epsilon = k^{3/2}/L$ which represents the rate of energy dissipation by viscous action. Dissipation takes place into small eddies but, as the amount of energy available for viscous dissipation is in fact controlled by the energy extracted by the large-scale eddies, ϵ is nevertheless considered as a parameter characterizing the *large-scale* motion. When introducing k and ϵ, the eddy-viscosity determining equation 2.19 becomes

$$\nu_t \propto \frac{k^2}{\epsilon} \tag{2.20}$$

  The equations for k and ϵ which constitute the famous $k-\epsilon$ model (eq. A.15 and A.17) are derived from the Navier-Stokes equations. The exact equations include complex correlations whose behaviour is little known and which are modelled under somewhat drastic assumptions. Both the k and ϵ equations account for advective transport by the mean flow and by the fluctuating velocities (modelled by a diffusion term). The k equation (A.15) represents furthermore the effects of the interaction with the mean flow (net energy production), of viscous dissipation (energy destruction) and of buoyancy forces (stable stratification dampens turbulence while unstable stratification favours momentum and mass transfer). The ϵ equation (A.17) covers both the generation of vorticity through the

energy cascade (i.e. generation of smaller eddies, which implies a reduction of the overall effective length scale) and its viscous destruction (disappearance of small eddies, thus increase of the length scale). The $k-\epsilon$ model includes several empirical constants for which a standard set was determined (formula A.18, from (Rodi, 1980)), both through experiments and by application of the model to the simulation of a number of well-documented laboratory shear flows. A depth-averaged version of the $k-\epsilon$ model (eq. A.19 to A.21 and A.25), suitable for the calculation of shallow flows, especially in rivers, has been developed by Rodi and co-workers (Rastogi & Rodi, 1978; Rodi, 1980; Rodi *et al.*, 1981).

While they have been applied to a variety of flows with reasonable success, the two-equations models do not perform well under all circumstances. Indeed, as they consider a unique length and velocity scale, and consequently (cf eq. 2.20) an isotropic eddy viscosity, they cannot properly account for strong turbulence anisotropy (e.g in shallow water bodies, the turbulent motion is much more confined in the vertical than in the horizontal direction). Some models include a direction-dependent empirical correction of the eddy viscosity computed by eq. 2.20 . However, for achieving some generality, it may be necessary to describe the behaviour of the individual stresses and fluxes.

Turbulent stress/flux models (cf. A.3.4)

Exact equations for the turbulent stresses and fluxes can be derived from the Navier-Stokes equations (cf eq. A.26). Besides advection by the mean flow, the stresses undergo diffusive transport and are affected by viscous dissipation and by interaction both with mean flow and with fluctuating pressure. These terms require some proper closure for the equations to be workable. The kinetic energy k and the dissipation rate ϵ intervene in this modelling.

Complete stress equation models can prove burdensome, as, for general flows, it is necessary to consider 6 components for the stresses and one additional equation for ϵ. If scalar transport has to be computed, there are 4 supplementary equations, for the 3 components of the fluxes, and for the scalar fluctuation intensity $\overline{\varphi^2}$ which intervenes in the turbulent fluxes equations as does k in the turbulent stresses equations. This represents a set of 7 to 11 partial differential equations ! Consequently, several simplifications are made in practice.

The most immediate is to write complete equations for some components of the stresses only. For instance, for the computation of secondary motion in rivers, it is common, either to simply neglect the turbulent stresses in the streamwise direction (with respect to the longitudinal advective terms), either to express these with the help of the simple eddy viscosity concept : only the stresses related to transverse and vertical velocities are explicitly modelled.

Another popular method is to discard the gradients of the turbulent stresses in their gover-

ning equations. Thus, these become total differential equations, which are much easier to solve. Such models are termed algebraic stress/flux models (Rodi, 1980; ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988).

2.2.3 Application to typical river problems

A first state-of-the-art review relative to the application of turbulence models in hydraulics was achieved by Rodi (Rodi, 1980) and has been recently updated (ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988). From these reviews, we can note the following facts :

- Few developments in turbulence modelling have been achieved in the frame of surface water flows studies. The models in use are generally borrowed from the field of mechanical and aeronautical engineering.
- The applied models usually belong to the $k - \epsilon$ family. Little or no empirical coefficient adjustment with respect to the published values is performed.
- While there are numerous examples of turbulence modelling applied to typical engineering hydraulics problems, there are significantly fewer applications to environmental problems.

In rivers, the vertical length scale is usually much smaller than the horizontal one. Thus, it is tempting to neglect vertical components or variations of flow and concentrations and to focus on the forecast of depth-averaged (DA) distributions. However, neglecting three-dimensional effects can sometimes lead to erroneous predictions. There are at least two situations of practical interest where caution must be exercised : the near-field calculation of discharges (cf. A.4.1) and the modelling of areas where secondary currents are important (cf. A.4.2). As regards these cases, it is useful to refer to the impressive body of work achieved by Rodi and co-workers (Rastogi & Rodi, 1978; Rodi *et al.*, 1981; Naot & Rodi, 1982; Demuren & Rodi, 1983; Demuren, 1983; Pavlovic & Rodi, 1985; Demuren & Rodi, 1986; Keller & Rodi, 1988) : it helps to figure out the limit of applicability of depth-averaged models.

The **near-field of a discharge** is defined as the region where it significantly influences the flow field. There we can observe all or part of the following phenomena : flow deflection upstream of the discharge, setting up of back flow zones, vertical motions due to density differences between river and effluent, ... The extent of the near-field depends on site characteristics (i.e. flow ratios, discharge geometry). Its order of magnitude in large and shallow rivers is typically a few hundred meters long. The question is to assess whether an approximate description of this area, which

should theoretically be described by 3D tools, has consequences over the accuracy of forecasts in the downstream region. Rodi et al. investigated two kinds of *steady* discharges :

1. The fate of *coaxial slot-discharges* was simulated both with 3D and DA $k-\epsilon$ and related mean-flow models (Rastogi & Rodi, 1978). The models performance was assessed by comparing the forecasted depth-averaged velocities and isotherms with respect to laboratory data. It appears that in rough channels (with a rugosity typical of natural rivers), 3D and DA models yield fairly close predictions except for discharges with a very low Froude densimetric number (i.e. characterized both by a relatively slow entering velocity and large density differences with the receiving river).

2. The fate of *side discharges* (sewage effluent in the river Neckar and hydrothermal effluent in the river Rhine) was simulated by a chain of depth-averaged models and compared with field data collected in the first kilometer downstream the discharges (Rodi *et al.*, 1981). With respect to flume applications, one of the model constants was modified (cf eq. A.25 in A.3.3). Except for one case in the Rhine river, where the hot discharge was completely submerged, the depth-averaged $k-\epsilon$ model was found to perform fairly well, even in the immediate vicinity of the outfall.

3. The 3D $k-epsilon$ model mentionned in point 1 has been extended in order to deal with the case of side isothermal discharges and compared successfully to flume experiments for both smooth and rough bed conditions (Demuren & Rodi, 1983). The extension involved mainly a careful treatment of wall and open boundaries conditions. No tuning of the turbulence model parameters was applied, nor needed. However, the good agreement was reached only if a high-order numerical scheme (QUICK, cf (Leonard, 1979)) was used to solve the equations. Otherwise, numerical errors were found to blunder the predictions.

The tentative conclusion proposed by Rodi (Rodi, 1980; ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988), namely that "a depth-averaged $k-\epsilon$ model is generally adequate enough to deal with the near-field", though not definitely confirmed, has been supported by other works (e.g. cooling discharge in the river Loire (Lebosse, 1987)).
(nb : for details about the above studies, please refer to appendix A.4.1).

Secondary currents can develop in any kind of cross-section. They are more marked in irregular cross-sections (e.g. compound channels which combine a main channel with a shallow flood plain) and meandering reaches. The computation of the detailed features of the secondary motion is generally far beyond our practical interests. The important point is rather to assess how it influences the development of the mean flow depth-averaged or even cross-section averaged quantities (e.g. the distribution of streamwise velocity or bed shear stress) or how it contributes to and increases the weight of dispersion terms.

- 3D turbulence closure models have been applied rather frequently to the prediction of flow in experimental (straight) compound channels. Hitherto, the conclusion of these applications is far from being clear. No model seems able to reproduce every detail of the secondary motion. While some researchers strongly advocate the use of algebraic stress model (ASM) written for every stress component, some others have shown that simplified ASM or standard $k-\epsilon$ models perform just as well as regards the forecast of streamwise velocities and bed stresses (cf. A.4.2). (Keller & Rodi, 1988) even present examples where a 2D $k-\epsilon$ depth-averaged model provide a reasonable assessment of the flow, apart from very limited areas of the cross-section (namely, the steep boundary between the main channel and the flood plain).

- The development of flow and transport in meandering channels has been somewhat less studied.

  (Demuren, 1983; Demuren & Rodi, 1986) applied satisfactorily a 3D $k-\epsilon$ model, where some anisotropy of the eddy viscosity was empirically introduced, to the interpretation of experiments on two kinds of strongly meandering flumes (cf. A.4.2).

  The depth-averaged $k-\epsilon$ model developed by Rodi and co-workers was also compared to these data (Pavlovic & Rodi, 1985). In other applications (e.g. (Rodi *et al.*, 1981; Keller & Rodi, 1988)), the turbulent stresses and fluxes computed by the $k-\epsilon$ model have seemingly accounted adequately for the effect of dispersive terms purposely omitted in the mean-flow and scalar equations. Here, it was not the case : a transverse dispersion term needed to be added explicitly to the scalar transport equation. This term was modelled with the help of the three-dimensional model introduced in (Demuren, 1983). Then, (Pavlovic & Rodi, 1985) noted a a reasonable agreement between the observations and the model outcomes.

  An even simpler model was satisfactorily applied by (Mary, 1982) to the simulation of an heated discharge into a laboratory S-shaped flume (cf. A.4.2). The model was based on St-Venant equations (cf 2.4.5) : no refined turbulence modelling was applied but turbulent and dispersive terms were both represented by a diffusion-like operator. As in (Pavlovic & Rodi, 1985), the thermal dispersion coefficients were evaluated with analytical formulae relating them to the vertical distribution of velocity. Yet, this distribution was not supplied by a 3D model but by empirical, experimental relationships.

The provisional conclusion of this review is once again encouraging as regards the depth-averaged $k-\epsilon$ model : it appears to provide a reasonable assessment of the main flow characteristics in moderately irregular natural channels, for steady flows. As regards strongly meandering channels, both 3D and 2D $k-\epsilon$ models need to be somehow empirically modified or completed. However, there is no strong evidence indicating that more complex turbulence models perform far better.

Summary *There exists a large range of possible turbulence closure models. The simplest ones rely on the eddy viscosity concept, which postulates an analogy between turbulent transport and molecular diffusion, and on a parameterization which relates the eddy viscosity to local mean flow gradients. The more advanced ones involve a direct simulation of part of the turbulent spectrum. Intermediate models do more or less justice to the transport and history effects which govern turbulence and to its anisotropic features.*

The application of turbulence closure models to environmental hydraulics problems is still far from being widespread. As regards two problems of particular interest in fluvial hydraulics, namely the near-field calculation of discharge and the representation of the effects on pollutant transport of secondary currents in bends, encouraging results have been observed, albeit only in steady-state flow conditions, with the help of a $k - \epsilon$ depth-averaged closure model.

2.3 Dispersion modelling

This section is restricted to dispersion modelling in the field of fluvial hydraulics, for non-stratified flows. As stated above, in this context, flows are mostly two-dimensional in plan and their vertical component may often be discarded. Besides, bed friction appears to play a dominant part in the generation of turbulence and secondary motions. As we shall see later on, this feature is reflected in the various empirical formula which aim at assessing dispersive parameters.

Moreover, we focus on the intermediate and far-field downstream discharges, i.e. areas where the spreading is mainly controlled by the receiving water body hydraulics and not by the discharge intrinsic dynamics.

An introduction to dispersion modelling in other situations may be found for instance in (Fischer *et al.*, 1979) (mixing in unconfined water bodies like oceans, mixing in reservoirs, dilution of jets and plumes in coastal waters) or (Tassin, 1986; Vinçon-Leite, 1991) (vertical mixing in lakes) ...

As for turbulent modelling, we recall hereafter only the main steps in dispersion modelling. The more curious reader should refer to appendix B. In particular, we report in B.2 the complete development of Taylor's analysis of the dispersion process and its subsequent extension by Fischer, whereas in section 2.3.1 we only discuss the hypotheses underlying these approaches. Similarly, 2.3.2 is but a summary of the review of formula and mixing experiments, included in B.3.

2.3.1 An advective process modelled by a diffusion operator

In the case of turbulence modelling, the emphasis was put on the assessment of the turbulent *stresses* and the modelling of the turbulent *fluxes* was but a by-product of this effort. In contrast, the dispersion process has been mostly studied in the frame of the analysis and simulation of the fate of pollutants in water bodies. Thus, the modelling of the dispersive contributions in the flow equations is this time derived from the modelling of dispersive terms in scalar equations.

Dispersion is essentially an *advective process* as it stems mainly from *differential advection* with respect either to depth or cross-sectional averaged velocities (see expression of dispersive terms in eq. 2.10 to 2.12). Turbulent transport is often modelled by a diffusion-like operator (see 2.2.2) on the ground that the randomness of turbulent fluctuations is analogous, albeit at a larger scale, to the random walk of fluid molecules which causes molecular diffusion. It turns out that dispersive terms are generally modelled by a diffusion operator too, although the variations of velocity with respect to its mean certainly do not exhibit the kind of randomness we find in turbulent and molecular motion.

In contrast to the development of turbulence modelling, which, albeit partly empirical, relies on the introduction of more and more physics in the models, the theoretical developments relative to modelling of dispersive terms are rather sparse and the justification for their representation by a diffusion operator is poorly substantiated. The main steps were achieved by :

- Taylor (Fischer *et al.*, 1979), by studying dispersion in a pipe at asymptotically large times, demonstrated that the mass transport in the streamwise (x) direction was proportional to the mean concentration gradient in this direction. Thus, the evolution of mean concentration $\overline{C}$ was found to obey the well-known one-dimensional advection-dispersion equation :

$$\frac{\partial \overline{C}}{\partial t} + \overline{U}\frac{\partial \overline{C}}{\partial x} = K\frac{\partial^2 \overline{C}}{\partial x^2} \tag{2.21}$$

  where $\overline{U}$ stands for the mean velocity and K for a *dispersion coefficient* which is a rather complicated integral function of the velocity distribution over the cross-section (cf eq. B.9).

- Elder (Elder, 1959) applied Taylor's analysis to flow in an infinitely wide channel. Assuming that the vertical profile of velocity was governed by the log-law, he proposed first an evaluation of the vertical turbulent diffusivity, then an evaluation of the longitudinal dispersion coefficient, by assuming that the dispersive terms were due to vertical non-uniformities only (cf B.3.1.)

- Fischer (Fischer, 1967; Fischer *et al.*, 1979) revisited Taylor's analysis, providing a sounder demonstration for it. He notably extended this analysis to the computation of longitudinal dispersion coefficients in rivers, in the context of cross-sectional averaged, one-dimensional

representations (cf B.3.3). He highlighted why Elder's formula was giving poor results in field studies, pointing out that a natural river could not be likened straightforwardly to an infinitely wide channel and that the main contribution to longitudinal dispersion derives from transverse and not vertical velocity variations. He proposed an exact formula for the computation of the dispersion coefficient, which is formally similar to Taylor's formula (cf eq. B.25).

Taylor's or Fischer's developments are based on the same set of hypothesis :

Hyp. 1 The studied flow is steady, uniform and strongly unidirectional. Velocity deviations from their (cross-sectional or depth) mean are small.

Hyp. 2 One considers a pollutant cloud far away from its source so that :

- concentration deviations from their cross-sectional mean are small;
- the mean concentration distribution is a smoothly, slowly varying function, both in space and time.

In that case, the concentration distribution has reached some kind of steady-state equilibrium over the water column or the cross section.

Unfortunately, in real rivers, the assumptions relative to the flow are seldom fulfilled :

- Flow is rarely stationary. However, it is often slowly varying, apart from exceptional events (floods, highly transient input ...), so that this is not the major objection.
- Natural rivers exhibit variability, so that the assumption of uniformity does not hold.
- Velocity deviations from their mean can be far from being negligible, especially when the bed or banks contain large scale irregularities, or when the river is meandering, all things which favour the development of strong secondary currents within the cross section.

Besides, achieving the quasi-homogeneity postulated in Hyp. 2 takes some time. There is an *initial period* after the pollutant injection, or, equivalently, an *initial zone* downstream of the pollutant source, needed so that the concentration deviations become small enough. It is only after this delay or when the pollutant flows out of this zone that the evolution of concentration can eventually be represented by a dispersion equation. On dimensional grounds, it appears that this initial time is proportional to a ratio of the form L^2/K where L is the typical length scale over which homogeneity must be nearly achieved (depth or river width) and K a mixing coefficient (either the vertical turbulent diffusivity or a transverse dispersion coefficient). Unfortunately,

the proportionality factor is influenced by local, site specific, irregularities, so that it is always advisable to check in situ, by direct measurements, the extent of the initial mixing period or length.

In summary, from a "philosophical" point of view, the use of a diffusion operator and related dispersion coefficients to account for the dispersive terms in the flow and transport equations is somewhat dubious. Its theoretical justification is built on sandy ground and the limits of applicability are frequently ignored. However, this approach is widespread as adequate dispersion coefficients can frequently be empirically evaluated.

2.3.2 Empirical formulae for dispersion coefficients

2.3.2.1 Foreword

The evaluation of dispersion coefficients is based partly on theoretical analysis, which succeeded mainly in disclosing which factors control dispersion, and partly on the interpretation of laboratory or field experiments. The values of dispersion coefficients reported in the literature are quite scattered and the related empirical formulae for forecasting them are numerous. Such confusion may be ascribed partly to the variety of experiments and studied sites, partly to the difficulties inherent to data collecting and processing :

- First, it can be difficult, especially when dealing with field experiments, to ensure a satisfying *representativeness* of the experiments and a correct *reproductibility* of the results and to *control completely the experimental conditions* (see 2.4.2).
- Secondly, *measurements generally do not allow one to distinguish between the effects of turbulent mixing and dispersion*, so that the proposed dispersion coefficient generally accounts for both contributions to the pollutant spreading.
- Lastly, *the coefficient evaluation depends on the tool used to interpret the data.*
  - Some authors consider only the moments of the dye cloud at different locations and times. They compute the dispersion coefficient by assuming that the dye cloud variance grows linearly with time and that the growth rate is proportional to the dispersion coefficient (a relation which, in fact, stands only if the pollutant obeys indeed the advection-dispersion equation and if the flow is perfectly steady and uniform).
  - Others use a mathematical model simulating the advection-dispersion equation and determine the dispersion coefficient by tuning the model with respect to observations, in that case complete pollutographs. The value of the coefficient depends undoubtedly on the detail of the hydraulics description in the model : some simply assume the flow

to be uniform; the most sophisticated have been using stream-tube models (cf. (Cunge *et al.*, 1980b)) which allow description to some extent of the differential convection in the streamwise direction but do not take into account any possible transverse velocity.

2.3.2.2 Usual formulae

Elder (Elder, 1959) demonstrated that, in an infinitely wide channel, the vertical turbulent diffusivity follows a parabolic profile over the depth (cf B.3.1). Such a proposition was confirmed by field experiments in straight flumes or natural rivers (however, the constants defining the parabola are not exactly as prescribed by Elder (Nezu & Rodi, 1986)). The depth-averaged value of this diffusivity, K_z, was found to be proportional to the product $U_* h$ where h denotes the water-depth and U_* the shear friction velocity at the river bed :

$$K_z = 0.067\, U_* h \tag{2.22}$$

From this, Elder deduced that the streamwise dispersion coefficient K_x was also proportional to $U_* h$:

$$K_x = 5.93\, U_* h \tag{2.23}$$

Pursuing the same line of thought, he investigated the dependence between the transverse dispersion coefficient, K_y, and $U_* h$ through mixing experiments in straight and shallow flumes and proposed :

$$K_y = 0.23\, U_* h \tag{2.24}$$

Elder's conclusion relative to the strong correlation between K_y and $U_* h$ was confirmed by other researchers. However, the magnitude of the proportionality coefficient $\alpha_y = K_y / U_* h$ depends on the river's geometric features. Indeed, bottom and sidewall irregularities, as well as bends, favour the development of secondary currents which increase transverse spreading, thus leading to a corresponding increase in the dispersion coefficient. It is possible to classify the open channels into three groups (cf B.3.2) :

straight flumes and man-made channels

$0.1 \leq \alpha_y \leq 0.3$ (average given by Elder's relation 2.24)

slightly irregular or gently meandering rivers

$0.4 \leq \alpha_y \leq 0.8$ (average value 0.6)

sharply curving channels (laboratory and field)

α_y can take values up to 3 or 4. Fischer (Fischer, 1969; Fischer *et al.*, 1979) and later Yotsukura et al. (see (Fischer *et al.*, 1979; Holley, 1987)) suggested empirical formulae to account for the effects of bends, where the ratio of the bend radius of curvature to either river depth or width intervenes.

As regards the streamwise dispersion coefficient, we must consider two cases : the coefficient as understood in depth-averaged two-dimensional models, which is related only to vertical non-uniformities, and the coefficient used in one-dimensional model, which accounts for any non-uniformity over the whole cross-section. To our knowledge, the only formula available for the first case is that of Elder, i.e. 2.23.

When looking at the literature dealing with the onedimensional streamwise dispersion coefficient, it appears, first that Elder's relation 2.23 dramatically underestimates it, secondly that the correlation between K_x and $U_* h$ is poor : the proportionality factor varies within a factor of 100 or so.

As mentionned above, Fischer (Fischer *et al.*, 1979) demonstrated the dominant part played by transverse velocity variations in the development of longitudinal dispersion and came to the conclusion that K_x was indeed width dependent (cf B.3.3). He proposed :

$$K_x = 0.011 \frac{U^2 B^2}{U_* h} \tag{2.25}$$

where U denotes the mean velocity and B the width. Albeit not perfect, this formula has been found to agree with observations within a factor of five or so (Fischer *et al.*, 1979; Marivoet & Van Craenenbroeck, 1986; Mouchel, 1989; Rigaudière, 1992), which is reasonable enough, considering the approximations made in its derivation. A host of other formulae for K_x have been suggested (see (Lassale, 1992)). Yet, none seems to surpass the others, and Fischer's 2.25 remains the favourite of most practicioners.

Summary *The dispersive terms are usually represented with the help of a diffusion operator. The theoretical justification for this modelling approach is rather poor. However, it has proved to work reasonably well in practice.*

The dispersion coefficients appearing in this operator have been studied with the help of laboratory or field dye-tracing experiments, where the relative contribution of turbulent mixing and dispersion is usually impossible to weight. As regards depth-averaged representations, the related streamwise and tranverse dispersion coefficients are found to be correlated to the product $U_* h$ where h denotes the water-depth and U_* the shear friction velocity at the river bed.

2.4 Additional considerations for model choice

We have reviewed in the previous sections different approaches available to model the turbulent and/or dispersive terms, and consequently to close the three-dimensional or depth-averaged two-dimensional flow and transport equations. From a theoretical viewpoint, some methods,

especially as regards turbulence modelling, seem more firmly physically-based than others, and thus could be deemed superior. However, when applying hydrodynamic models to large water bodies, we are faced with specific problems that one usually does not meet when simulating plain laboratory experiments.

2.4.1 Interaction between conceptual and numerical modelling

For solving the flow partial differential equations, we need to transfer them on a discrete frame with the help of finite difference or finite element approximations, so that their solution reduces to a problem of matrix algebra. Thus, we need to apply a spatial and temporal discretization to the studied domain. The physical processes which occur at a smaller scale (either in time or space) than the computational grid are "filtered out" from the flow description. Then, one has to decide if it is relevant to re-introduce the effects of these "subgrid motions" on the explicitly resolved scales and, if so, how to do it. This problem of subgrid scale motion modelling is formally similar to the turbulence closure problem, except generally for a scale difference. *Ideally, the chosen time and space steps should correspond to a clear-cut boundary between quite different physical processes, so that their conceptualization can be unambiguous.*

In simple laboratory experiments, the distinction between the "mean flow" and the superimposed "turbulent fluctuations" can generally be made easily enough and moreover, as the studied domain is of limited extent, it is usually possible to set the time and space steps so that they suit the physics without the computation becoming too cumbersome. Besides, boundary conditions, which may strongly influence the interior flow pattern, are perfectly known. On the other hand, *geophysical systems are characterized by a* **continuum** *of interacting motions, with different length and time scales,* and which are *stimulated by poorly known forcings.* Existence of this continuum is particularly obvious when considering for instance marine variability, where we find everything ranging from phenomena dominated by global scale climatic processes (time scale, $T_s \simeq 1$ year) to 3D eddy turbulence ($T_s \simeq$ a few seconds), with intermediary processes such as tidal ($T_s \simeq 1$ hour) and seasonal ones ($T_s \simeq 1$ month) (Nihoul *et al.*, 1989b). *The different time and length scales are usually known only approximately so that drawing a line between the different processes is somewhat tricky.* Besides, as we are dealing with large domains, computational efficiency requires that :

- we discard eventually some components of the flow and transport phenomena (e.g. the vertical ones);
- we choose relatively large space and time steps;
- we respect some relationships between these steps which are imposed on mathematical grounds (i.e. to ensure the stability and convergence of the algorithms in use).

It may be difficult to reconcile these constraints with a sound physical approach.

Having some previous idea about the dimensionality of the problem and the importance of subgrid scale motion could guide us as regards the relevance of more or less complex approaches. Unfortunately, this seems quite difficult to assess "a priori".

For instance, shallowness does not necessarily warrant that a plain depth-averaged model brings sufficient insight into the flow dynamics, as illustrated by the study of the Hudson-Rarity estuary (Oey *et al.*, 1985a; Oey *et al.*, 1985b; Oey *et al.*, 1985c; ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988).
By the way, the same work illustrates both the interest lying in refined turbulence modelling and the ambiguity in doing so. Oey et al. quasi-three dimensional model neglects all turbulent stresses due to interaction of horizontal velocity fluctuations but still considers the vertical gradients of momentum ($\overline{uw}$ and $\overline{vw}$ stresses intervening respectively in the U and V momentum equations) and salt flux ($\overline{\varphi w}$). These are computed with the help of the eddy viscosity-diffusivity assumption and a two-equations model which dependent variables are the kinetic energy and a turbulence length-scale (Oey *et al.*, 1985a). The model agrees reasonably well with the observations. Yet, one may wonder if the actual discrepancies are to be ascribed to the turbulence model (for instance, the neglect of some terms ...) or to a lack of spatial resolution : even if fine for an estuary, the horizontal grid size is far from being small ($\Delta x = \Delta y = 530m$!).

Some researchers, e.g. Abbott, have been advocating the use of simple criteria.
When considering the numerical modelling of depth-averaged free surface flows, Abbott (Abbott *et al.*, 1981) suggested that the importance of the various sub-grid scale components may be characterized in many situations by the dimensionless factor $h/\Delta x$ where h is the water depth and Δx a typical mesh size. Then, he reckoned that when this number is very small, flow processes, including circulations, are mainly controlled by the distributions of bathymetry, the convective momentum, the bed resistance and the wind stress, and obviously the initial and boundary conditions, so that the subgrid scale modelling is relatively unimportant and ineffective.

Apart from relying on empirical criteria or on the comparison of different models to determine the true influence of subgrid scale motion, a third approach has been followed by Nihoul and co-workers. They discussed (Nihoul *et al.*, 1989b) how the formalism of turbulence modelling can be extended to the modelling of fluctuations related to processes whose time scale is up to a day or a week (tides, storm surges for instance) and how the combined effect of the Reynolds stresses due to turbulent motion and of the "Leonard stresses" (see (Abbott *et al.*, 1981)) due to advective subgrid-scale motion can be effectively accounted for. Application to the modelling of shallow stratified seas with a $k - \epsilon$ model or a k model (where the turbulence

length scale is empirically prescribed) is subsequently illustrated with reasonably encouraging results (Nihoul *et al.*, 1989b; Beckers, 1991).

In the case of fluvial hydraulics and depth-averaged models, there is a common feeling that, except in special areas (e.g. discharge near-fields) pressure gradient and bottom friction are the dominant factors in the flow momentum equations. On the other hand, for a correct appraisal of scalar transport, there is clearly a need for a precise assessment of dispersive and turbulent terms. How this can be achieved is another question.

Is it practical to model with refinement subgrid scale motion ? The reported applications of depth-averaged $k-\epsilon$ models, while encouraging, deal with steady-state situations (Rodi *et al.*, 1981; Lebosse, 1987). The stationarity alleviates considerably the computations. Besides, on a theoretical point of view, the proposed treatment of dispersive terms (implicitly included in the turbulent contribution), which proved already dubious (and was corrected) in the case of mass transfer in meandering channels (Pavlovic & Rodi, 1985), has neither been justified in the case of transient flows. In such cases, models using empirical diffusion coefficients accounting both for turbulence and dispersion, if carefully tuned, may prove as useful as DA $k-\epsilon$ models (Mary, 1982) while, at the time being, more sophisticated turbulence models do not seem quite able to bring outstanding improvement in the understanding of the flow dynamics (cf A.4.2).

In summary, it turns out that the application of refined turbulence closure models to typical environmental problems is neither straightforward nor undoubtedly warranted because :

- on one hand, there is a variety of cases for which the interest of actual turbulence models with respect to empirical diffusion models has not been assessed;
- on the other hand, *computational constraints often prevent us from making a proper distinction between the effects of different subgrid scale motions*, some effectively dependent on physical phenomena such as turbulence, others purely related to a lack of spatial and temporal resolution.

2.4.2 Data availability

Another important problem lies with the availability or, more often, the lack of suitable data. We can first distinguish between data related to the description of the studied domain and data related to the description of flow and transport patterns within this domain.

In the first category we find geometric and bathymetric data, information related to the bed composition and roughness, ... For instance, it is plain that if the bathymetry of a domain is poorly known, interest in a very refined subgrid motion modelling is dubious, as, implicitly, it

needs to be tuned in order to account for gross uncertainties plaguing the mean motion quantities (e.g. exact depth). The assessment of the bed rugosity also raises a problem. There are some formulae which relate the friction coefficient to the size of bed materials or eventually to the bed features (dunes, etc ...) However, these formulae are laboratory based and besides, such detailed bed characteristics are not easy to monitor in situ. Consequently, the assessment of rugosity usually stems, in rivers, from calibration of monodimensional backwater calculations and, in oceans or coastal areas, from rules of the thumb or once again, from a calibration exercise. Bed resistance properties define, for instance, the boundary conditions of turbulence models so that approximations of these can cast doubt on model forecasts.

In the second category, we find any information characterizing directly a feature (velocity, flow or depth measurements) of the flow pattern or describing a consequence of the flow development (temperature, species concentrations).

- Obviously, indirect measures such as the latter are more difficult to interpret. Indeed, a species distribution results from the action of the flow but also from the initial distribution, boundary conditions (various sources and sinks) and biogeochemical reactions, so that the action of flow alone is sometimes difficult to single out. Unfortunately, sampling water and measuring its contents often proves easier than making a direct assessment of velocities. When possible, the use of conservative (chemically inert) species, such as in dye-tracing experiments, is preferable for calibrating flow and transport models.

- In addition, much effort must be made in order to collect *representative* data.

  - The first problem lies in the randomness inherent in any flow or transport features. Instrumental records made in exactly the same conditions - although qualitatively similar - cannot generally be superposed. Only the ensemble average of a large number of recordings made in identical conditions appears to be reproducible. Unfortunately, it is barely possible to proceed so. A typical example is a dye-tracing experiment. Of course, as we are injecting a great quantity of dye, it will somewhat smooth out the differences between individual paths for individual fluid particles. However, what we are observing is only one among the possible trajectories of the dye cloud. Nothing tells us it corresponds to an "average" behaviour. Nevertheless, we usually calibrate a more or less complex transport model with this kind of data and expect it can help us to forecast a "typical" pollutant transfer.

  - Moreover, without previous knowledge of the flow and transport dynamics, it is not straightforward to establish a good sampling frequency, both in space and time. Most successful experiments generally go together with the development of a model of the system. Even if uncalibrated, the model can often bring qualitative insights into the system dynamics, their time scales and marked heterogeneities. Yet, practical

considerations (time and money) often preclude the application of a comprehensive monitoring program.

- Finally, the observations may be influenced, blurred, by some external forcing variables (open boundary conditions, wind effects, accidental discharges, ...) whose importance is neither immediately nor easily appraised.

Common sense dictates that one avoids including in a model assumptions and corresponding sub-models that one cannot really check, unless it turns out to improve significantly the portability of the model. For instance, depth-averaged measurements do not allow one to distinguish between turbulent motion and dispersive motion effects. In that case, refined turbulent modelling should be considered only if a simpler model, where turbulence and dispersion effects are accounted for by a common diffusion operator, either fails to reproduce observed data or can be calibrated in different situations but at the cost of such variability of the tuning parameters that extrapolation or interpolation to other conditions is foolhardy. If and only if the refined model can deliver fair forecasts without site-specific calibration (e.g. the $k - \epsilon$ model with its standard set of constants), is it then preferable !

2.4.3 Model objectives and further applications

The requirements imposed on a model also depend on the use we intend to make of it.

A model can be considered for instance purely as a research tool which merely helps one to formulate one's conceptions and beliefs about the functioning of a system. The modeller could then be interested primarily in checking that his description ensures a qualitative, rather than quantitative, agreement between the simulated and observed behaviours. For instance, he could simply want the model to give a good picture of the essential dynamics, changes of states, ... of the system, discounting a very precise assessment of the various quantities that characterize these different stages, and focussing only on the fact that they can or not exceed some threshold.

On the other hand, when applied to engineering problems (design of an outfall, implementation of a policy discharge control), the demands concerning the predictive ability of the model can be significantly different and more stringent. However, if the model needs to be used frequently and by unskilled users, different from the model builders, some balance between accuracy, efficiency and robustness must be achieved, frequently to the detriment of accuracy.

Besides the faithfulness with which the model reproduces observations made in controlled conditions, one must consider which kind of scenario it will simulate. Let us assume we want to analyze the fate of sewage effluents. In dry weather, it is possible to achieve a good assessment

of the pollutant input. On the other hand, forecasting accurately the volume and quality of combined storm sewer overflows from any slightly complex sewer network is still a challenge. In the second case, discrepancies between simulation and observations in the receiving water body can be ascribed equally to model failures or to poor knowledge of the pollutant inputs. In such situations, a simple model which allows straightforward and rapid analysis of the sensitivity of forecasts both to model assumptions and scenario uncertainties is more advisable than a complex one whose formal coherence and accuracy can lead to a blind trust in dubious predictions.

2.4.4 The Seine River case study

We have stressed that the choice of an efficient modelling strategy is tightly linked to the features of the problem at hand. It is now high time to introduce the framework in which we intend to develop and apply a surface water flow and transport model.

Our work is part of a research program, the Piren-Seine, devoted to the study of water quality problems in the Seine River. In particular major problems are caused by combined sewer overstorms (CSO) occurring in Paris area. We sum up hereafter the characteristics of the CSO problem.

River features The Seine river cannot be considered as a truly natural water body. Indeed, its course has been drastically modified in order to allow for barge navigation. Evolutions of the bathymetry are surveyed and the navigation channel is maintained through dredging works.

The flow is regulated first by large reservoir dams, located about one hundred kilometers upstream of Paris, which reduce high flows, floods, and, at the contrary, ensure water availability during low flows (May to September). Secondly, navigation locks distributed along the river control water elevation so that barges can pass. Consequently, the hydraulic behavior of the Seine river during low flows is closer to that of a series of pools pouring into each other than to that of a naturally flowing river.

Across and downstream of Paris, the Seine is rather shallow, the average water depth being 5 meters (except during autumn or winter floods) while the river width varies between 80 (on each side of an island) and 200 m. Width to depth ratio is thus in the range of 15 to 40. Because of dredging, the bottom slope is nearly null.

With the help of the reservoir dams, low flows discharge keeps between 80 and 120 m^3/s. The Seine is slow flowing, with average velocities about 10 cm/s for a 100 m^3/s discharge.

Various dye-tracing experiments indicate that the typical mixing length required for the homogeneization of side discharges is 5 to 10 kilometers, depending on flow rate and reach

geometry.

The CSO problem Paris and its suburbs form a heavily urbanized area of 10 million inhabitants. Its extensive sewer network has the following characteristics :

- A large part of it, notably the network which runs below the town of Paris itself, is a combined system.
- The network has been originally designed in order to send all wastewater to the treatment plant of Achères, located 50 km downstream of Paris. Achères is a huge plant which releases into the Seine River a daily average treated flow of 27 m^3/s. New facilities have been recently built upstream of Paris (Valenton, Noisy-le-Grand) but Achères, and the five large collectors which supply it, remain the central pieces of the sewage system. These large collectors have contact points with the Seine River (for instance, some pass below the river through siphons) which are favourite locations for overflows !
- Due to the urbanization, runoff is very important. During the big storm events (occurring most frequently at the end of June and in September), the sum of overflows may double the flow rate in Seine. Single overflows with peak discharge of about 50 m^3/s have been observed.

The impact of CSO has been first highlighted by fishkills, linked to oxygen depletion induced by the degradation of the large amounts of organic matter brought by the CSOs. CSOs contain also a significant amount of heavy metals and hydrocarbons, which cause longer-range impacts.

Monitoring the CSOs impacts The CSOs influence on the river has been surveyed first with the help of fixed monitoring stations (O2, temperature, nutriments, solids ...), operating continuously during the low flows of 1991 and 1992. Besides, several events have been looked at more closely, the polluted masses issued from the CSOs being sampled as they were flowing down the river.

A qualitative observation of these measurements, backed by simplified modelisation of the biological processes, allowed to conclude the following :

- Oxygen depletion, and a drastic perturbation of the O2 diurnal cycle, lasts two or three days after a big storm. This indicates that part of the pollution is transported at a slower pace than the water. It might correspond to the solid fraction of the pollutants, which undergoes a process of deposition and re-suspension.

- Impact are not strictly proportional to the magnitude of the overflows, smaller overflows being sometimes nastier than larger ones. This might be due to the fact that stronger runoff carries off larger, heavier solids, which, once in the quiet river, deposit more easily.
- The initial state of the receiving water, notably its eutrophication, also appears to play an important part in controlling the magnitude of the impact.
- Classical models like Streeter & Phelps one cannot represent these effects.
- As concerns heavy metals, CSOs influence, both in the short and long-range, is obvious.

The right choice ? Now, a truly quantitative assessment of impacts requires more detailed tools, notably as regards hydraulic description. Indeed, due to the number of CSOs, their proximity, their transient features, the relatively long time and distance required for homogeneous mixing, the representativity of measurements, both at fixed stations and along the river course, is unclear. Do they represent an average situation ? Do they over/under-estimate the troubles ? It seems essential to understand flow patterns and the mixing of CSOs plumes if we want to study properly, quantitatively, the biogeochemical processes which control oxygen and heavy metals fate.

The minimal requirement is a two-dimensional tool (due to the geometry of overflows, it may be assumed they mix quickly over the water column). This tool must of course be able to deal with transient flows. Yet, as we shall see later on, while flow rates and bathymetry are fairly well known, the level of uncertainty attached to water level measurements is notable. Besides, we have few data relative to typical velocity profiles in the Seine, none as regards turbulence characteristics. Consequently, it does not seem proper to start with a refined turbulence modelling.

In conclusion, we considered it would be right to undertake first of all the development of a model based on the two-dimensional St-Venant shallow water equations, which equations we recall in next section.

2.4.5 The St–Venant model : a logical choice for depth-averaged situations

To sum up the comments made in sections 2.4.1 to 2.4.4, it appears that *considerable gain can be achieved when tailoring a model to a specific application rather than aiming at universality.* Adapting a whole model to some application is a process formally similar to the one which leads to the choice of a numerical algorithm. In the latter case, one must ensure that numerical by-products, for instance numerical dispersion, remain negligible with respect to "natural" physical variability and physical dispersion of the simulated variables. In the first case, *the model should be developed so that the consequences of its formulations are either easily distinguished from, or*

significantly smaller than, the consequences of realistic and relevant changes in initial system conditions, inputs or forcing variables.

Consequently, *the first step when modelling depth-averaged flows, and the logical choice in the case of the Seine River study (cf 2.4.4), is usually to apply the St-Venant shallow water equations* which rely on the following simplifications :

- The bed resistance depends on the depth-averaged velocities according to a quadratic power law (cf 2.1.3).
- When applied to river flows, the wind stress and Coriolis force are usually neglected : the bed friction remains the only momentum sink.
- Both in the momentum and scalar transport equations, the turbulent and dispersive contributions are represented by a single diffusion operator, whose coefficients are given by one of the formulae presented in 2.3.2.2 or B.3.

According to these assumptions, eq. 2.9 to 2.11 become :

$$\frac{\partial h}{\partial t} + \frac{\partial hU}{\partial x} + \frac{\partial hV}{\partial y} = 0 \tag{2.26}$$

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} + V\frac{\partial U}{\partial y} + g\frac{\partial \zeta}{\partial x} = S_x - f_b U \tag{2.27}$$

$$\frac{\partial V}{\partial t} + U\frac{\partial V}{\partial x} + V\frac{\partial V}{\partial y} + g\frac{\partial \zeta}{\partial y} = S_y - f_b V \tag{2.28}$$

and the scalar transport equation is (neglecting the surface and bottom fluxes, i.e. for a conservative, passive, species) :

$$\frac{\partial \phi}{\partial t} + U\frac{\partial \phi}{\partial x} + V\frac{\partial \phi}{\partial y} = S_{xy} \tag{2.29}$$

In the above equations, we have dropped, for the sake of clarity, the overbars which denote depth-averaged quantities.
Terms $-f_b U$ and $-f_b V$ represent momentum sink by bed friction. f_b refers to the friction factor we introduced in section 2.1.3 (eq. 2.13 to 2.15).
S_x, S_y and S_{xy} are the dispersive and turbulent terms, whose expression is of the form :

$$\begin{aligned} S_x = \frac{1}{h}\Bigg[& \frac{\partial}{\partial x}\left(hD_{xx}\frac{\partial U}{\partial x}\right) + \frac{\partial}{\partial y}\left(hD_{yy}\frac{\partial U}{\partial y}\right) \\ & + \frac{\partial}{\partial x}\left(hD_{xy}\frac{\partial U}{\partial y}\right) + \frac{\partial}{\partial y}\left(hD_{yx}\frac{\partial U}{\partial x}\right)\Bigg] \end{aligned} \tag{2.30}$$

D_{xx}, D_{yy}, D_{xy}, D_{yx} ($\mathrm{m}^2.\mathrm{s}^{-1}$) are the dispersion coefficients for the momentum equations, Γ_{xx}, Γ_{yy}, Γ_{xy}, Γ_{yx} their counterparts in the scalar transport equation (named here the advection-dispersion equation). Although these coefficients encompass much more than turbulent contributions, it is often assumed that the Γ and D are linked by the same kind of relationship which

binds together eddy diffusivity and viscosity, namely :

$$\Gamma = \frac{D}{\sigma_t} \tag{2.31}$$

where σ_t is the turbulent Prandtl or Schmidt number, for which the values usually adopted are either 0.5 or 1.

Empirical formulae commonly in use are suited to forecast the dispersion coefficients in the streamwise direction ($D_{\|}$, $\Gamma_{\|}$) and in the orthogonal direction ($D_{\perp}$, $\Gamma_{\perp}$). Yet, the coordinate directions are not necessarily lined up with the streamlines. If θ is the angle (positive counter-clockwise) of the local streamline (parallel to the local velocity) with respect to the x axis, we have the following relationships :

$$D_{xx} = D_{\|} \cos^2\theta + D_{\perp} \sin^2\theta \tag{2.32}$$

$$D_{yy} = D_{\|} \sin^2\theta + D_{\perp} \cos^2\theta \tag{2.33}$$

$$D_{xy} = D_{yx} = \left(D_{\|} - D_{\perp}\right) \sin\theta \cos\theta \tag{2.34}$$

and similar formula relating the Γ to $\Gamma_{\|}$ and $\Gamma_{\perp}$.

In this work, we have been following the usual path by choosing the two-dimensional St-Venant equations as the core of our surface flow and transport model. Yet, in the forthcoming applications, we shall try to bear in mind the numerous simplifications which led to this choice and try to check as thoroughly as possible whether they are always acceptable.

2.5 Résumé français : "Écoulements à surface libre et transport dissous"

L'emploi de modèles mathématiques et numériques pour résoudre différents problèmes d'hydraulique s'est largement développé dans les trente dernières années. Ces outils s'avèrent être d'indispensables compléments aux mesures sur le terrain ou aux expérimentations sur modèle physique. Ils permettent en effet d'interpréter les données recueillies, voire d'en optimiser la collecte, et surtout possèdent un potentiel certain d'extrapolation. Compte tenu de la puissance sans cesse croissante et des coûts sans cesse décroissants des ordinateurs, l'utilisation de modèles assure entre autres une réduction notable des frais d'étude.

En hydraulique de surface, la variété des cas modélisés est grande. Nous les regrouperons schématiquement dans les deux catégories suivantes :

- En **ingéniérie hydraulique**, on s'intéresse aux écoulements dans des géométries artificielles, sur/autour de structures construites par l'homme. Par exemple, on traite les écoulements au voisinage de jetées, de piles de pont, dans des bassins de stockage/sédimentation, des ports; on calcule le panache en champ proche d'effluents issus de centrales thermiques ou d'usines d'épuration dans le but d'optimiser la configuration des diffuseurs, etc ...
- Dans le cadre des problèmes d'environnement (**hydraulique environnementale**), on se préoccupe d'écoulements dans le milieu naturel : rivières, lacs, estuaires, régions côtières, océans ...

Dans la seconde catégorie, les domaines d'étude considérés sont souvent nettement plus grands que dans la première. Les échelles de temps et d'espace caractérisant les phénomènes hydrauliques et leurs fluctuations significatives peuvent de ce fait être largement différentes.
Par ailleurs, pour les problèmes d'environnement, le calcul des écoulements constitue rarement une fin en soi. On se préoccupe plus généralement du devenir de substances chimiques et le processus de transport/dilution n'est qu'un des facteurs qui gouvernent leur sort. Par conséquent, la précision avec laquelle on prend en compte écoulement et transport doit être en premier lieu cohérente avec le degré de détail retenu pour simuler les processus biogéochimiques.
L'approche modélisatrice doit également être cohérente au regard de la qualité des données disponibles : certaines données relatives au milieu naturel sont moins accessibles, ou moins fiables, que les données concernant un problème d'ingéniérie, par exemple tout ce qui concerne la bathymétrie, la géométrie des domaines.
Enfin, les sollicitations auxquelles sont soumis les systèmes étudiés sont diverses. Certains écoulements devront être décrits de façon tridimensionnelle, d'autres seront caractérisés correctement par des grandeurs moyennées sur la verticale, voire sur la section mouillée; dans certains cas une

approche stationnaire suffira, pour d'autres il faudra absolument saisir l'instationnarité ...

La grande variété des problèmes rencontrés explique le foisonnement des modèles développés. En effet, les équations de Navier-Stokes, équations "universelles" régissant les fluides, ne peuvent actuellement être résolues à un coût raisonnable. L'art du modélisateur consiste donc tout d'abord à dériver de ces équations d'autres équations plus aisément manipulables. Ceci suppose d'intégrer les phénomènes en temps et en espace, voire d'en négliger certains. Dans ce processus d'intégration, on introduit des termes inconnus, qui traduisent l'effet sur la conservation de la masse et/ou de l'énergie des phénomènes que l'on s'est refusé à décrire car ils se produisent à une échelle plus fine que notre grille d'observation. Quantifier ces termes, de façon adéquate par rapport au problème traité, pose ce qu'on appelle un problème de fermeture des équations. Selon la méthode retenue on débouche sur des modèles différents.

Dans ce chapitre, **nous commencerons par introduire brièvement les équations de Navier-Stokes** (section 2.1). Nous les intégrerons progressivement, dans le temps puis dans l'espace, en équations tri- puis bidimensionnelles. Ce faisant, nous introduisons les termes inconnus mentionnés ci-dessus. Nous indiquerons succintement quelle est leur signification physique (turbulence, dispersion horizontale, dissipation d'énergie par frottement ...).

Le problème de fermeture des équations est abordé dans les deux sections suivantes, la section (2.2) consacrée à la turbulence, la suivante (2.3) à la dispersion. En ce qui concerne la **modélisation de la turbulence**, nous nous sommes limités à mentionner les approches les plus couramment employées en hydraulique à surface libre. Dans ce domaine d'application, on a tendance à emprunter, adapter les modèles de fermeture développés dans d'autres branches de la mécanique des fluides plutôt qu'à innover. C'est pourquoi dans leur grande majorité les modèles utilisés sont basés sur l'approche classique, statistique, de la turbulence (cf 2.2.2). Cette approche repose sur la décomposition des variables hydrauliques en une partie moyenne et une partie fluctuante. Compte tenu de la présence des termes advectifs non-linéaires dans les équations de Navier-Stokes, le moyennage introduit des corrélations inconnues (moyennes des produits de fluctuations) qui composent ce que l'on dénomme le tenseur de Reynolds.

1. Dans les modèles basés sur le concept de **viscosité turbulente**, les contraintes de Reynolds sont exprimées en fonction des gradients du champ de vitesse moyen si bien que les termes turbulents se retrouvent représentés par un opérateur de dispersion, dont l'importance est contrôlée par la viscosité turbulente. La viscosité turbulente est une propriété locale de l'écoulement, contrairement à la viscosité moléculaire qui est une propriété intrinsèque du fluide. Elle est supposée proportionnelle au produit d'une échelle de longueur L et d'une échelle de vitesse $\widehat{V}$ caractéristiques. C'est la façon d'exprimer ces deux échelles qui différencie les modèles de cette catégorie.

   - Dans les modèles de longueur de mélange, L et $\widehat{V}$ sont calculées directement à par-

tir du champ moyen, ou de caractéristiques géométriques (distance à la paroi, par exemple).

- Dans les modèles à 1 équation, on calcule $\widehat{V}$ à l'aide d'une équation de transport dérivée des équations de Navier-Stokes.
- Dans les modèles à 2 équations, dont le populaire $k-\epsilon$, on emploie également une équation de transport pour estimer L.

2. Les modélisations basées sur le concept de viscosité turbulente sont souvent qualifiées de trop simplistes, notamment à cause de l'hypothèse d'isotropie de la turbulence. Pour pallier cette insuffisance, on a proposé le développement d'équations de transport pour *chaque* composante du tenseur de Reynolds, dans le cadre des **modèles des tensions de Reynolds**, aussi appelés **modèles du second ordre**.

Une revue des applications aux problèmes d'hydraulique fluviale (cf sec. 2.2.3) (dilution de panaches en champ proche, calcul de courants secondaires) laisse à penser qu'une modélisation de type $k-\epsilon$ fournit souvent des résultats raisonnables, la supériorité de modèles plus sophistiqués restant à prouver.

La **dispersion** est un phénomène d'origine advective : en effet, les termes dispersifs représentent la contribution au transport de quantité de mouvement ou de masse due aux déviations par rapport au mouvement moyen calculé (cf sec. 2.3.1). Ainsi, dans un modèle bidimensionnel plan, la dispersion correspond théoriquement à l'effet des non-uniformités de la vitesse sur une colonne d'eau. Cependant, comme on passe par une étape de discrétisation avant la résolution, la dispersion recouvrira également, de fait, l'effet des circulations sous-maille dans le plan ! On peut donc s'attendre à ce qu'elle soit partiellement dépendante de la finesse du maillage.

La dispersion est fréquemment modélisée par un opérateur de diffusion. Le cadre dans lequel cette modélisation a été rigoureusement justifiée est restreint par rapport à ses nombreuses applications. La coutume perdure compte tenu de ses succès pratiques !

De nombreuses expériences indiquent que, dans un cadre bidimensionnel plan, en rivière, les coefficients de dispersion longitudinale (i.e. parallèlement à l'écoulement) et transverse sont corrélés au produit de la vitesse de frottement au fond et de la hauteur d'eau (cf sec 2.3.2), le coefficient de proportionnalité dépendant de la géométrie du cours d'eau (bief rectiligne ou méandré).

Finalement, nous passons en revue les **considérations pratiques** qui influencent le choix d'un modèle, notamment la qualité des données disponibles pour l'implanter, et le type d'utilisations envisagées (cf sec 2.4). Nous résumons en particulier les principales **caractéristiques du problème des rejets pluviaux urbains en Seine**, premier cadre d'application pour les modèles

que nous allons développer (sec 2.4.4).
Les rejets pluviaux induisent des perturbations hydrauliques trés importantes en Seine (doublement du débit de base en cas de forts orages). Compte tenu de la configuration des déversoirs d'orage et de la faible profondeur de la Seine, on peut estimer que l'homogénéisation des panaches sur la verticale est rapide. Par contre, la distance de bon mélange en travers de la rivière est, selon différentes expériences de traçage, de l'ordre de 5 à 10 km. Le caractère transitoire des écoulements, la densité et proximité des points de rejet, l'hétérogénéité transverse, rendent difficile à interpréter les mesures de qualité de l'eau (oxygène, micropolluants) effectuées pour suivre ces rejets. La quantification des phénomènes biogéochimiques exige une évaluation correcte des phénomènes de transport et dilution. Pour ce faire, le minimum requis semble être un modèle bidimensionnel plan transitoire.

Si la bathymétrie de la Seine est bien connue (c'est une rivière navigable dont le chenal est de ce fait bien surveillé), il n'en est pas de même de son hydraulique. Si les débits sont accessibles avec une bonne précision et un pas de mesure réduit (toutes les 2 heures), grâce à des stations de mesure à ultrasons, les lignes d'eau sont relevées approximativement, en des points trés espacés (les barrages de navigation, distants en moyenne d'une trentaine de kilomètres) et peu régulièrement (2 à 3 fois par jour et en cas de manoeuvre des barrages). On ne dispose pas non plus de profils de vitesse caractéristiques. Enfin, compte tenu de la taille des biefs que l'on souhaite modéliser (une dizaine à une quarantaine de kilomètres), l'adoption d'un maillage trés fin n'est pas souhaitable. Dans ces conditions, il semble assez illusoire d'appliquer un modèle comprenant une description détaillée du phénomène de turbulence. **On se contentera donc de choisir comme base de notre modèle d'écoulement les équations bidimensionnelles de St-Venant** (section 2.4.5). **Le transport dissous sera décrit par l'équation bidimensionnelle d'advection-diffusion.**

La suite de ce travail sera consacrée à l'étude de méthodes de résolution fiables de ces équations.

Chapter 3

Introduction to computational methods

Flow and transport models differ not only by their equations, i.e. by the way each relevant phenomenon is represented, but also by the numerical techniques applied to solve these equations.

Fluid dynamics are described by a set of partial differential equations (PDE) involving space and time derivatives of the dependent variables. The purpose of numerical methods is to approximate them, as closely as possible, by a system of discretized equations. The unknowns are the values of the variables (velocities, depth, concentrations) at a finite number of fixed locations (nodes) in the studied domain.

The objective of this chapter is certainly not to deliver a detailed discussion about the merits of one or another scheme (which will in fact be tackled in parts II and III of this report). We shall merely try to give a brief survey of the methodologies available, first to build up the system of discretized equations, then to solve it. We also introduce the technical vocabulary relative to the development of numerical schemes and the study of their behaviour.

3.1 A common framework : Weighted Residual Methods

Albeit finite difference methods have been originally developed following a quite different line of thought than finite element or finite volume methods, these different approaches applied in the field of computational fluid dynamics can now be introduced within the same framework, as subclasses of the general weighted residual methods (WRM), as suggested by (Fletcher, 1991), chap. 5 .

Let us consider a PDE, written in the general form :

$$\mathcal{L}(f) = 0 \tag{3.1}$$

where $\mathcal{L}$ denotes a space and time differential operator. As the starting point of WRM, we find the assumption that the PDE solution can be represented analytically :

$$f(\vec{x}, t) = \sum_{1}^{J} a_j(t)\, \phi_j(\vec{x}) \tag{3.2}$$

The analytical functions ϕ_j are often termed *trial functions*. We note that, with the help of eq. 3.2, WRM allow us to assess the solution value at any point of the computational domain $\mathcal{D}$. Of course, unless J is made arbitrarily large, forcing the solution of the PDE to obey 3.2 introduces some error. Consequently, while an exact solution f^* satisfies indeed eq. 3.1, the approximate solution 3.2 verifies only

$$\mathcal{L}(f) = R \tag{3.3}$$

where R is referred to as the *equation residual*. The objective is then to make R as "small" as possible on the computational domain.

How do we measure the "size" of a function and its smallness ? This is usually done by introducing the L_2 norm on the space $\mathcal{F}$ of integrable functions on $\mathcal{D}$ ($\mathcal{D}$ is a subspace of $\Re$, $\Re^2$ or $\Re^3$ according to the dimension of the problem) :

$$\|f\| = \sqrt{\int_{\mathcal{D}} f^2 \, d\Omega}$$

Similarly, the scalar product of two functions is defined by :

$$< f, g > = \int_{\mathcal{D}} f g \, d\Omega$$

The fact that a function R is null can be expressed alternatively by $\|R\| = 0$ or,

$$\forall g \in \mathcal{F}, \quad < R, g > = 0 \tag{3.4}$$

R is compelled to be small by satisfying 3.4, not for any g, but for a *representative subset* of $\mathcal{F}$, i.e the *weighting (test) functions*, W_m, $m = 1, M$.

There are different possible choices for the W_m, for instance :

1. **Collocation method**

$$W_m(\vec{x}) = \delta(\vec{x} - \vec{x}_m) \tag{3.5}$$

where δ denotes the Dirac function (1 if $\vec{x} = \vec{x}_m$, zero elsewhere). This choice implies that R is zero and eq. 3.1 satisfied for a set of points over the domain $\mathcal{D}$.

Finite difference methods (FDM) are typically collocation methods, introduced without referring to the approximation 3.2. FDM usually discretize the computational domain with the help of an orthogonal grid. The collocation points correspond to the vertices (nodes) of this grid. The PDE at each node is transformed into an algebraic equation as the partial derivatives are replaced by finite differences which involve the value of the dependent variables f at the node and at its neighbours.

2. **Subdomain method**

The domain $\mathcal{D}$ is divided in sub-regions $\mathcal{D}_m$, which may eventually overlap, and whose juxtaposition covers the entire domain. Then, W_m are defined so that :

$$W_m(\vec{x}) \neq 0 \text{ (frequently } W_m(\vec{x}) = 1) \text{ if } \vec{x} \in \mathcal{D}_m, \quad W_m(\vec{x}) = 0 \text{ if not} \tag{3.6}$$

Consequently, the PDE is satisfied over small sub-regions rather than at definite locations.

The PDE accounts for a physical law. Should this law correspond to a conservation statement (a frequent case in fluid dynamics), the subdomain method amounts to applying the conservation principle to a *finite* surface or volume. When we take the limit of shrinking this *control volume* to a point we obtain the PDE again. **Finite volume methods** (FVM) are subdomain methods where the explicit introduction of an approximate solution like (3.2) is dropped.

When completely developed on an orthogonal grid, finite difference and finite volume representations can be quite similar. However, whereas FDM start with difference approximations for individual derivatives and make up the representation for the complete PDE by adding them, FVM work out a discrete statement of the *complete* physical conservation principle, i.e. of the *entire* PDE : consequently, they appear unable to provide an *independent* algebraic representation of just a derivative alone.

We shall not discussed further on FVM. However, we would like to point out the following attractive features of such methods :

- FVM provide naturally conservative schemes.
- Secondly, as illustrated for instance in (Fletcher, 1991) chap. 5.2, they allow one to discretize complicated computational domains with curvilinear surfaces or volumes without relying on a complex change of coordinates, as is needed for finite difference or element methods.
- Lastly, they allow one to express the conservation principle even when the studied problem involves shock, discontinuities : *continuous* differential equations fail locally to describe this kind of phenomena. Unfortunately, continuous PDE and related solutions are necessary for applying the formalism of finite difference or element methods.

3. **Galerkin method**

   The Galerkin method amounts to taking the weighting functions from the same family as the trial functions :

$$W_m(\vec{x}) = \phi_m(\vec{x}) \tag{3.7}$$

   Most **finite elements methods** follow the Galerkin approach. The ϕ_m are chosen so that *(i)* they are defined on *limited* sub-regions of $\mathcal{D}$ (and zero elsewhere); *(ii)* they have simple analytical forms (low-order polynomials); *(iii)* they are linearly independent. Consequently, they generate a vectorial sub-space $\mathcal{F}_h$ of $\mathcal{F}$. We replace the problem of solving eq. 3.1 on $\mathcal{F}$ by the approximate problem of solving it on $\mathcal{F}_h$. Given a clever definition of the ϕ_m, an increase of their number (i.e. a corresponding increase in spatial resolution), the solution on $\mathcal{F}_h$ is a fair enough approximation of the "true" solution.

   The above introduction to the FEM is quite brief from a mathematical viewpoint. However, FEM methods can be placed on a sound mathematical foundation by relating them to the problem of minimizing some quadratic form (linked for instance to the energy of the studied system) and by giving them a variational interpretation. Such formalism applies straightforwardly to elliptic partial differential equations and has then been extended to hyperbolic ones. The interested reader should refer to specialized textbooks (Raviart, 1981; Goussebaile *et al.*, 1986).

The following two sections give more details of the development of FDM and FEM.

3.2 Basics of Finite Differences Methods (FDM)

3.2.1 Discretization

FDM have been first relying on the discretization of the computational domain with the help of an orthogonal grid. The figures 3.1 and 3.2 give examples of orthogonal computational grids,

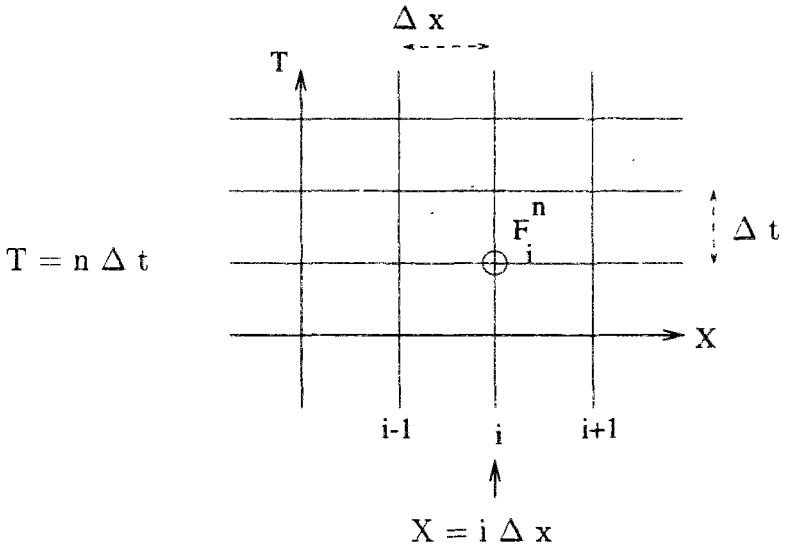


Figure 3.1: One-dimensional finite difference grid

respectively in one-dimensional and two-dimensional situations. The associated notations are :

$$f_i^n = f(x_i, t_n) = f(x_0 + i\Delta x, t_0 + n\Delta t)$$
$$f_{i,j}^n = f(x_i, y_j, t_n) = f(x_0 + i\Delta x, y_0 + j\Delta y, t_0 + n\Delta t)$$

where Δt is the time step, Δx (Δy) the space step along the $x-$ ($y-$) direction. x_0 and (x_0, y_0) denote the origin of the space discretization, respectively in one and two-dimensional situations. t_0 is the starting (initial) time of the computation. Points defined by $x = x_i$ or $\vec{x} = (x_i, y_j)$ are *nodes* of the computational grid.

In both examples, the mesh size is uniform. However, the space steps $x_i - x_{i-1}$ and $y_j - y_{j-1}$ can be taken to be variable. This allows some grid refinement. Yet, $x_i - x_{i-1}$ cannot be reduced in a limited area only, independently of the remainder of the studied computational domain. If only local refinement is desired, it can be achieved through the nested grid technique : one defines first a "coarse" grid which covers the whole computational domain, then a "fine" grid wich covers a limited rectangular sub-region. The PDE is solved first on the coarse grid : this provides boundary conditions for the resolution on the interior fine grid.

The computation of flow fields in and around complex shapes involves computational boundaries which are but crudely approximated by an orthogonal grid, unless the mesh size is drastically reduced. The imposition of boundary conditions in such problems typically implies a local loss of accuracy, whose influence on interior nodes can eventually decay but slowly ... or worse,

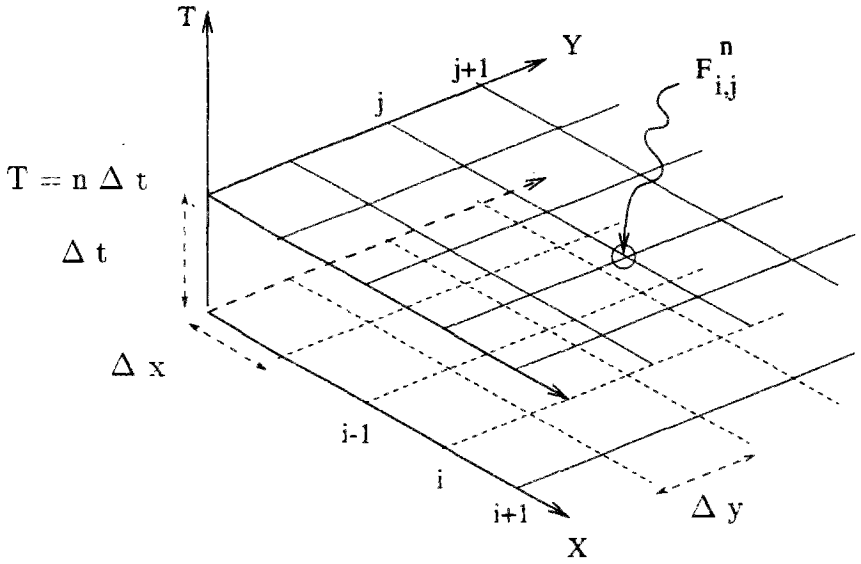


Figure 3.2: Two-dimensional finite difference grid

be amplified !

In order to extend the effectiveness of FDM, it is possible (and common) to use generalised curvilinear coordinates (Fletcher, 1991) (chap.12). More precisely, we introduce a transformation (also termed "mapping") of the physical space (x, y, z) into some generalised-coordinate space (ζ, ν, ξ) such that the computational boundaries in the physical space coincide with coordinate lines in the generalised-coordinate space. The governing equations are expressed in terms of the new coordinates as independent variables. Discretization and computation are then performed in the transform of the computational domain.

The modified equations contain additional terms (Fletcher, 1991) (chap.12), first-order coordinate derivatives such as $\frac{\partial x}{\partial \zeta}$, their cross-products and even higher-order coordinate derivatives. Discretization of such terms represents an additional source of error in the numerical resolution.

A desirable feature of the curvilinear grid is that it should be well-conditioned, i.e. smoothly varying, close to orthogonal and with local grid aspect ratios (terms such as $(x_{i+1} - x_i)/(x_i - x_{i-1})$) close to unity. For such mappings, some of the additional terms eventually vanish.

When generating a grid, the values of the generalised coordinates (ζ, ν, ξ) are fixed on the boundaries first. Then, computing a function $\zeta(x, y, z)$ all over the computational domain can be interpreted as a boundary value problem : desirable properties of the mapping (e.g. zero values of some additional terms in the equations) are expressed as partial differential equations that the (ζ, ν, ξ) must satisfy. An alternative approach is referred to as algebraic mapping. The boundary nodes are "intelligently" distributed along the boundaries with the help of "stretching functions"

(Fletcher, 1991) (chap. 13), which allow grid refinement in regions where high gradients of the dependent flow variables are expected. Then, one uses explicit, algebraic, interpolation formulae between the boundary nodes to determine the interior grid nodes. Grid generation techniques are reviewed extensively in such books as (Roache, 1985) (chap. 6), (Anderson *et al.*, 1984) (chap. 10) and (Fletcher, 1991) (chap. 13).

3.2.2 Differentiation

Several procedures are available for developing finite difference approximations of the space and time derivatives. The most common are :

- **combination of Taylor series expansions**
  Taylor series are written for each dependent variable at each grid node. By weighting them appropriately, it is possible to express any derivative with some truncation error which is a power of the time or space steps. The exponent of this power law defines the *order* of the finite difference approximations. For instance,

$$\begin{aligned} f(x_i,\, t_n+\delta) &= f_i^n + \delta \frac{\partial f}{\partial t}^n_i + \frac{\delta^2}{2}\frac{\partial^2 f}{\partial t^2}^n_i + O\left(\delta^3\right) \\ f(x_i+\delta,\, t_n) &= f_i^n + \delta \frac{\partial f}{\partial x}^n_i + \frac{\delta^2}{2}\frac{\partial^2 f}{\partial x^2}^n_i + \frac{\delta^3}{6}\frac{\partial^3 f}{\partial x^3}^n_i + O\left(\delta^4\right) \end{aligned}$$

  so that, equating δ either with the time step Δt or the space step Δx (and its opposite), we can obtain :

$$\frac{\partial f}{\partial t}(x_i,\, t_n) = \frac{f_i^{n+1} - f_i^n}{\Delta t} + O\left(\Delta t\right) \tag{3.8}$$

$$\frac{\partial f}{\partial x}(x_i,\, t_n) = \frac{f_{i+1}^n - f_{i-1}^n}{2\Delta x} + O\left(\Delta x^2\right) \tag{3.9}$$

$$\frac{\partial^2 f}{\partial x^2}(x_i,\, t_n) = \frac{f_{i+1}^n - 2f_i^n + f_{i-1}^n}{\Delta x^2} + O\left(\Delta x^2\right) \tag{3.10}$$

  $O\left(\delta^m\right)$ refers to a polynomial expression which lower-order term is proportional to δ^m . (nb : Eq. 3.8 is said to correspond to *forward time differencing*, eq. 3.9 and 3.10 to *centred space differencing*)

  The use of uneven space steps complicates the Taylor series expansion. For instance, if one wants to obtain once again a second-order approximation of $\frac{\partial f}{\partial x}$ for $\Delta x_1 = x_i - x_{i-1} \neq \Delta x_2 = x_{i+1} - x_i$, relation 3.9 has to be replaced by :

$$\frac{\partial f}{\partial x}(x_i,\, t_n) = \frac{1}{\Delta x_1 + \Delta x_2}\left[\frac{\Delta x_1}{\Delta x_2}\left(f_{i+1} - f_i\right) + \frac{\Delta x_2}{\Delta x_1}\left(f_i - f_{i-1}\right)\right] \tag{3.11}$$

  By combining the truncation errors related to each derivative approximation, one eventually assesses the whole truncation error (TE) with which the PDE is approximated by its finite difference representation.

It is worth noting that a Taylor series expansion assumes the dependent variables to be continuous and differentiable functions.

- **polynomial fitting**

  It is assumed that the solution to the PDE can be approximated locally, i.e. in the vicinity of each node, by a polynomial. The coefficients of the polynomial are evaluated by fitting it to the neighbouring nodes. Once the polynomial is defined, it can be derived straightforwardly as any analytical function. Beyond the second-order polynomial, the expressions obtained (for the first and second order space derivatives) are not identical to those from corresponding order Taylor series expansion.

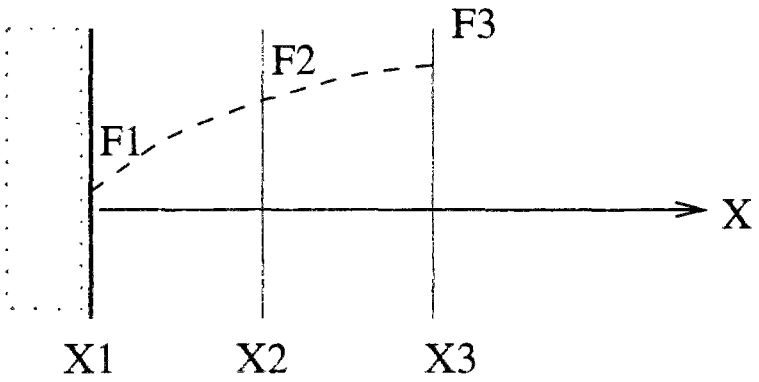


Figure 3.3: Quadratic fitting in boundary vicinity

While it can be used to derive the entire finite difference representation, polynomial fitting is most commonly employed in the treatment of boundary conditions. For instance, let us consider figure 3.3. Assuming that function f is approximated by a second-order polynomial $\mathcal{P}_2$ over the interval $[x_1, x_3]$ (and that, for the sake of simplicity, $x_3 - x_2 = x_2 - x_1 = \Delta x$), we obtain that :

$$\frac{\partial f}{\partial x}(x_1) = \frac{-f_3 + 4f_2 - 3f_1}{2\Delta x} + O\left(\Delta x^2\right) \tag{3.12}$$

Such a relation could help us to express a boundary condition on the space derivative in terms of interior grid points. When applying this method, it is particularly important to ensure that the hypothesis made in deriving the boundary derivatives is consistent with the differentiation method applied at interior nodes (Roache, 1985) (sec. III.C). In the above example, we could check for instance that deriving $\mathcal{P}_2$ at node 2 yields an approximation to $\frac{\partial f}{\partial x}(x_2)$ which is equal to the approximation presented in eq. 3.9. Consequently, relation 3.12 is consistent with the application of centred space differencing at interior nodes. This will not be the case for other interior node methods.

A last method is the **integral method** which relies on local integration of the PDE and computation of the subsequent integrals with the help of the *mean-value theorem.* It is much less widely applied. The interested reader is referred to textbooks by (Roache, 1985; Anderson *et al.*, 1984).

3.2.3 Equivalent linear system of equations

Let us consider first a one-dimensional situation. $\vec{f}$ denotes the vector whose components are the nodal values f_i ($1 \leq i \leq \mathcal{N}$). From eq. 3.8, we have simply :

$$\frac{\partial \vec{f}}{\partial t} = \frac{1}{\Delta t}\left(\vec{f}^{n+1} - \vec{f}^{n}\right) = \frac{1}{\Delta t}\underline{I}_{\mathcal{N}}\left(\vec{f}^{n+1} - \vec{f}^{n}\right)$$

where $\underline{I}_{\mathcal{N}}$ is the identity matrix. It can be shown that, with the help of finite difference approximations, spatial differentiation reduces also to a linear operation on vector $\vec{f}$. Combining expressions 3.9 and 3.12, we obtain for instance :

$$\frac{\partial \vec{f}^{n}}{\partial x} = \underline{A}_x \vec{f}^{n} \tag{3.13}$$

where $\underline{A}_x$ is a square matrix (dimension $\mathcal{N}$) whose expression reads :

$$\underline{A}_x = \frac{1}{2\Delta x}\begin{bmatrix} -3 & 4 & -1 & & & 0 \\ -1 & 0 & 1 & & & \\ & -1 & 0 & 1 & & \\ & & & . & . & . \\ & & & -1 & 0 & 1 \\ 0 & & & 1 & -4 & 3 \end{bmatrix}$$

Expressing other derivatives in the same way, it is straightforward to check that, if the PDE is a linear one, it can be replaced by a linear equation whose dependent variable is the vector $\vec{f}$.

When one aims at computing new values of the function f at time $T + \Delta t$, knowing all dependent variables at times $t \leq T$, one can either develop the spatial derivatives as a function of past and present values of f only, either as a function of advance values of f too (i.e. of f at time $t = T + \Delta t$). We would write for instance, instead of eq. 3.13

$$\frac{\partial \vec{f}^{n}}{\partial x} = \gamma \underline{A}_x \vec{f}^{n+1} + (1-\gamma)\underline{A}_x \vec{f}^{n} \quad (\ 0 \leq \gamma \leq 1\ ) \tag{3.14}$$

In the first case, the value of f at $T + \Delta t$ at a grid node turns out to depend merely on past values of f and can be computed independently, individually. In the second case, the assessment of $f\,(T + \Delta t)$ requires the *simultaneous* solutions of finite difference equations at several grid

nodes. The first kind of schemes are called **explicit** ones, the second are **implicit** ones. Explicit schemes typically reduce to systems of the form

$$\underline{\vec{f}}^{n+1} = \underline{B}\,\underline{\vec{f}}^{n} \tag{3.15}$$

while implicit schemes lead to a matrix inversion problem :

$$\underline{A}\,\underline{\vec{f}}^{n+1} = \underline{B}\,\underline{\vec{f}}^{n} \tag{3.16}$$

The linear systems arising from finite difference representations are usually sparse ones : as the finite differentiation involves only the neighbours of a node, most of the components of the matrix are zero. In one-dimensional situations, the matrices involved are generally tridiagonal or pentadiagonal. Unless some space splitting is applied (see section 3.6), the matrix band width is more important in multidimensional situations since, according to the numbering of the nodes, such neighbouring nodes as (i, j) and $(i, j+1)$, for instance, may rank a few lines apart in vector $\vec{f}$. In the case of multivariable problems, the dimension and complexity of $\vec{f}$ are further increased, as $\vec{f}$ must encompass the nodal values of each dependent variable.

PDEs are not always linear. However, without loss of generality, we can assume that most equations encountered in fluid mechanics and heat or mass transfer can be written in the form

$$\frac{\partial \vec{F}}{\partial t} + \frac{\partial \vec{H}}{\partial \vec{x}} = 0 \tag{3.17}$$

where $\vec{H}$ is a function of the vector of variables $\vec{F}$. Introducing the Jacobian matrix

$$J = \frac{\partial \vec{H}}{\partial \vec{F}}$$

we can locally linearize the system 3.17 by holding J constant while the vector $\vec{F}$ is advanced through a single time step : in the vicinity of each gride node M, 3.17 is approximated by a linear PDE :

$$\frac{\partial \vec{F}}{\partial t} + J_{\mathrm{M}} \frac{\partial \vec{F}}{\partial \vec{x}} = 0 \tag{3.18}$$

Then, finite differentiation can be applied to eq. 3.18.

3.3 Basics of Finite Element Methods (FEM)

3.3.1 Discretization

To implement an FEM, one needs first to discretize the computational domain into sub-areas, termed *elements*. The elements have a simple geometrical shape : triangle or quadrangle in two

dimensions, tetrahedron, prism or parallelepiped in three dimensions. They are defined so that the intersection between two elements is either void, limited to a point (element summit) or limited to a common side or face (respectively in two and three dimensions). The size of the elements can be space varying, allowing easily local grid refinements, notably in flow regions where hydraulic variables display strong gradients. A detailed presentation of the main families of finite elements (the P family associated with triangles and tetrahedrons, the Q-family related to the quadrangles and parallelepipeds) can be found in (Raviart, 1981; Goussebaile *et al.*, 1986).

As in the case of finite difference discretization (cf section 3.2.1) it is possible to introduce distorted rectangular or triangular elements with the help of appropriate curvilinear mappings (mostly isoparametric transformations, see for instance (Goussebaile *et al.*, 1986), chap. 3 , (Fletcher, 1991) section 5.5.3). This extends furthermore the ability of FEM to handle complicated computational domains.

3.3.2 Choice of shape and test functions

One then has to define the trial functions ϕ_m which are referred to as **shape functions** in the FEM vocabulary. This is done abiding by the following rules :

1. Each ϕ_m is zero except for a restricted area of the computational domain (its "definition" domain), corresponding either to one element only or to a small number of contiguous elements.

2. The ϕ_m are built by introducing into each element a finite number of points, called *nodes*. The ϕ_m are piecewise polynomial functions uniquely defined by their values at the nodes. Consequently, the order of the ϕ_m depends on the number of available nodes.

   The ϕ_m are chosen such that their values are 1 at a specific node and zero at all other nodes.

   The ϕ_m are continuous functions. This property is not necessarily satisfied by their derivatives.

With such shape functions, relation 3.2 becomes

$$f(\vec{x}, t) = \sum_{1}^{J} f_j \phi_j(\vec{x}) \tag{3.19}$$

where J denotes now the total number of nodes and f_j the value of f at node number j.

Figure 3.4 gives two examples of shape functions defined on rectangular elements. In the first case, the nodes correspond merely to summits of the elements (4 nodes/element) : the resulting shape function is bilinear. In the second case, there are additional nodes which correspond to

midside points and to the element centre (9 nodes/element) : it allows the shape function to become biquadratic. (NB : the term *primary nodes* usually denote nodal points corresponding to an element summit, *secondary nodes* are the other ones).

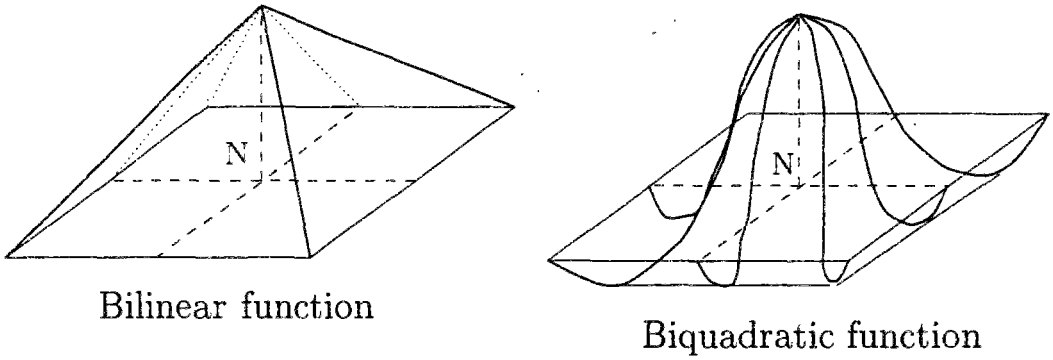


Figure 3.4: Examples of 2D shape functions

At the beginning, in the early seventies, FEMs followed the Galerkin approach, namely they used the same shape and **test** functions (cf section 3.1). However, these methods displayed major drawbacks when applied to convection-dominated transport problems. Indeed, they lead naturally to "centred" approximation of the advective terms, such as centred FDMs using approximation 3.9, and suffer the same kind of problems : inability to deal with pure advection, appearance of oscillations unless some *very* restrictive condition on the space step size is enforced (Roache, 1985; Baptista *et al.*, 1984). In the late seventies, several attempts were made to take into account the physics behind advective terms. In these Petrov-Galerkin methods, the weighting functions are not equal to the trial functions but they are obtained from them by a change in shape that increases the relative weight of upstream information in a way that depends on the element geometry and the flow characteristics (Brooks & Hughes, 1982; Donea *et al.*, 1987; Noorishad *et al.*, 1992; Miller & Rabideau, 1993), (Goussebaile *et al.*, 1986) (section II.3.1). More recently, more advanced methods have been proposed (Hervouet, 1984), (Goussebaile *et al.*, 1986) (section IV) where the test functions are obtained from the shape functions by solving a transport equation with the help of the characteristics method. A last alternative is to use splitting techniques (cf sec. 3.6) and to apply the FEM formalism only to non-advective terms of the equations (Baptista *et al.*, 1984; Hervouet & Watrin, 1988).

3.3.3 Construction of the equivalent linear system

We shall deal hereafter with the general procedure for interior nodes. For the sake of simplicity, the following steps of the FEM derivation will be illustrated by assuming that the operator $\mathcal{L}$ is linear, that we deal with a steady-state problem and that shape and test functions are the

same. As $\mathcal{L}$ is linear, we have :

$$R = \mathcal{L}\left(\sum_1^J f_j\phi_j\right) = \sum_1^J f_j\mathcal{L}(\phi_j)$$

so that,

$$\int_{\mathcal{D}} \phi_m R\, d\Omega = 0 \iff \sum_1^J f_j \int_{\mathcal{D}} \phi_m \mathcal{L}(\phi_j)\, d\Omega = 0 \tag{3.20}$$

Introducing the square matrix $\underline{A}$ (dimension J) whose components are $a_{i,j} = \int_{\mathcal{D}} \phi_i \mathcal{L}(\phi_j)\, d\Omega$ and the vector $\vec{f}$ (dimension J) whose components are the nodal values f_j, we see that the solution of the PDE reduces to the solution of the system :

$$\underline{A}\,\vec{f} = 0 \tag{3.21}$$

Thanks to the choice of the shape functions, most of the $a_{i,j}$ are zero : the intersection of the domains of definition of ϕ_i and ϕ_j is not void only if nodes i and j are neighbours, i.e. belong to the same or contiguous elements. With a clever numbering of the nodes, $\underline{A}$ has a limited band width and can be stored with efficiency. Evaluating the $a_{i,j}$ amounts to computing polynomial functions (products of the ϕ_j and their derivatives) and to evaluating their integrals. This can be done by using approximate formulae (Gauss quadrature formula) or by using specific software which deals with formal algebraic calculus (Hervouet & Watrin, 1988).

When dealing with unsteady problems the usual practice is to discretize only the spatial terms, whereas the time derivatives are approximated by classical finite differences. Let us assume the operator $\mathcal{L}$ is such that $\mathcal{L}(f) = \frac{\partial f}{\partial t} + L_x(f)$ where L_x is a spatial differential operator only :

- $\frac{\partial f}{\partial t}$ at time t_n is approximated by $(f^{n+1} - f^n)/\Delta t$
- We adopt an implicit scheme, so that $L_x(f)$ at time t_n is developed as $L_x(f) = L_x\left(\gamma f^{n+1} + (1-\gamma) f^n\right)$. The weighting factor γ ($\gamma \in [0,1]$) is termed the "implicitation parameter".

Consequently, R develops into :

$$\frac{\partial f}{\partial t} + L_x(f) = \sum_1^J \frac{f_j^{n+1} - f_j^n}{\Delta t}\phi_j + \gamma \sum_1^J f_j^{n+1} L_x(\phi_j) + (1-\gamma)\sum_1^J f_j^n L_x(\phi_j)$$

Introducing the square matrix $\underline{G}$ (dimension J), frequently termed the *mass matrix*, whose components are $g_{i,j} = \int_{\mathcal{D}} \phi_i \phi_j\, d\Omega$, and the vectors $\vec{f}^{n+1}$ and $\vec{f}^n$ whose components are the nodal values of solution f respectively at t_n and $t_{n+1} = t_n + \Delta t$, we obtain now the system :

$$\left(\underline{G} + \gamma\Delta t\,\underline{A}\right)\vec{f}^{n+1} = \left(\underline{G} - (1-\gamma)\Delta t\,\underline{A}\right)\vec{f}^n \tag{3.22}$$

If the operator $\mathcal{L}$ is not linear, the PDE is locally linearized, as in the case of the FDM development (cf section 3.2.3), prior to applying the FEM formalism.

3.3.4 Treatment of boundary conditions

We shall conclude this summary presentation by mentioning another important feature of FEM : FEM naturally allow a flexible discretization of the computational domain and, as a by-product, a close approximation of the physical boundaries; they also lead generally to a more satisfying mathematical treatment of boundary conditions than do FDM.

Let Γ be the boundary of the computational domain $\mathcal{D}$. If boundary conditions are of the Dirichlet type, i.e. $f_{|\Gamma} = g$ there is simply no need to introduce any discretized equation at the boundary nodes : the nodal values derive directly from knowing g.

Boundary conditions can also be of the Von Neumann type, namely $\frac{\partial f}{\partial n}_{|\Gamma} = g$ where n denotes the direction orthogonal to the boundary. We shall illustrate the advantage of a finite element formulation on the example of the heat diffusion equation :

$$\frac{\partial f}{\partial t} - K \Delta f = 0$$

For the sake of simplicity, we assumed the diffusion coefficient K to be uniform. Δ denotes the Laplacian operator, $\Delta f = \mathrm{div} \nabla f$, div and ∇ being respectively the divergence and gradient operators.

Let Γ_m be the boundary of $\mathcal{D}_m$, domain of definition of shape function ϕ_m. Applying the Stokes formula :

$$\int_{\mathcal{D}} u \, \mathrm{div} \vec{v} \, d\Omega = - \int_{\mathcal{D}} \nabla u . \vec{v} \, d\Omega + \int_{\Gamma} u \, \vec{v} . \vec{n} \, ds \tag{3.23}$$

we have

$$\int_{\mathcal{D}} \phi_m \Delta f \, d\Omega = = - \int_{\mathcal{D}_m} \nabla \phi_m . \nabla f \, d\Omega + \int_{\Gamma_m} \phi_m \nabla f . \vec{n} \, ds \tag{3.24}$$

The first integral on the right-hand side of eq. 3.24 is the linear combination of elementary integrals such as $\int_{\mathcal{D}_m} \nabla \phi_m . \nabla \phi_j \, d\Omega$. If node m is an interior node, the whole of D_m belongs to the computational domain and, according to the definition of ϕ_m, ϕ_m is zero on Γ_m so that the second integral disappears. If node m is a boundary node, $\mathcal{D}_m$ denotes in fact the intersection of the definition domain of ϕ_m with the computational domain. Noting that $\frac{\partial f}{\partial n} = \nabla f . \vec{n}$ we have :

$$S_m = \int_{\Gamma_m} \phi_m \nabla f . \vec{n} \, ds = \int_{\Gamma \cap \Gamma_m} \phi_m \frac{\partial f}{\partial n} \, ds = \int_{\Gamma \cap \Gamma_m} \phi_m g \, ds \tag{3.25}$$

If g has an analytical form, the last expression can be straightforwardly computed. If g is given only at nodes, before the integration proceeds, g is approximated along the boundaries with the appropriate shape functions. In both cases, it appears that *such boundary conditions merely*

introduce source terms in the linear discretized equations. This means that the form of systems such as described by eq. 3.22 becomes in fact

$$\underline{M}\,\vec{f}^{n+1} = \underline{M'}\vec{f}^{n} + \vec{S} \tag{3.26}$$

In the example of the heat diffusion problem, we have :

$$\begin{aligned} m_{i,j} &= \int_{\mathcal{D}_i \cap \mathcal{D}_j} [\phi_i \phi_j + K\gamma\Delta t\, \nabla\phi_i.\nabla\phi_j]\, d\Omega \\ m'_{i,j} &= \int_{\mathcal{D}_i \cap \mathcal{D}_j} [\phi_i \phi_j - K(1-\gamma)\Delta t\, \nabla\phi_i.\nabla\phi_j]\, d\Omega \\ S_i &= K\,\Delta t \int_{\Gamma \cap \Gamma_i} \phi_i \left[\gamma g^{n+1} + (1-\gamma) g^n\right] ds \end{aligned}$$

3.4 Consistency, stability and convergence

3.4.1 The desirable properties of numerical schemes

The definitions of consistency, stability and convergence were introduced in relation to finite-difference representations but the same concepts can be extended to finite-element representations :

- **Consistency**
  A finite difference numerical representation is said to be consistent if the truncation error vanishes as the finite-difference mesh size (both in space and time) approaches zero. This means that the limit of the difference representation is indeed the continuous differential equation. Consistency implies more than merely a good limit behaviour of each finite difference approximation to individual derivatives.

- **Stability**
  As explained above, the application of a finite difference representation results in replacing the PDE by a system of algebraic equations. Stability is the tendency for any spontaneous perturbations (such as round-off errors, noise in initial or boundary conditions, ...) in the solution of this system of equations to decay.

- **Convergence**
  A convergent finite-difference scheme is defined mathematically as one in which all values of the finite difference solution approach the parent continuum differential equation solution as the mesh size approaches zero.

  Lax's theorem states that : given a properly posed initial value problem governed by a system of *linear* PDE, a finite difference representation of this system is convergent if and only if it is both consistent and stable.

In practical situations, the Lax theorem is frequently assumed to hold even for *non - linear* PDE.

In a perfect world, it would be possible to conduct a complete prior analysis of a numerical scheme's properties before programming it. However, practical problems typically involve variable coefficients, nonlinearities and complicated types of boundary conditions, in brief all kinds of things whose influence can generally not be appraised in the frame of a nice theoretical analysis. Besides, consistency and convergence studies are concerned with the behaviour of the scheme at the limit when the space and time steps reduce to zero, whereas, in practical applications, the grid size is usually far from being negligible. Consequently, analysing a scheme always requires running it on test cases, as relevant as possible with respect to the planned applications.

Nevertheless, we shall mention herafter the main tools available for getting a first hint at a scheme's behaviour. This presentation is deliberately brief. The interested reader will find in Appendix C a more detailed account of methods suitable for the study of FDMs, which have been used in the context of this work.

3.4.2 Consistency study

Finite difference methods Consistency is studied by considering the truncation error (TE) of the finite-difference representation. TE is a polynomial expression of the space and time steps, the coefficients being space and time derivatives of dependent variables. Notation TE $= O(\Delta x^n, \Delta t^m)$ means that the lower-order terms of the truncation error are n^{th} and m^{th} order ones respectively for the space and time step. Obviously, as soon as $n \geq 1$ and $m \geq 1$, the truncation error vanishes as the mesh sizes decrease and the related scheme proves to be consistent.

Considering that if $n > m$, δ^n tends to zero quicker than δ^m as δ decreases, it can be tempting to prefer high-order finite difference approximations of the derivatives as they result in higher-order truncation error. However, the formal superior accuracy of high-order schemes can be misleading. Indeed, in order to achieve cost-effectiveness, both the number of nodes and the number of iterations throughout a simulation need to remain limited. Consequently, both the space and time steps keep somewhat coarse.
Deriving higher-order approximations usually implies to involve more neighbours of a node. If the space step is too big, it may just become meaningless, from a physical viewpoint, to introduce an explicit influence of these farther off neighbours by the way of our numerical algorithm.
Besides, using more nodes can significantly increase the complexity of the resulting algebraic equations, notably for boundary nodes, and consequently raise the computational cost especially if the scheme is implicit. For multivariable problems, the use of a staggered grid (i.e. each variable defined on different, intertwined, grids) may efficiently improve the approximation (cf

C.2).

Finite element methods Consistency is generally assessed in the frame of variational analysis. The differential operator $\mathcal{L}$ should first possess the properties (e.g. positiveness, symmetry) which allow to give a variational interpretation of the PDE. Then, the properties of the vectorial space $\mathcal{F}_h$ generated by the combined choice of elements and shape functions need to be appraised.

The subscript h refers to the maximum dimension of any element of the discretization. Consistency means that the series of $\mathcal{F}_h$ are such that, whatever the required level of precision, any continuous and differentiable function defined over computational domain $\mathcal{D}$ can be satisfactorily approximated with some element belonging to one of the $\mathcal{F}_h$.

As already mentioned (cf. 3.3.1), suitable families of finite elements have already been identified and thoroughly studied. Examples of the assessment of approximation error linked to a particular finite element discretization (the counterpart of truncation error for FDM) can be found for instance in (Goussebaile *et al.*, 1986), section 3.1.5 or (Fletcher, 1991), sections 5.3 and 5.4 .

3.4.3 Stability and convergence

3.4.3.1 Linear PDE

Finite difference methods The two most common techniques used to study the stability of FDM are the Von Neumann (cf C.3.1) and matrix (cf C.3.2) methods. Let f^* be the "ideal", error-free, solution of a finite difference equation (FDE), f the numerical solution of the system of algebraic equations associated with the FDE, and $\xi = f^* - f$ the corresponding error. It can be shown that, if the algebraic equations produced by discretization are linear, the error terms ξ satisfy the same set of equations as f^*.

1. **Von Neumann method**
   We consider a Fourier series expansion of the error ξ. Then, the decay or amplification of each mode is considered separately to determine stability or instability.

   For instance, in a one-dimensional case, we write

$$\xi_i^n = \xi(x_0 + i\Delta x, t_0 + n\Delta t) = \sum_k \mathcal{A}_k (t_0 + n\Delta t) \exp j\omega_k . i\Delta x \tag{3.27}$$

($j^2 = -1$). The ω_k are the wave numbers : the related wavelengths $\wedge_k = 2\pi/\omega_k$ are multiple of the space step ($\wedge_k = k\Delta x$). The elementary component corresponding to $\wedge_k$ is given by :

$$\varphi^n_{k,i} = \mathcal{A}^n_k \exp j\theta_k i \tag{3.28}$$

(with phase angle $\theta_k = \omega_k \Delta x$ and modulus $\mathcal{A}^n_k = \mathcal{A}_k\,(t_0 + n\Delta t)$)

As the PDE is linear, elementary components obey the same FDE as the global solution f^* and the complete error ξ. Developing the FDE at each node, and eliminating the common factor $\exp j\theta_k i$, leads for each k to an equation of the kind :

$$\mathcal{A}^{n+1}_k = \mathcal{G}_{\text{num}} \mathcal{A}^n_k \tag{3.29}$$

The **amplification factor** $\mathcal{G}_{\text{num}}$ is a trigonometric function of θ_k and its features depend on the discretization parameters and on such factors as diffusivities or flow velocities for instance.

If the errors are to remain bounded, we must have

$$|\mathcal{G}_{\text{num}}| \leq 1 \quad \text{for all } \theta \tag{3.30}$$

Inequality 3.30 expresses the stability criterion of the FDE. It generally gives way to inequalities relating the space and time steps and the PDE coefficients.

Von Neumann analysis can also be applied to a *system* of linear PDEs with several variables. The study of the Fourier series expansion of the error *vector* leads to stability criteria analogous to 3.30 except that $\mathcal{G}_{\text{num}}$ is no longer a scalar but is a square matrix (dimension : the number of variables). Now stability requires that the *the spectral radius of* $\mathcal{G}_{\text{num}}$ *be less than 1.* (nb : the spectral radius is the euclidian norm of the matrix or, alternatively, the maximum eigenvalue modulus).

2. **Matrix method**

In this technique, we express the set of equations governing the error propagation in matrix form and examine the properties of the associated matrix.

As already mentioned, the errors obey the same FDE as the problem solution so that the evolution of vector $\underline{\xi}$ (components ξ_i) is governed by relations like 3.15 or 3.16. The matrices describing the behaviour of $\underline{\xi}$ can differ slightly from matrices related to the evolution of solution $\vec{f}$ because of boundary conditions (cf section C.3.2). In summary, we obtain something like

$$\underline{\xi}^{n+1} = \underline{G}\,\underline{\xi}^n \quad \text{for } n = 0,\ 1,\ 2 \ldots \tag{3.31}$$

the dimension of square matrix $\underline{G}$ being approximately equal to the number of grid nodes.

Now we take interest in the evolution of the norm of vector $\underline{\xi}$. It remains bounded if and only if the spectral radius of matrix $\underline{G}$ is less than unity.

While the matrix method allows naturally to include in the stability analysis the study of the influence of boundary condition treatment, it leads to the manipulation of much more complicated objects (matrices instead of scalars) than the Von Neumann approach. In particular, when the PDE is multidimensional or when the coefficients of the linear PDE are space dependent, the $\underline{G}$ matrix and the subsequent search for its eigenvalues can become quite complex and uneconomical. Consequently, the Von Neumann method is the preferred one.

The theoretical scope of application of the Von Neumann method is limited (linear PDEs with constant coefficients !). It can nevertheless be applied to PDEs with variable coefficients but it leads then to a series of *local* criteria, which are necessary but not always sufficient to ensure stability. Similarly, if the coefficients are unsteady, the stability criteria will also result in time-dependent requirements on the discretization parameters.

For a very simple linear PDE (e.g. the pure advection equation), it is possible to express the **exact** amplification factor of the PDE. It can then be compared with the FDE amplification factor $\mathcal{G}_{\text{num}}$. Consequently, the damping ($\mathcal{G}_{\text{num}} < 1$) and phase shift induced by the FDE can be studied precisely for each wavelength. Such analysis is called the study of **dissipation** and **dispersion** phenomena respectively (cf section C.4).

Finite element methods Due to the formalism of finite element methods, their stability is studied in a context similar to the application of the matrix method to FDM. While the variational interpretation of the PDE may allow us to demonstrate some properties (e.g. positiveness, symmetry ...) of the system matrix without looking closely at each of its component, the FEM matrices are often significantly more complex than matrices produced by finite difference discretisation. Consequently, a direct assessment of their eigenvalues and spectral radius is even less straightforward.

In fact, solving linear systems corresponding to FEM implies generally to practice matrix algebra at a higher-level than for FDM systems. The study of the convergence of an FEM can then be considered in a more general framework : the study of different methods, direct or iterative ones, available to inverse linear systems. A review of these methods is the object of section 3.5 and, in a more extended version, of Appendix D.

3.4.3.2 Non-linear PDE

In fact, the techniques described in 3.4.3.1 are applied after linearising locally systems such as eq. 3.17 into systems such as eq. 3.18. However, the study of the linearised system provides only necessary requirements for the stability of the original problem. The strategy generally applied

is then to respect the linear stability criteria while allowing for some "safety margin".

Splitting (cf section 3.6 further down) both alleviates and complicates the analysis of a numerical scheme. Indeed, on one hand, the individual operators associated with each fractional step can be easier to study than the global one. On the other hand, one has then to combine the various stability criteria. Caution must be exercised in doing so (cf C.3.1). Notably, the reach of each criteria can be different (cf C.3.3), according to the property of the related operator (linear operator, constant coefficients operator, etc ...). Finally, the influence of boundary condition treatment eventually becomes quite obscure.

3.5 Solving finite difference and element systems

As explained in 3.2.3 and 3.3.3, when applied to flow equations FDM and FEM (and FVM also) typically produce a system of equations that can be written, for each time step, as

$$\underline{A}(\vec{f})\,\vec{f} = B \tag{3.32}$$

where $\vec{f}$ is the vector of $\mathcal{N}$ unknown nodal values. The regular matrix $\underline{A}$ contains the algebraic coefficients arising from discretization. If the related PDE is non-linear, $\underline{A}$ depends on the solution $\vec{f}$ itself. $\underline{A}$ is typically sparse. B is made up of the coefficients associated with discretization and of known values of $\vec{f}$ (e.g. past and boundary values).

The following sections 3.5.1 and 3.5.2 review briefly some approaches to the solution of eq. 3.32. A more extended presentation is given in Appendix D.

3.5.1 Iterative methods for solving linear systems

We shall first examine the case of a linear PDE (or system of PDE). Solving 3.32 amounts then to a matrix inversion problem :

$$\underline{A}\,X = B \tag{3.33}$$

The solution of 3.33 can be achieved, either directly, either iteratively. If the dimension of matrix $\underline{A}$ (i.e. the number of computational nodes) is large, direct inversion generally proves to be rather uneconomical unless the matrix has a *very* special structure (e.g. triangular, tridiagonal, block-tridiagonal, pentadiagonal matrices). The most common direct method is the Gauss one (cf D.1) which consists of transforming the initial linear system into a triangular one by performing a linear combination of the various relations of the system.

Hereafter, we deal only with iterative methods, as they are the most used.

3.5.1.1 General methods

The general structure of iterative techniques is illustrated by rewriting $\underline{A} = \underline{M} - \underline{N}$ where $\underline{M}$ should be a regular matrix, close to $\underline{A}$ in some sense (i.e. $\|\underline{M}\| \simeq \|\underline{A}\|$), but easy to invert. We have then :

$$\underline{A}X = B \iff X = \underline{M}^{-1}\underline{N}X + \underline{M}^{-1}B \tag{3.34}$$

so that we can consider building a series $X^{(k)}$ approximating the solution X by :

$$X^{(k+1)} = \underline{M}^{-1}\underline{N}X^{(k)} + \underline{M}^{-1}B \tag{3.35}$$

$$\text{or } X^{(k+1)} = X^{(k)} - \underline{M}^{-1}R^{(k)} \tag{3.36}$$

where $R^{(k)}$ is the vector of equation residuals at the kth step of the iteration,

$$R^{(k)} = \underline{A}X^{(k)} - B \tag{3.37}$$

We have $R^{(k+1)} = \underline{N}\,\underline{M}^{-1}R^{(k)}$, so that the scheme described by 3.35 or 3.36 converges only if the spectral radius of $\underline{N}\,\underline{M}^{-1}$ is less than 1.

The various iterative methods differ in their choice of $\underline{M}$. Let us write $\underline{A} = D - L - U$ where D is the diagonal matrix, L and U strictly lower and upper triangular matrices respectively.

- **Jacobi method** amounts to :

$$\underline{M} = D \quad \text{and} \quad \underline{N} = L + U \tag{3.38}$$

- **Gauss - Seidel method** corresponds to :

$$\underline{M} = D - L \quad \text{and} \quad \underline{N} = U \tag{3.39}$$

- In the **successive overrelaxation method (SOR)** a *relaxation parameter* λ ($\lambda \neq 0$) is introduced :

$$\underline{M} = \frac{1}{\lambda}D - L \quad \text{and} \quad \underline{N} = \left(\frac{1}{\lambda} - 1\right)D + U \tag{3.40}$$

Convergence is not easily studied for the general case. However, for some special situations, we have the following important results (Fletcher, 1991) :

- If $\underline{A}$ is diagonally dominant (i.e. $|a_{ii}| \geq \sum_{j \neq i} |a_{i,j}|$), Jacobi and Gauss-Seidel methods do converge.
- The Gauss-Seidel method typically converges twice as fast as the Jacobi method.

- The condition $\lambda \in]0,2[$ is necessary for SOR to converge. If $\underline{A}$ is positive definite, this condition proves to be sufficient.
- The SOR rate of convergence is sensitive to the choice of λ. Let μ be the spectral radius (maximum eigenvalue modulus) of matrix $\underline{N}\,\underline{M}^{-1}$. The optimum choice would be

$$\lambda_{\text{opt}} = \frac{2}{1 + (1-\mu)^{1/2}}$$

Finding μ explicitly can be expensive. Consequently, the preferred strategy is to obtain an estimate of μ as the iteration proceeds and subsequently to update λ.

The SOR scheme with a good estimate of λ_{opt} is considerably more efficient than either the Jacobi or Gauss-Seidel methods.

The iterative methods presented above may prove rather slow. More efficient algorithms can be devised when the matrix $\underline{A}$ has some additional useful properties (see gradient methods). Alternatively, the convergence of Jacobi, Gauss-Seidel and relaxation methods can be improved by resorting to multigrid techniques, whose principles are indicated in D.3.2.

Sometimes, $\underline{A}$ stems from the discretization of a spatial operator which can be easily split along the different spatial directions, so that for instance we could write in two-dimensions $\underline{A} = A_1 + A_2$ where each A_i corresponds strictly to a one-dimensional operator. **ADI (Alternate Direction Implicit) methods** consider two different decompositions of $\underline{A}$ ($r > 0$)

$$\underline{A} = (A_1 + rI) + (A_2 - rI) \tag{3.41}$$

$$\underline{A} = (A_2 + rI) + (A_1 - rI) \tag{3.42}$$

and use them alternatively, so that the iteration proceeds as :

$$(A_1 + rI)\ X^{(k+1/2)} = B - (A_2 - rI)\ X^{(k)} \tag{3.43}$$

$$(A_2 + rI)\ \ X^{(k+1)} = B - (A_1 - rI)\ X^{(k+1/2)} \tag{3.44}$$

Each step requires only the solution of much simpler (generally tridiagonal), monodimensional systems. As explained in section D.4.1, if $\underline{A}$, A_1, A_2 are positive and at least one of the A_i is definite, ADI methods are convergent.

3.5.1.2 Gradient methods

If matrix $\underline{A}$ is symmetric, positive and definite, $\underline{A}X - B$ can be interpreted as the gradient ($\nabla \mathcal{J}$) of quadratic function $\mathcal{J}$:

$$\mathcal{J}(X) = \frac{1}{2} X^t \underline{A} X - B^t X$$

and solving the linear system is equivalent to finding a minimum of $\mathcal{J}$.

One effective method for finding the minimum of a convex function is to move in the vectorial space by following as closely as possible the direction in which $\mathcal{J}$ decreases most. Such a direction is of course variable and is given at any point by the (opposite) local gradient direction (cf D.3.1). Gradient methods are based on this idea :

- $X^{(k)}$ being known, we look for the following iterate in the gradient direction :

$$\begin{aligned} X^{(k+1)} &= X^{(k)} - \rho \nabla \mathcal{J}\left(X^{(k)}\right) \quad \text{where } \rho \geq 0 \\ \text{i.e. } X^{(k+1)} &= X^{(k)} - \rho\left(\underline{A} X^{(k)} - B\right) \qquad (3.45) \\ \text{or } X^{(k+1)} &= X^{(k)} - \rho R^{(k)} \qquad (3.46) \end{aligned}$$

- ρ may be chosen constant or its value can be optimised at each time step, for instance $\rho = \rho_{\text{opt}}$ so that,

$$\mathcal{J}\left(X^{(k+1)}\right) = \min_{\rho} \mathcal{J}\left(X^{(k)} - \rho \nabla \mathcal{J}(X^{(k)}) \right) \qquad (3.47)$$

Simple gradient methods as presented above are based on a *local* optimisation criterion only. **Conjugate gradient methods** represent an improvement : instead of minimizing $\mathcal{J}$ at each iteration in the local gradient direction, we aim to minimize it in the vectorial space generated by the local and all previous gradients. This allows to take advantage of past information and guarantees that the minimum will be reached in a finite number of iterations (maximum number : the total space dimension $\mathcal{N}$). Besides, it can be demonstrated (Labadie, 1986; Fletcher, 1991) that a conjugate gradient iteration can be programmed simply (cf. the set of relations D.20 to D.23 in D.3.1) and requires neither additional storage nor significantly more computational time than simple gradient methods.

The **method of decomposition with coordination** can be implemented if $\underline{A}$ can be split according to different directions, each resulting one-dimensional operator and the related matrix being symmetric, positive and definite. Then,

$$\underline{A} X = B \iff (A_1 + A_2) X = B_1 + B_2 \qquad (3.48)$$

Let us introduce the functions

$$\mathcal{J}_i(X) = X^t A_i X - B_i \quad \text{for } i = 1, 2 \qquad (3.49)$$

The problem 3.48 is equivalent to the minimization of function $\mathcal{J} = \mathcal{J}_1 + \mathcal{J}_2$. The form of $\mathcal{J}$ suggests splitting the minimization problem into two parts, corresponding to each $\mathcal{J}_i$, $i = 1, 2$. In fact, solving eq. 3.48 amounts to solving the problem of constrained minimization :

$$\min_{X_1, X_2 \, (X_1 = X_2)} \mathcal{J}_1(X_1) + \mathcal{J}_2(X_2) \tag{3.50}$$

As detailed in D.4.2, this can be done iteratively, each iteration implying only the solution of a series of monodimensional problems.

3.5.1.3 Preconditioning

Preconditioning (cf D.3.1) is an important stage in the implementation of gradient methods, as it can drastically improve their efficiency.

First let us define the *condition number* of a matrix. Let us consider different kind of perturbations of the system $\underline{A}X - B$ and the resulting perturbation δX on solution X : first a perturbation δB of the right-hand side term, secondly a perturbation δA of the matrix itself. It can be demonstrated that we have respectively

$$\frac{\|\delta X\|}{\|X\|} \leq \|\underline{A}\| \, \|\underline{A}^{-1}\| \, \frac{\|\delta B\|}{\|B\|} \tag{3.51}$$

$$\frac{\|\delta X\|}{\|X + \delta X\|} \leq \|\underline{A}\| \, \|\underline{A}^{-1}\| \, \frac{\|\delta A\|}{\|\underline{A}\|} \tag{3.52}$$

The influence of any perturbation appears to depend strongly on the *condition number* $\text{cond}(\underline{A}) = \|\underline{A}\| \, \|\underline{A}^{-1}\|$ which should preferably be set as small as possible. The value of $\text{cond}(\underline{A})$ depends on the chosen norm but it can be shown that, whichever it is, $\text{cond}(\underline{A}) \geq 1$ (Labadie, 1986). For a symmetric matrix and with the euclidian norm, $\text{cond}(\underline{A})$ reduces to the ratio between maximum and minimum eigenvalues.

For most gradient methods, the influence of the condition number may be evaluated explicitly. For instance, for the standard conjugate gradient method, we have (Labadie, 1986; Goutal & Hérard, 1990) :

$$\| X - X^{(k)} \| \leq 2 \left(\frac{\sqrt{\text{cond}(\underline{A})} - 1}{\sqrt{\text{cond}(\underline{A})} + 1} \right)^k \| X - X^{(0)} \|$$

As $\text{cond}(\underline{A})$ departs from its minimum value (1), the performance of the algorithm decreases.

Preconditioning refers to the replacement of the original system $\underline{A}X = B$ by $\underline{C}\,\underline{A}X = \underline{C}B$, where matrix $\underline{C}$ is chosen so that $\underline{C}\,\underline{A}$ "contains the same amount of informations" than $\underline{A}$, keeps all the nice properties of $\underline{A}$ (symmetry, positiveness ...) but has a lower condition number.

The efficiency of a preconditioning method depends on the balance between the increased complexity its implementation induces and the superior convergence rate it brings. It depends

obviously on the properties of the original linear system and consequently on the discretisation and differentiation applied to the original PDE.

3.5.2 Non-linear systems

The solution of non-linear systems such as 3.32 requires to linearize them on each time interval. The most immediate method for linearizing 3.32 from time t^n to $t^n + \Delta t$ is to set :

$$\underline{A} \simeq \underline{A}(\vec{f}^n) \tag{3.53}$$

Should the time step Δt be small enough, this often proves to be a fair enough approximation. However, if $\underline{A}$ is fast-varying, the upper allowable limit for Δt (which keeps 3.53 safe) can be unfit as regards numerical cost-effectiveness. With a bigger time step, it is necessary to adopt an other strategy : $\underline{A}$ and the solution of 3.32 from t^n to $t^n + \Delta t$ are approximated through an iterative process. Once again, the problem reduces to the task of solving a linear system, this time a new one at each iteration. The outlines of such approach are given hereafter and in section D.5.

3.5.2.1 Newton's methods

The solution of 3.32 is then viewed as the problem of finding the root that sets to zero the residual :

$$R(X) = \underline{A}(X)\,X - B \tag{3.54}$$

The Newton-Raphson approach applies to typical problems such as :

$$G(X) = 0 \iff g_i\,(x_1, x_2, \ldots x_N) = 0 \;\; \forall i = 1, 2, \ldots N \tag{3.55}$$

where functions g_i are not linear. By expanding each g_i into a Taylor series and neglecting terms of second-order and above (cf D.5), it can be shown that the root may be approached through the following iterative process :

$$[\nabla G]\,\delta X^{(k)} = -G\left(X^{(k)}\right) \tag{3.56}$$

$$X^{(k+1)} = X^{(k)} + \delta X^{(k)} \tag{3.57}$$

where the Jacobian matrix $[\nabla G]$ is defined by :

$$[\nabla G]_{i,j} = \frac{\partial g_i}{\partial x_j}\left(X^{(k)}\right) \tag{3.58}$$

This cycle can be applied straightforwardly to $G = R$ as defined by 3.54. (Notation $\underline{J}^{(k)}$ denotes from now on the estimation of the Jacobian of R at iteration k.) Usually, the cycle is initialized

by setting $X^{(0)} = \vec{f}^n$. Each iteration implies solving a *linear* system of equations 3.56. This can be done by resorting to some of the techniques described summarily in 3.5.1 or with more details in sections D.1 to D.4.

Newton's method is like a double-edged sword : on the one hand it converges quickly as soon as the current iterate falls nearby the root (with a quadratic reduction of the error); on the other hand it can diverge spectacularly if the initial guess, or an intermediate approximation, goes beyond the bounds of some specific radius of convergence (Fletcher, 1991) (chap. 6). In order to minimize the risks of throwing the next approximation too far apart from the root, the corrections δX are not implemented in full : eq. 3.57 is replaced by

$$X^{(k+1)} = X^{(k)} + \rho^{(k)} \delta X^{(k)} \tag{3.59}$$

Eq. 3.59 bears a formal similarity with a gradient iteration like 3.46, solution $\delta X^{(k)}$ of 3.56 being interpreted as a search direction. $\rho^{(k)}$ can be optimised (cf. D.5), the criterion being for instance to reduce as much as possible the resulting residual mean square, which is analogous to what is done in the optimal gradient method (cf. eq. 3.47) .

3.5.2.2 Quasi-Newton methods

The main contribution to the execution time of Newton's methods is the solution of system 3.56, namely

$$\underline{J}^{(k)} \delta X^{(k)} = -R^{(k)} \tag{3.60}$$

One first strategy to alleviate this stage is to "freeze" $\underline{J_k}$ during some subsequent δk iterations. Thus, it needs to be factorised only once every δk steps. Yet, such simplification may lower the accuracy and hinder the convergence. An alternative approach is provided by Quasi-Newton methods (Fletcher, 1991) (section 6.1), (Goutal & Hérard, 1990) (section 3.9) where the combination of 3.60 and 3.59 is replaced by

$$X^{(k+1)} = X^{(k)} - \rho^{(k)} \underline{H}^{(k)} R^{(k)} \tag{3.61}$$

$\underline{H}^{(k)}$ is an approximation to $\left(\underline{J}^{(k)}\right)^{-1}$. $\underline{H}^{(k)}$ is modified at each iteration so that it approaches more and more closely $\left(\underline{J}^{(k)}\right)^{-1}$ as $X^{(k)}$ converges to the exact solution. To different updating formulae for $\underline{H}$ correspond different Quasi-Newton methods.

(Fletcher, 1991) (sec. 6.1) reports that the efficiency of Quasi-Newton methods often depends on $\underline{J}$ having special properties such as positive definiteness. Consequently, these methods need to be tailored case-by-case. It turns out that some of them can be interpreted as preconditioned conjugate gradient algorithms (Goutal & Hérard, 1990) (sec. 3.9) .

3.6 Splitting

The use of operator splitting and related *fractional step methods* is widespread in computational fluid dynamics. The purpose of such methods, first introduced by (Yanenko, 1968), is to replace a complex calculation by a sequence of simpler ones. Their implementation requires, first to split the governing equations into elementary parts, secondly to solve successively each part within each time step. As mentioned in (Usseglio-Polatera & Chenin-Mordojovich, 1988), there exist different splitting criteria.

3.6.1 Time splitting

Splitting can be introduced purely for numerical reasons, in order to stabilize a scheme, avoid tedious calculations, etc ... Time splitting can apply both to one-dimensional and multi-dimensional operators.

An example is given by predictor-corrector schemes. Let us consider the general equation :

$$\frac{\partial y}{\partial t} = f(y, t) \tag{3.62}$$

In order to achieve second-order accuracy, it could be integrated by the trapezoidal rule, namely

$$\frac{y^{n+1} - y^n}{\Delta t} = \frac{1}{2}\left[f\left(y^{n+1}, t^{n+1}\right) + f(y^n, t^n)\right] \tag{3.63}$$

$$\text{or} \quad \frac{y^{n+1} - y^n}{\Delta t} = f\left(\frac{y^{n+1} + y^n}{2}, t^{n+1/2}\right) \tag{3.64}$$

However, both relations 3.63 and 3.64 require an iterative treatment if f is not linear. In contrast, the scheme described by

$$\frac{y^{n+1/2} - y^n}{\Delta t/2} = f(y^n, t^n) \tag{3.65}$$

$$\text{and} \quad \frac{y^{n+1} - y^n}{\Delta t} = f\left(y^{n+1/2}, t^{n+1/2}\right) \tag{3.66}$$

is also second-order accurate and does not require any iteration (nb : higher-order approximation could be achieved by more general Runge-Kutta methods).

Another example is the alternate direction implicit (ADI) method. Let us consider the general form of linear systems associated with FDM (cf 3.2.3) :

$$\underline{A}X = B$$

Let us assume furthermore that the problem is two-dimensional and that $\underline{A}$ can be partitioned into two matrices A_i, $i = 1, 2$ corresponding to one-dimensional operators in each space direction. ADI schemes read typically :

$$\frac{X^{n+1/2} - X^n}{\Delta t} = \frac{1}{2}\left[B - A_1 X^{n+1/2} + A_2 X^n \right] \tag{3.67}$$

$$\frac{X^{n+1} - X^{n+1/2}}{\Delta t} = \frac{1}{2}\left[B - A_1 X^{n+1/2} + A_2 X^{n+1} \right] \tag{3.68}$$

Steps 3.67 and 3.68 are respectively equivalent to iterations 3.43 and 3.44 with the choice $r = 2/\Delta t$. As indicated in 3.5.1.1, this scheme is convergent if all matrices are positive and if one A_i at least is definite.

Each step requires solving a set of one-dimensional linear systems along each coordinate line (lines parallel to the first coordinate for step (3.67), to the second one for step (3.68)). Considering the usual derivatives involved in fluid dynamic PDEs and their usual discretization, these one-dimensional systems are most often governed by tridiagonal matrices.

The ADI approach has been made famous by (Leendertse, 1970) and is still a popular basis for numerous FDM flow models.

3.6.2 Space splitting

Space splitting is customary in multidimensional problems : it consists of separating the partial space derivatives according to the relevant space directions. As in ADI methods, the problem is then reduced to a succession of one-dimensional problems. Space splitting is generally less straighforward to implement for FEM than for FDM, and not so cost-saving.

Examples of such schemes are methods relying on an approximate factorisation of the differential operator. For instance, let us consider a diffusion equation such as

$$\frac{\partial f}{\partial t} = K_x \frac{\partial^2 f}{\partial x^2} + K_y \frac{\partial^2 f}{\partial y^2} \tag{3.69}$$

(for the sake of simplicity K_x and K_y are assumed constant). Let us discretize $\frac{\partial f}{\partial t}$ simply by $(f^{n+1} - f^n)/\Delta t$. We adopt an implicit scheme : each spatial derivative is a weighted average of its value at times t^n and t^{n+1} respectively. A_1 and A_2 correspond respectively to the discretization of operators $K_x \frac{\partial^2}{\partial x^2}$ and $K_y \frac{\partial^2}{\partial y^2}$, I denotes as usual the identity matrix. Consequently, eq. 3.69 becomes :

$$\left[I - \gamma \Delta t\, A_1 - \gamma \Delta t\, A_2 \right] f^{n+1} = \left[I + (1-\gamma)\Delta t\, A_1 + (1-\gamma)\Delta t\, A_2 \right] f^n \tag{3.70}$$

($\gamma \in]0, 1]$ implicitation parameter). The left-hand side operator can be approximated as follows :

$$I - \gamma \Delta t\, A_1 - \gamma \Delta t\, A_2 \simeq \left[I - \gamma \Delta t\, A_1 \right] . \left[I - \gamma \Delta t\, A_2 \right] \tag{3.71}$$

Going back to the original PDE, we could check that such approximation is satisfactory if

$$\frac{\partial^4 f}{\partial x^2 y^2} \ll \frac{\partial^2 f}{\partial x^2} \quad \text{and} \quad \frac{\partial^4 f}{\partial x^2 y^2} \ll \frac{\partial^2 f}{\partial y^2} \tag{3.72}$$

Then, the resolution could proceed in two stages :

$$[\, I - \gamma \Delta t\, A_1 \,]\, f^{n+1/2} = [\, I + (1-\gamma)\Delta t\, A_1 + (1-\gamma)\Delta t\, A_2 \,]\, f^n \tag{3.73}$$

$$[\, I - \gamma \Delta t\, A_2 \,]\, f^{n+1} = f^{n+1/2} \tag{3.74}$$

This is similar to ADI methods, except that we need to compute the right-hand side terms of equations only once.

As illustrated when comparing ADI and factorisation methods, the same multidimensional operator can be split differently, depending on whether we favour time or space splitting.

3.6.3 Process splitting

Process splitting is based on a recognition of the different *physical processes* described by the governing equations. The successive fractional steps correspond to the solution of each individual process. Process splitting is equally convenient for FDM and FEM. As regards FDM, the combination of process and space splitting is commonplace.

Process splitting allows one to apply to each process the numerical treatment best suited to the properties of the differential operator which describes it. Peculiarities of one operator can no longer damage the treatment of another.

On the other hand, one must note that, with process splitting, basic processes which in fact occur at the same time are considered to *succeed* within each time step. One may intuitively guess that the validity of such assumptions should vanish for too big a time step.

Process splitting is notably applied to the scalar transport equation : the hyperbolic part of the equation, associated with advection, is split from the parabolic part, associated with diffusion. This gives birth to the so-called *eulerian - lagrangian methods* (Baptista *et al.*, 1984). The advective part is usually treated with the characteristics method (see the following part) : the scheme is then free of stringent stability conditions which plague most of the FD or FE treatment of advective operators. Provided the spatial interpolation formula applied at the foot of each characteristic line is carefully chosen, such schemes also prove remarkably accurate (cf chapters 6 and 7). The remains of the equation are easily dealt with, either in FDM or FEM situations. For instance, once the operator is ridden of advective terms, the matrices produced by FE discretization easily fulfill the desirable requirements of symmetry and positiveness.

Process splitting may also apply to the flow equations. For instance, the two-dimensional St-Venant equations introduced in the previous chapter (sec. 2.4.5) can be written :

$$\frac{\partial \phi}{\partial t} + \vec{U}.\nabla\phi = S_{xy} \quad (3.75)$$

$$\frac{\partial h}{\partial t} + \vec{U}.\nabla h + h\mathrm{div}\vec{U} = 0 \quad (3.76)$$

$$\frac{\partial U}{\partial t} + \vec{U}.\nabla U + g\frac{\partial \zeta}{\partial x} + f_b U = S_x \quad (3.77)$$

$$\frac{\partial V}{\partial t} + \underbrace{\vec{U}.\nabla V}_{\mathrm{I}} + \underbrace{g\frac{\partial \zeta}{\partial y} + f_b V}_{\mathrm{III}} = \underbrace{S_y}_{\mathrm{II}} \quad (3.78)$$

(nb : Most notations were introduced in the previous chapter. In addition, $\vec{U}$ denotes the velocity vector, whose components are U and V respectively along the x and y axis.)

Terms I, II and III correspond respectively to *advection, mass and momentum diffusion* and *propagation and friction* processes. When splitting is applied, FDM usually treat I, II, III apart (e.g. (Mary, 1982; Benque *et al.*, 1982; Dan N'Guyen, 1988; Dan N'Guyen, 1993)) whereas, in FEM, terms II and III can be solved within the same fractional step (e.g. (Goutal, 1987; Hervouet & Watrin, 1988)).

3.6.4 Advantages vs. shortcomings

According to (Usseglio-Polatera & Chenin-Mordojovich, 1988), the main, and considerable, advantages of splitting techniques are the following :

- *Process splitting increases the potential for numerical accuracy.*
  Each operator can be treated independently with the best suitable technique. Notably, the splitting may contribute to build intermediate systems of equations which are well-behaved (e.g. symmetric, positive), i.e. which are suitable for the application of efficient solution techniques (cf sec. 3.5).

- *Splitting produces cost-effective numerical codes*, either because stability requirements can be relaxed (e.g. advection and process-splitting) allowing notably the use of bigger time steps, either because simpler algorithms and numerical objects are being manipulated (e.g. reduction of the problem dimension by space splitting).

- When process splitting is applied, *it results naturally in modular codes.* This can help physical analysis, as the different basic processes governing flow and transport are easily tuned differently, even neglected, so that their respective importance can readily be evaluated.

However, some shortcomings have also been clearly identified :

- First, let us note that one category of schemes produced by splitting has been extensively studied, namely the ADI schemes. Their flaws when applied to irregular (as regards bathymetry and geometry) computational domains have been demonstrated for a long time (Weare, 1979; Stelling *et al.*, 1986).

  Figure 3.5, extracted from (Stelling *et al.*, 1986), shows the numerical domain of influence of any perturbation occurring at point P, in the case of a zigzaging channel and in the case of an island. The discretization follows the usual staggered grid (cf C.2). We can check that this domain is purely constricted by the discretisation, so that eventually, for large time steps, the physical and numerical regions of influence (or alternatively the physical and numerical perturbation propagation speeds) no longer coincide. This is particularly dangerous as, due to its stability, the ADI scheme will indeed produce results at large time steps, it will not blow up, but these results will be deprived of any physical meaning.

  (Weare, 1979) demonstrated that the typical treatment of no-slip conditions in staggered grid ADI schemes could also lead to the introduction of spurious friction and momentum dissipation, should the boundary be irregularly shaped.

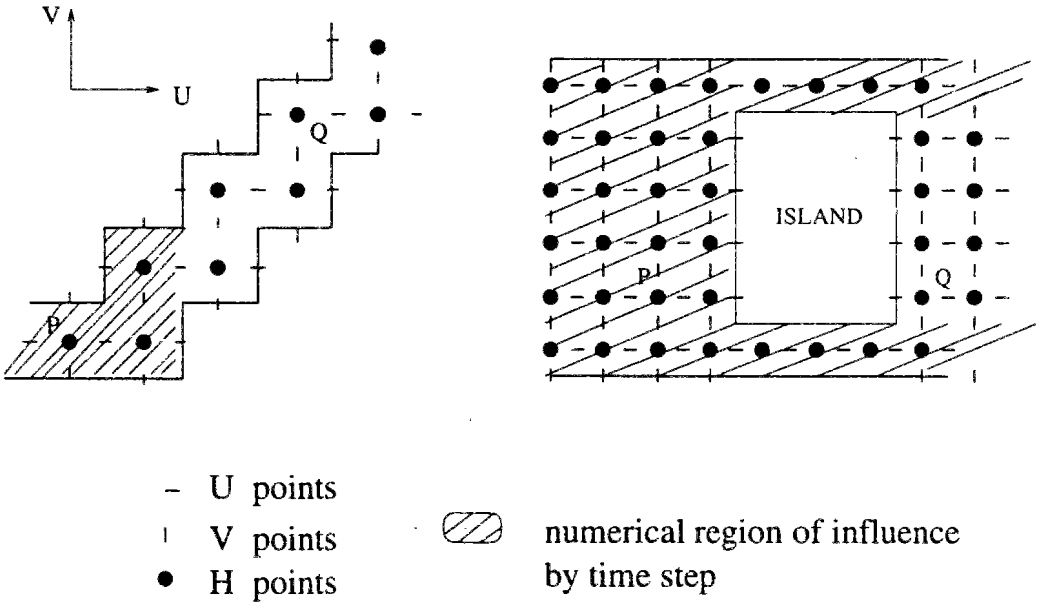


Figure 3.5: Computational domains of influence

- More generally, the major drawbacks are linked to the implementation of boundary conditions.

Splitting introduces additional intermediate variables whose values (or gradients) need to be prescribed on all or part of the boundaries. Yet, physical laws or, alternatively, available data are related only to variables at entire time steps.

Besides, most boundary conditions, when applied at corners of the computational domain, introduce coupling, either between different variables, either between grid nodes located along different coordinate lines of the computational domain. This naturally jeopardizes the application of splitting, unless (mostly empirical) modifications of the boundary conditions are introduced.

This topic will be further detailed in the following section.

In summary, splitting is a powerful tool but it should be applied with care. Usually, the conditions under which a fractional step code performs reasonably well cannot be completely derived from an individual analysis of each operator resulting from the splitting and must be numerically estimated (i.e. with the help of test cases). Besides, empirical adaptations, notably at boundaries, must generally be introduced in order to allow for the full benefits of the splitting techniques. Assessing the soundness of such formulations also requires some numerical testing.

3.7 Boundary conditions

As pointed out by Roache (Roache, 1985) (chap. 3), the importance of boundary conditions in flow computations can immediately be appraised when considering that all the various flow patterns of common fluids are essentially solutions of the same Navier-Stokes equation : the flow solutions are distinguished only by their boundary and initial conditions.

Attempts to determine realistic, accurate and stable boundary methods can prove highly frustating. Indeed, it has been found from numerical experiments (Roache, 1985) that the adequacy of any boundary condition can depend on the flow features, the chosen interior-point method, other boundary conditions and sometimes initial conditions. Any endeavour at devising a "universal" method appears to be doomed. Consequently, the treatment of boundary conditions is usually tailored case-by-case according to the category of studied flows and to the features of the specific code used. This choice often relies on empirical information and on the modeller's own experience. Thus, it is generally difficult to expose in a nice theoretical framework, which probably explains why relevant literature on the subject is rather sparse with respect to papers focussing on interior-points methods.

An exhaustive review of problems linked to boundary conditions stands clearly far beyond the scope of this section. In the following, we shall merely attempt to make the reader aware of the complexity of the subject. The examples we mention are restricted to the case of depth-averaged

(shallow water) flow equations (eq. 3.76 to 3.78).

3.7.1 Physical and mathematical approach

First, the initial conditions and the number and kind of boundary conditions must be such as to make the mathematical problem "well-posed". As quoted by (Abraham *et al.*, 1981), the main aspect of well-posedness is that small changes in boundaries conditions would lead to small changes only in the interior area solution. The development of catastrophic flow changes under smoothly-varying boundary conditions may often be a sign of ill-posedness.

The number of boundary conditions required to ensure well-posedness depends on the type of equations and on the actual flow conditions (Abraham *et al.*, 1981; Goutal, 1987). As regards the case of shallow water two-dimensional equations, a pioneer work was achieved by (Daubert & Graffe, 1967). The shallow-water equations (3.76 to 3.78) have three dependent variables (water-depth or free surface elevation and the two components of the depth-averaged velocity). The advective terms (terms I) are usually not negligible so that the type of the equations depends mainly on whether the diffusive operator (terms II) is included or not in eq. 3.77 and 3.78 : in the latter case, the system is hyperbolic, in the first case, the velocity equations become parabolic. The auxiliary (i.e. boundary and initial) conditions needed for different families of PDE can be formally analyzed (Fletcher, 1991) (chap. 2), notably with the methods of characteristics (Abbott, 1979) (chap. 3). This leads to the following results (table extracted from (Abraham *et al.*, 1981)) :

Table 3.1: Number of open boundaries cond. for $\neq$ types of shallow water eq.

Type of flow	Hyperbolic system		Incomplete Parabolic system	
	Inflow	Outflow	Inflow	Outflow
Sub-critical	2	1	3	2
Super-critical	3	0	3	2

In fluvial hydraulics, the flow situations are often sub-critical and the effect of momentum diffusion, even if taken into account in the interior area, is generally neglected at the open boun-

daries. The preferred choice is usually to prescribe the velocity or unit discharge distribution at the inflow and the free surface elevation at the outflow.

At closed boundaries, the conditions must be essentially consistent with the fact that there cannot be any mass or momentum flux across closed boundaries. Thus, the velocity component normal to the boundary, denoted U_n, is zero :

$$U_n = 0 \tag{3.79}$$

The free surface gradient normal to walls is also zero :

$$\frac{\partial \zeta}{\partial n} = 0 \tag{3.80}$$

However, different conditions may apply to the tangential component U_t of the velocity. On walls, as the water is a viscous fluid, its velocity is theoritically zero :

$$U_t = 0 \tag{3.81}$$

Yet, in order to describe accurately the velocity profile along the wall, the choice of a no-slip boundary condition should be combined with a local refinement of the computational grid (Mary, 1982; Stelling & Wang, 1984). In many applications, the practical grid size is much too large and a perfect-slip boundary condition is used instead, namely :

$$\frac{\partial U_t}{\partial n} = 0 \tag{3.82}$$

Another, intermediate, option is to assume first that the velocity profile in the wall boundary layer follows a logarithmic profile and secondly that the mesh size next to the wall is fine enough so that the first interior node belongs to this boundary layer. Then, we have

$$\frac{\partial U_t}{\partial n} = a\, U_t \tag{3.83}$$

Boundary conditions need sometimes to be adapted in order to take into account the peculiarities of the domain geometry. At sharp corners of the computational grid (cf fig. 3.6), there are notably several ways to specify the normal direction. The simplest one is to take the average of the normals to each side of the protruding corner, respectively $\vec{n_1}$ and $\vec{n_2}$. Another could be to introduce there the normal to the real domain boundary (which is but approximated by the computational grid).

Let us assume we have chosen to apply a perfect slip (eq. 3.82) or partial slip (eq. 3.83) condition (so that velocities are not forced to be zero along the walls). This is acceptable apart from cases as illustrated in fig. 3.7, where the flow separates in the vicinity of corner M. Considering perfect or partial slip along each side of the protruding corner, in the immediate

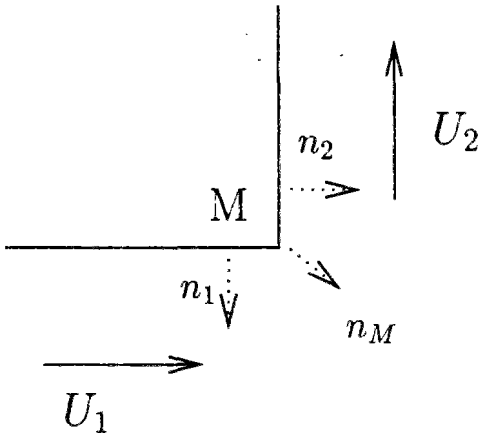


Figure 3.6: Flow around a corner

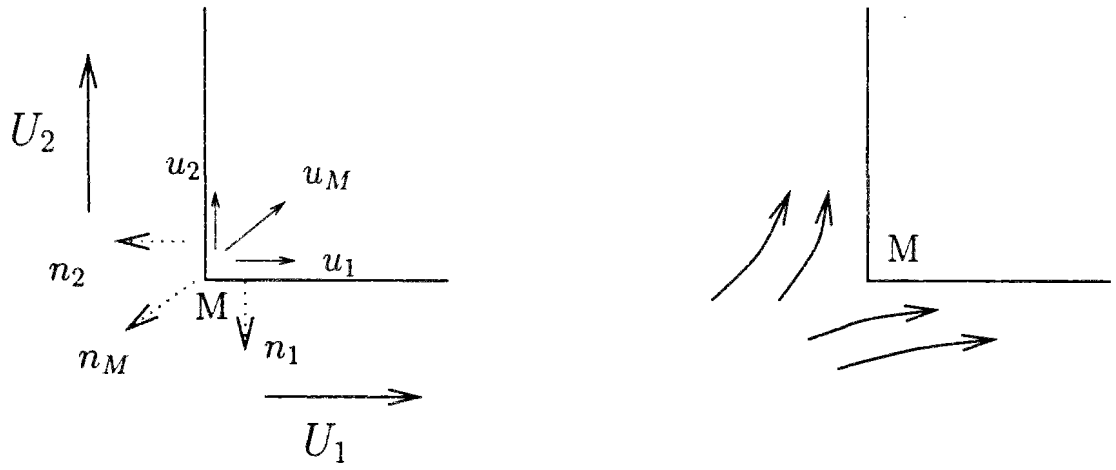


Figure 3.7: Flow separation

vicinity of M, could lead to inconsistent situations, whatever the choice of the normal direction (velocity directed outside the computational domain, see fig. 3.7) : the only way out is to postulate explicitly that the velocity is zero at M. Fortunately, this corresponds to our physical intuition as regards what occurs very locally in a separation zone.

3.7.2 Incompleteness of boundary conditions

3.7.2.1 Lack of data

As mentioned above, the open boundaries conditions are generally of the Dirichlet type, i.e. the *values* of the dependent variables must be prescribed. Yet, as soon as the problem deals with a natural water body, especially when it is large, the detailed knowledge of inflow and outflow distributions is generally beyond our grasp, for lack of means or adequate monitoring tools. Consequently, it becomes necessary to "imagine" appropriate boundary conditions.

Example 1 A typical situation is the case when we aim at modelling two-dimensional flow in a river reach knowing only the upstream global discharge and the mean downstream water elevation (or some law which relates it to the discharge). Such conditions are sufficient only from the point of view of one-dimensional calculations. Two-dimensional computations require that we postulate some velocity or unit discharge distribution at the inflow boundary. This can be done in several ways. The simplest is to postulate a uniform velocity distribution. Alternatively, an approximate distribution can be built according to the bathymetry of the inflow section, as is done in streamtube models (Holly, 1975; Holly, 1979). In both cases anyway, it is advisable to test different approaches and check their influence on interior solutions. It is safer, if possible, to place the computational domain limit some distance upstream of the area of interest, in order to further limit the consequences of these approximate boundary conditions.

Example 2 Another typical example is the simulation of coastal flows. The effect of the large-scale circulation outside the studied area has to be brought in through boundary conditions. This implies prescribing the transport in and out of the area or, alternatively, prescribing the baroclinicity and free surface elevation along the open boundaries as a function of time. Observational data are usually insufficient to build properly this kind of function. Although it can prove effective when computing the effect of *steady* forcings, the use of "clamped" (i.e. Dirichlet) conditions on the elevation has long been identified as a major source of error when simulating the *transient* response of estuaries or coastal areas : often the area reponds less energetically than it should. What is needed in fact is a boundary condition which should make the boundaries transparent to outgoing transients, yet permits the background tidal and mean elevations

to be prescribed and maintained. Such a condition, termed a *radiation boundary condition*, has been devised (Orlanski, 1976) and is since widely used (Blumberg & Kantha, 1985; Beckers, 1991; Dan N'Guyen, 1993). It reads :

$$\frac{\partial \zeta}{\partial t} + c\frac{\partial \zeta}{\partial n} = -\left(\frac{\zeta - \zeta_k}{T_f}\right) \tag{3.84}$$

where c is the phase speed of the wave ($c = \sqrt{g\,h}$) and n denotes as usual the direction normal to the boundary. The right-hand term of the equation implies that f is forced to some equilibrium value f_k with a time scale of the order of T_f . f_k itself can be a time-dependent function which integrates the main features of the large-scale and tidal motion. When T_f tends to zero, eq. 3.84 reduces to a clamped condition whereas, when T_f is large with respect to the time scale of the simulated phenomena, the open boundary becomes transparent to waves travelling at the phase speed c.

The previous examples illustrate how the prescription of adequate boundary conditions requires either commonsense and a good deal of trial and error (example 1), or gaining further insight into the dynamics of the studied flows and to build from scratch a formulation which respects these dynamics (example 2).

3.7.2.2 Splitting and the introduction of auxiliary variables

To illustrate the consequence of splitting on boundary conditions requirements, we shall consider the application of *process splitting* to eq. 3.76 to 3.78. As indicated above, the resolution may be fractioned in three steps :

1. Resolution of terms I leads to solving :

$$\frac{\partial f}{\partial t} + \vec{U}.\nabla f = 0 \quad \text{for } f = U, V, h \tag{3.85}$$

   $\frac{\partial f}{\partial t}$ is expressed as $\left(f^{n+1/3} - f^{n}\right)/\Delta t$ where $f^{n+1/3}$ denotes the intermediate advected variables.

2. Resolution of terms II leads to solving :

$$\frac{\partial f}{\partial t} = S \quad \text{for } f = U, V \tag{3.86}$$

   where S denotes a diffusion operator and $\frac{\partial f}{\partial t} = \left(f^{n+2/3} - f^{n+1/3}\right)/\Delta t$

3. For the last step, the remains of the equations are :

$$\frac{\partial h}{\partial t} + h\mathrm{div}\vec{U} = 0 \tag{3.87}$$

$$\frac{\partial \vec{U}}{\partial t} + g\nabla\zeta + f_b\vec{U} = 0 \tag{3.88}$$

where the time derivatives have the following meaning :

$$\frac{\partial h}{\partial t} = \frac{h^{n+1} - h^{n+1/3}}{\Delta t}$$

$$\frac{\partial \vec{U}}{\partial t} = \frac{\vec{U}^{n+1} - \vec{U}^{n+2/3}}{\Delta t}$$

Solving the first step by the method of characteristics implies only spatial interpolation between known (f^n) values as long as the characteristics remain inside the computational domain. However, as soon as the characteristics exit this domain (which is the case at the inflow boundaries), the space interpolation has to be replaced by a time interpolation. Consequently, the resolution of this step implies prescribing any f at the inflow boundaries. This is one condition more than implies the well-posedness of the global problem (cf 3.7.1). Usually the lacking condition concerns the water depth. One could consider different ways to overcome this difficulty, for instance :

- to "freeze" the water depth (i.e. to set $h^{n+1/3} \simeq h^n$);
- to compute an approximate value of h^{n+1} by developing explicitly the continuity equation 3.76;
- to drop the characteristics method for boundary nodes and apply instead a special finite differencing to eq. 3.85, which makes use of downstream nodes ...

A catalogue of extra boundaries conditions for difference schemes approximating hyperbolic equations such as

$$\frac{\partial U}{\partial t} = a\frac{\partial U}{\partial x} \quad \text{or, more generally,} \quad \frac{\partial U}{\partial t} = \frac{\partial F}{\partial x}$$

may be found in (Shokin & Kompaniets, 1987), together with their stability requirements. From a practical point of view only, some of them appear to be handy for the inclusion in an algorithm based on the method of characteristics. However, this does not mean that they will prove equally convenient from a physical viewpoint. (Shokin & Kompaniets, 1987) illustrate this on a test case dealing with a non-viscous compressible gas. The gas is confined in a pipe bounded on the left by a solid wall and is submitted on the right side to a piston moving with constant velocity. The corresponding velocity and pressure equations are solved with the well-known Mac Cormack scheme and different treatments of the boundaries conditions are applied, with

dramatically different results. One method ranks far above the others : it consists in solving for the boundary nodes the characteristic equations derived from the full hyperbolic system (see their expression in (Abbott, 1979) for instance) and to transfer the Riemann invariants along them in order to compute the boundary values of the dependent variables. Cunge (Cunge, 1988) also considers this method to be the sole "pure" approach to the treatment of boundaries for hyperbolic equations. Yet, he points out that it is seldom applied because it complicates the algorithm.

While step 1 has consequences for the *open* boundary conditions only, the solution of step 2 implies prescribing conditions on *every* boundary, as the related operator is parabolic. Usually, the problem of open boundaries is solved by neglecting diffusion, i.e. assuming that $\vec{U}^{n+2/3} = \vec{U}^{n+1/3}$. As regards closed boundaries, (Mary, 1982) suggests working on the remaining equations of step 3 in order to study the features of $\vec{U}^{n+2/3}$ at such boundaries. Eq. 3.88 yields :

$$\vec{U}^{n+2/3} = \vec{U}^{n+1} + g\Delta t\,\nabla\zeta + \Delta t f_b \vec{u} \tag{3.89}$$

The two last terms of the right-hand side could be developed as a combination of $\vec{U}^n$, $\vec{U}^{n+1}$, ζ^{n+1} and ζ^n. In order to derive boundary conditions on $\vec{U}^{n+2/3}$, we multiply eq. 3.89 by unit vectors respectively normal and parallel to the boundary ($\vec{n}$ and $\vec{\tau}$). From 3.79 and 3.80, we deduce that

$$U_n^{n+2/3} = \vec{U}^{n+2/3}.\vec{n} = 0 \tag{3.90}$$

while

$$U_t^{n+2/3} = \vec{U}^{n+2/3}.\vec{\tau} = \vec{U}^{n+1}.\vec{\tau} + g\Delta t\,\frac{\partial\zeta}{\partial\tau} + \Delta t\, f_b \vec{U}.\vec{\tau} \tag{3.91}$$

A no slip condition on $\vec{U}$ (cf eq. 3.81) would imply that

$$U_t^{n+2/3} = g\Delta t\,\frac{\partial\zeta}{\partial\tau} \tag{3.92}$$

On the other hand, should $\vec{U}$ satisfy 3.83, we could differentiate 3.91 in the direction normal to the wall, which leads to :

$$\frac{\partial U_t^{n+2/3}}{\partial n} = aU_t^{n+1} + \Delta t\left(\frac{\partial f_b}{\partial n} + af_b\right)U_t \tag{3.93}$$

At this stage of the calculation, we do not know the variables at time t^{n+1}. Consequently, in order to reduce either eq. 3.92 or eq. 3.93 to a workable form, their right-hand sides need to be made somehow explicit. Yet, several options are available, eg. for eq. 3.93 : expressing everything as a function of variables at time t^n which leads to a Von Neumann condition; expressing everything as a function of $U_t^{n+2/3}$ which leads to a relation similar to eq. 3.83; neglecting or not f_b and its spatial derivative, etc ... Each option gives birth to a different condition and it does not seem possible to guess a priori which is best.

Similarly, splitting the resolution according to different space directions also introduce intermediate variables for which boundary conditions are needed.

3.7.3 Numerical "translation"

In brief, the implementation of boundary conditions can lead to the following problems :

1. undesirable coupling between different variables;
2. undesirable coupling between different directions;
3. need for developing special formulations at the boundary nodes.

3.7.3.1 Multivariable boundary conditions

The first problem is met with the boundary conditions on velocity. Indeed, as soon as the boundary of the domain is not parallel to one of the coordinate lines, applying conditions such as eq. 3.79, 3.90, 3.82, 3.83 or 3.91 links the two components of the depth-averaged velocity. When applying a straightforward orthogonal FDM discretization, the boundary of the studied domain is artificially compelled to follow the coordinate lines piecewise and coupling only occurs at corners. With an FEM discretization, the elements may follow more exactly the real boundary and the boundary conditions may use the "true" normals. Yet, more nodes are concerned by coupling. (nb : Should an FDM use Boundary-Fitting generalised Coordinates, it can respect as faithfully the true boundary conditions as FEM)

This problem is particularly annoying when no other terms couple each component equation. For instance, the diffusion equations described by eq. 3.86 could be solved independently for U and V were it not for the boundary conditions. If one wants to maintain this independence, alternative boundary conditions must be used, e.g. :

(a) (Hervouet & Watrin, 1988) suggests using in TELEMAC, a FEM code, the following relationships :

$$\frac{\partial U}{\partial n} = a\,U \tag{3.94}$$

$$\frac{\partial V}{\partial n} = a\,V \tag{3.95}$$

Besides, eq. 3.94 and 3.95 may be locally replaced, respectively by $U = 0$ or $V = 0$, when one boundary happens to be parallel to the $y-$ or $x-$ direction. However, in general, these boundary conditions used for the sake of numerical simplicity will not coincide with the physical boundary conditions. Indeed, the combination of 3.94 and 3.95 implies 3.83 but does not ensure that 3.79 is satisfied.

(b) When dealing with a FDM code, the solution could be to begin by neglecting the effect of diffusivity in the vicinity of corners, i.e. to postulate there $\vec{U}^{n+2/3} = \vec{U}^{n+1/3}$. These Dirichlet conditions allow us indeed to uncouple the $U-$ and $V-$ equations and solve them separately. Then, once the $\vec{U}^{n+2/3}$ are known at *interior* grid points, the estimation of $\vec{U}^{n+2/3}$ at corners could be corrected, for instance by combining 3.90 and some development of 3.93.

However, the relevance of formulations such as (a) or (b) is most probably bound to depend on the kind of flow, the mesh size, ... and can be appraised only by comparing simulations and experiments.

3.7.3.2 Space coupling

The second problem is most crucial with FDM schemes which make a widespread use of space splitting. Yet, conditions such as 3.80 introduce at corner points like M (cf. figure 3.8) a dependency between the variable values at nodes located along orthogonal coordinate lines. Once again, different approaches are possible. One would be very similar to method (b) above : postulate that $\zeta^{n+1}(M) \simeq \zeta^n(M)$, solve the system for interior nodes, then correct $\zeta^{n+1}(M)$. Another method, as discussed by Roache when dealing with the vorticity equation (Roache, 1985) (sec. III.C), would be to apply discontinuous conditions at the corner, i.e. to apply $\frac{\partial \zeta}{\partial x} = 0$ when working with coordinate line MA and use $\frac{\partial \zeta}{\partial y} = 0$ when working with coordinate line MB. Roache argued that at a geometric singularity like M there was no reason to assume there would be strict continuity or single-valuedness of a quantity like the vorticity. Such reasoning could probably apply to the free surface elevation too. However, there is a risk in doing this : that the solution will be dependent on whether the calculation proceeds first along the $x-$coordinate lines or the $y-$ ones !

Once again, there is no clear answer to the problem, apart from numerically testing the influence of the different approaches. If no data (or insufficient ones) are available for assessing quantitatively the relevance of the algorithm, it is at least possible to check whether the simulated flow pattern is coherent, with respect to our physical intuition.

3.7.3.3 Specific discretizations

We already mentioned (sec. 3.3.4) a *very* attractive feature of FEMs, namely that an appropriate manipulation of integrals leads to expressing the boundary conditions either by adding source terms to the system of equations (Von Neumann conditions) or by modifying the system coefficients (mixed conditions as 3.94 or 3.95). FDMs do not offer such facility, unless they use

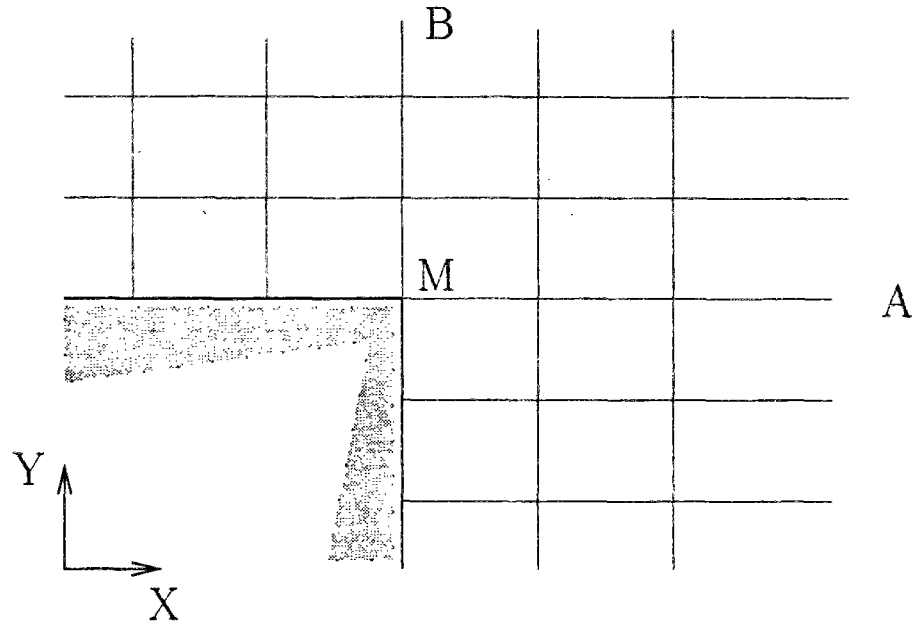


Figure 3.8: Angles of computational domain

Boundary-Fitting Coordinates (BFC).

Consequently, the discretization of boundary conditions is a further incentive to derive specific formulations (see for instance (Shokin & Kompaniets, 1987)) in order to express such quantities as $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, $\frac{\partial f}{\partial n}$ at boundary nodes in terms of interior nodes. The consequences may be to lower the order of the finite difference scheme (Roache, 1985) (sec. III.C). Besides (cf. 3.2.2), beyond formal precision, the main point is that one must ensure that the hypothesis made in deriving the boundary derivatives are consistent with the differentiation method applied at interior nodes.

3.8 Conclusion : it's a wide world !

As can be deduced from the above presentation and from a quick glance at recent congress proceedings (IAHR, 1993) or state-of-the-art papers (ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988) (part IV), there are a number of ways by which a solution technique to the flow and transport equations can be devised.

The preliminary issue is to define what an *efficient* solution technique is. An efficient, ideal model should be relevant, accurate and cost-effective. It appears that *efficiency is not a universal concept but a problem-dependent one* :

- Relevance has to do with the fact that the model portrays effectively the dominant physical processes. It depends both on the choice of the working equations and on the numerical treatment applied to each operator. As the respective importance of different processes varies according to the studied flow (cf chapter 2), we may expect the relevance of most models to be somewhat problem dependent.

- Accuracy can be appraised by different means and its evaluation cannot be carried out in an abstract or problemless context. As suggested by Ditmars (Ditmars *et al.*, 1987), applying a model to benchmark problems prior to real ones is highly advisable as it permits testing the model objectively, without the performance evaluation being blurred by troubles with field data monitoring. However, the benchmark problems need to be chosen carefully so that they allow us to check essential features of the model and not only superficial ones. This relevant choice is obviously dependent on the planned applications. Besides, accuracy depends on the quantities we choose to monitor it (reproduction of mean or extreme values, of localised or spatially averaged quantities, ability to describe long-term trends or alternatively instantaneous transient answers, ...) and its fair evaluation depends on the availability of these data.

- Cost-effectiveness has several aspects. The most obvious is associated with CPU time and storage requirements. However, as hardwares continuously progress, this point is probably less crucial than it was ten years ago. In our opinion, the most important point is the cost induced by the model development (including extensive testing). One has to decide to which extent it is worth "investing" skill, time and money in the model development, bearing in mind its possible applications.

As regards the development of a solution technique, it appears from the literature that the turning points are *the choice of a finite element or a finite difference approach* and, for scalar transport, *the treatment of the troublesome advective terms.*

The relative merits of FDM and FEM have been widely discussed for several years. Quoting Baptista (Baptista *et al.*, 1984), we may say that "it is usually recognized that

- FEM
  - rely on a strong mathematical foundation (variational analysis);
  - handle more efficiently complicated land boundaries and internal grid refinements;
  - are more consistent in the treatment of boundary conditions and in the set-up of interpolation procedures over the whole computational domain

- while FDM
  - are more intuitive to formulate and tend to require less memory capacity and CPU time, for a similar number of nodes;
  - result in significantly easier procedures concerning preparation and input of data."

In the past years, in the field of environmental hydraulics, FDM have often been preferred to FEM on the sole consideration of computational costs. However, with the development of ever more powerful computers and the design of specific algorithms taking full advantage of parallel processors (Peters, 1988; Hervouet, 1991), this advantage of FDM on FEM will probably become less significant. More important, in our opinion, is the fact that developing a *good*, up-to-date, FEM is undoubtedly more demanding than developing a FDM. For instance, it requires one to possess or develop a discretization tool when such operation can be handled manually for building FDM grids. It also requires generally a much better understanding of matrix algebra and matrix inversion problems (cf sec. 3.5).

Looking at the current evolution of the field of computational fluid dynamics (e.g. the applications in (IAHR, 1993)), it appears that the future belongs probably to finite element methods (provided they deal properly with the advective terms) in spite of the fact that the ever increasing use of flexible generalised coordinates, such as BFC, allows the FDM to compete fiercely. Whatever the future prospects, finite difference methods are presently attractive tools and provide a good introduction to the complexity of computational fluid dynamics. Pursuing the same line of thought as when selecting simply the 2D St-Venant equations as the core of our flow model and considering the possible applications and the available hardware, we decided to develop a finite difference algorithm as we judged we could achieve the development of such a tool sooner and safer than the development of a proper FEM. This is obviously an arbitrary choice. We thought that, beyond the choice of one or another approach, the relevant thing was to follow as closely as possible a proper methodology in developing the model (Ditmars *et al.*, 1987), (ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988) (part IV) and to master as much as possible the code behaviour. Consequently, a special emphasis has been put on the treatment of the troublesome advective terms (cf part II) and on benchmark testing of the whole code (cf chapters 9 and 10).

3.9 Résumé français : "Introduction aux méthodes numériques"

Les modèles d'écoulement et transport diffèrent non seulement par les équations sur lesquels ils reposent mais également par les méthodes numériques employées pour résoudre ces équations.

Dans ce chapitre, nous nous efforcerons de dresser un panorama succint des différentes méthodologies disponibles. Nous introduirons également le vocabulaire technique relatif à l'étude des schémas numériques. Notre objectif n'est pas de discuter en détail les mérites de telle ou telle méthode (ceci sera abordé dans les parties II et III de la thèse) mais uniquement de situer le cadre de nos travaux.

La dynamique des fluides est décrite par un jeu d'équations aux dérivées partielles (EDP) spatiales et temporelles. **La première étape des méthodes numériques est de les approximer, aussi fidèlement que possible, par un ensemble d'équations dont les inconnues sont les valeurs des variables hydrauliques et chimiques en un nombre fini de points** (noeuds) du domaine étudié et **en un nombre fini d'instants** durant l'évènement hydraulique considéré. L'ensemble d'équations obtenu est en particulier fonction de la localisation des noeuds, du choix des instants de calcul, c'est à dire de la **discrétisation** du domaine spatial ainsi que du temps.

- Les **méthodes aux différences finies (MDF)** (cf section 3.2.1), les premières apparues, ont été tout d'abord développées à partir d'une discrétisation orthogonale, c'est à dire que les noeuds de calcul sont répartis régulièrement le long de droites parallèles aux axes d'un repère orthonormal (à 1, 2, 3 dimensions). Toutefois, l'utilisation systématique d'une grille orthogonale serait assez restrictive : par exemple, ce n'est guère adapté à la description de domaines à la géométrie complexe. C'est pourquoi on a recours couramment à des grilles curvilignes. En fait, on effectue sur le domaine étudié un changement de coordonnées satisfaisant quelque propriété intéressante : par exemple que les frontières du domaine de calcul soient localement alignées avec les nouveaux axes de coordonnées ou encore que la densité des noeuds de calcul soit accrue dans les secteurs où l'on suspecte que les gradients des variables hydrauliques seront les plus forts. Les équations du fluide sont exprimées dans le nouveau repère. Par rapport aux équations originales, elles comprennent généralement des termes supplémentaires (dérivées des anciennes par rapport aux nouvelles coordonnées, dérivées croisées introduites puisque le repère n'est plus orthogonal, etc ...).

  En un noeud, les dérivées partielles sont approximées par des différences faisant intervenir la valeur des variables aux noeuds et en ses voisins (cf section 3.2.2) **ce en différents instants**. L'ensemble des noeuds concernés dépend du type de la dérivée partielle (premier, deuxième ordre) et du degré de finesse recherché pour l'approximation. **Les techniques les plus couramment employées pour développer les différences**

finies font appel, d'une part au développement en série de Taylor des fonctions vitesse, hauteur, concentration ...autour de chaque noeud, **et à la combinaison de ces séries, d'autre part à l'approximation locale** de ces mêmes fonctions **par des polynômes** d'ordre plus ou moins élévé, bâtis en fonction de leur valeur aux noeuds, puis dérivés.

Pour résumer, une équation gouvernant le fluide (par exemple, la conservation de la masse) se retrouve "traduite" sur le domaine étudié par autant d'équations algébriques qu'il y a de noeuds de calcul (cf sec 3.2.3). Ces équations relient la valeur "future" des variables hydrauliques en un noeud à leurs valeurs présentes et passées au noeud et en ses voisins. Si les équations aux dérivées partielles originales sont linéaires, les équations discrétisées forment un système linéaire (dont la taille est approximativement le nombre de noeuds de calcul multiplié par le nombre de variables impliquées). Si les équations originales ne sont pas linéaires, on procède généralement, préalablement à la discrétisation, à une étape de linéarisation locale.

- Les **méthodes aux éléments finis (MEF)** résultent d'une approche différente (cf 3.1), que nous allons décrire grossièrement.

  On raisonne dans l'ensemble des fonctions continues, différentiables, définies sur le domaine spatial étudié. Cet ensemble constitue un espace vectoriel de dimension infinie. Aux équations de l'hydraulique ou du transport dissous correspondent des opérateurs différentiels définis sur cet espace vectoriel, qu'annulent les fonctions solutions vitesse, hauteur d'eau, concentration, etc ...

  A la recherche des solutions "parfaites", on substitue la recherche de solutions approximatives, appartenant à un sous-espace de dimension finie, et "annulant partiellement" l'opérateur différentiel. Ce sous-espace est **engendré par un ensemble de "fonctions de base"**.

  Que veut-on dire par "annuler partiellement" ? La fonction nulle est la fonction dont le produit scalaire avec toute autre fonction est nulle. **On recherche des solutions approchées dont la transformée par l'opérateur** correspondant à l'équation **(dit le "résidu" de l'équation) ait un produit scalaire nul avec un ensemble pertinent de "fonctions test"** (dites aussi fonctions de pondération).

  Le choix des fonctions de base et des fonctions test détermine la méthode aux éléments finis (cf sec 3.3.2). Pour résumer, on dira que **la méthodologie la plus populaire est celle de Galerkin, qui consiste à choisir comme fonctions test les fonctions de base**, ces dernières formant une base orthogonale.

  Les fonctions de base sont de support réduit, i.e. elles sont nulles en dehors d'un sous-domaine restreint. Ces sous-domaines sont choisis de forme simple (triangle, quadrangle, ...), d'intersection vide, et leur juxtaposition couvre l'ensemble du domaine étudié : ce sont les

éléments qui le discrétisent (cf 3.3.1). Les fonctions ont également une forme analytique simple, généralement polynomiale.

En appliquant ce formalisme, on passe (voir 3.3.3), comme dans le cas des méthodes aux différences finies, des équations aux dérivées partielles de la dynamique des fluides à un ensemble d'équations **linéaires** (là encore, éventuellement après une étape de linéarisation des équations originelles, ou de réécriture des produits scalaires via des intégrations par partie). Les inconnues de ce système sont les coordonnées (à l'instant de calcul choisi) de la solution approchée dans le système des fonctions de base.

Les schémas numériques doivent posséder les propriétés suivantes (nb : définies ci-dessous dans le cadre des MDF) :

- la **consistence**, c'est à dire que la limite de l'approximation aux différences finies - quand pas d'espace et de temps tendent vers 0 - est effectivement l'EDP originelle.
- la **stabilité** Comme indiqué ci-dessus, l'application d'une MDF conduit à remplacer l'EDP par un système d'équations algébriques. La stabilité est la tendance à l'amortissement de toute perturbation de la solution de ce sytème (qu'il s'agisse d'erreurs d'arrondi, de "bruit" sur les conditions initiales ou aux limites, ...)
- la **convergence** Un schéma numérique est convergent s'il garantit que, au fur et à mesure que la discrétisation est affinée, sa solution tend vers la solution de l'EDP originelle.

La section 3.4 et plus encore l'annexe C sont consacrées à la description des principales techniques disponibles pour analyser a priori, analytiquement, les propriétés de consistence, stabilité et convergence des schémas numériques : méthode de Von Neumann (développement en série de Fourier et étude du facteur d'amplification - amplitude et déphasage - des séries) et méthode matricielle (étude des valeurs propres et du rayon spectral de la matrice correspondant au système d'équations issues de la MDF). Il faut cependant souligner que ces méthodes sont de portée limitée (problème des équations non-linéaires, prise en compte de l'influence des conditions limite sur la stabilité dans des cas d'étude réels ...). Si elles permettent de détecter de gros défauts, elles se doivent d'être complétées par l'analyse du comportement des schémas sur une batterie de tests (voir parties II et III du mémoire).

La section 3.6 introduit **les techniques d'éclatement d'opérateur** (dites aussi **méthodes à pas fractionnaires**) dont l'usage est trés répandu. Ces méthodes **ont pour objet de simplifier la résolution du système d'équations algébriques en le décomposant en systèmes plus simples ou présentant de meilleures propriétés de convergence.**
Dans les problèmes multidimensionnels, les opérateurs sont couramment éclatés selon les directions d'espace, ce qui permet de se ramener à la manipulation uniquement de systèmes d'équations monodimensionnelles.

Le procédé d'**éclatement selon les processus** est également fréquemment appliqué, notamment à la résolution de l'équation d'advection-dispersion (voir partie II). L'évolution des concentrations ou des variables hydrauliques est régie par des phénomènes physiques qui agissent de façon *simultanée*. Dans le cadre d'une méthode à pas fractionnaire on considère que, au sein d'un pas de temps, ces phénomènes se produisent *successivement*. Ainsi, pour calculer le nouvel état du système, on prendra en compte tout d'abord l'effet de l'advection, puis de la dispersion et - dans le cas de l'hydraulique - de la propagation et des frottements. A des processus distinctent correspondent généralement des opérateurs différentiels aux propriétés mathématiques différentes. La décomposition permet de traiter chacun indépendamment selon la technique la plus adaptée (par exemple, les méthodes aux caractéristiques pour l'advection). **Éclatement selon les processus et selon les directions d'espace peuvent bien sûr être combinés. L'éclatement d'opérateurs peut déboucher sur l'obtention de codes aisément modulables et efficients. Cependant, il complique certaines étapes, notamment la prise en compte des conditions aux limites** (cf section 3.7.2.2).

Comme indiqué plus haut, **la résolution des équations hydrauliques se ramène à celle d'un ou plusieurs** (si l'on a en particulier éclaté l'opérateur) **systèmes matriciels de grande taille, la matrice étant généralement assez creuse** (beaucoup de termes nuls) **et de largeur de bande limitée** (parce que le passage aux différences ou éléments finis relie l'évolution des variables en un noeud à son comportement en un nombre *restreint* de noeuds voisins). **Les grandes méthodes de résolution de ces systèmes matriciels sont évoquées en section 3.5 et en annexe D.**

- Dans les cas les plus favorables, on se retrouve avec des systèmes de type tridiagonal ou pentadiagonal (notamment en traitant des opérateurs de diffusion) pour lesquels existent des algorithmes d'inversion directe trés efficaces (algorithme de Thomas, cf D.1). Ce n'est malheureusement pas toujours le cas. En présence d'une matrice "quelconque", on renonce fréquemment à l'inversion directe, trop lourde tant en temps calcul qu'en place mémoire consommée, au profit de méthodes itératives.
- Le principe général des méthodes itératives est le suivant. La forme du système à résoudre est : $\underline{A}X = B$. On réécrit $\underline{A} = \underline{M} - \underline{N}$ où $\underline{M}$ est une matrice régulière, "proche" de $\underline{A}$ (par exemple au sens de la norme), mais facile à inverser. On a alors :

$$(\text{S1}) \quad \underline{A}X = B \iff (\text{S2}) \quad X = \underline{M}^{-1}\underline{N}X + \underline{M}^{-1}B$$

et on construit une suite qui approche la solution de (S2) itérativement par :

$$\begin{aligned} X^{(k+1)} &= \underline{M}^{-1}\underline{N}X^{(k)} + \underline{M}^{-1}B \\ \text{ou} \quad X^{(k+1)} &= X^{(k)} - \underline{M}^{-1}R^{(k)} \end{aligned}$$

où $R^{(k)}$ est le vecteur des résidus de l'équation à la $k$eme itération, soit $R^{(k)} = \underline{A}X^{(k)} - B$.

Les méthodes de Jacobi, Gauss-Seidel ou sur-relaxation (cf D.2) correspondent à différents choix de $\underline{M}$ (matrice diagonale ou diagonale inférieure extraite de $\underline{A}$, combinaison des deux, ...). On indique en D.3 des techniques qui permettent d'accélérer leur convergence.

- Parmi ces méthodes, on trouve les méthodes de Gradient (sections 3.5.1.2 et D.3.1), qui s'appliquent dans le cas où $\underline{A}$ est symétrique, définie, positive. En effet, l'expression $\underline{A}X - B$ peut alors être interprétée comme le gradient de la fonctionnelle quadratique $\mathcal{J}(X) = \frac{1}{2} X^t \underline{A} X - B^t X$ et l'annuler revient à chercher le minimum de $\mathcal{J}$. Pour trouver ce minimum de façon optimale, il faut à chaque itération suivre la direction du gradient local.
- Des méthodes itératives appropriées aux systèmes d'équations originellement non-linéaires (méthodes de Newton et quasi-Newton) sont brièvement présentées en 3.5.2 et D.5.
- Enfin, on notera que la rapidité de la convergence des méthodes itératives évoquées ci-dessus dépend de leur rayon spectral (norme de la plus grande valeur propre). Afin d'améliorer celui-ci, on peut choisir de substituer à la résolution de $\underline{A}X = B$ celle de $\underline{C}.\underline{A}X = \underline{C}.B$, $\underline{C}$ étant une matrice de *conditionnement* telle que le produit $\underline{C}.\underline{A}$ a un meilleur rayon spectral que $\underline{A}$. Les choix possibles pour $\underline{C}$ sont discutés en 3.5.1.3 et D.3.1.

Ce chapitre se termine par une sensibilisation aux problèmes posés par le traitement des conditions aux limites (section 3.7), **notamment l'insuffisance des conditions**, qu'elle soit réelle (i.e. due au manque de données de terrain) ou induite par la méthode de résolution (nécessité de dériver des conditions limite pour les étapes de calcul intermédiaires dans le cas d'application d'une technique à pas fractionnaires) **et le soin à apporter à leur "traduction numérique"** (perte de précision du schéma au voisinage des frontières, problèmes de couplage indésirables entre différentes variables et directions de l'espace aux "angles" du domaine de calcul, etc ...).

En conclusion, il apparaît tout à la fois que l'éventail des méthodes numériques disponibles pour la résolution des équations d'hydraulique et de transport est vaste mais également que le chemin est pavé d'embûches. **Nous avons opté a priori pour le choix de techniques de résolution aux différences finies, à pas fractionnaires, reposant sur l'éclatement des opérateurs en espace et selon les processus**, estimant qu'il nous serait possible de développer ainsi un code à moindre coût (de développement et fonctionnement). **Ce choix peut sembler arbitraire. Nous avons porté notre effort sur sa validation à l'aide d'une méthodologie rigoureuse, basée sur des tests intensifs pour la sélection des algorithmes de calcul. Ce travail fait l'objet des parties II et III du mémoire.**

Notation

- t time
- x, y, z coordinates (the directions x, y, z correspond respectively to the x_1, x_2, x_3 directions in tensor notations)
- U, V, W (also denoted U_1, U_2, U_3 in tensor notations) temporal mean velocities in x, y, z directions
- u, v, w (also denoted u_1, u_2, u_3 in tensor notations) turbulent fluctuating deviations from these mean values
- $\overline{U}$, $\overline{V}$ depth-averaged velocities in x and y directions
- P temporal mean of static pressure
- p fluctuating pressure
- ϕ temporal mean of scalar (concentration or temperature), φ the associated turbulent fluctuation
- $\overline{\phi}$ depth-averaged concentration or temperature
- g magnitude of gravitational acceleration
- ρ fluid density (the subscript r denotes a reference value, e.g. in a river with respect to a discharge)
- z_b bed elevation, with respect to a reference level
- h water depth
- $\zeta = h + z_b$ free surface elevation, with respect to a reference level
- q_s heat or mass flux (source or sink term) through surface
- τ_b bottom shear stress (τ_{bx} and τ_{by} its components, respectively along he $x-$ and $y-$ directions)

- U_* bed friction velocity, defined by $\tau_b = \rho U_*^2$
- C_f friction coefficient. When the friction is expressed by a quadratic law, τ_b modulus is computed by the relationship $\tau_b = \rho C_f \left(\overline{U}^2 + \overline{V}^2\right)$. C_f depends on the bed roughness.
- τ_s surface shear stress
- $\bar{\tau}_{xx}$ ($\bar{\tau}_{xy}$, $\bar{\tau}_{yy}$) depth-averaged value of the turbulent stress $-\rho\overline{u.u}$ (respectively $-\rho\overline{u.v}$ and $-\rho\overline{v.v}$).
- $\bar{J}_x$ ($\bar{J}_y$) depth-averaged value of the turbulent heat or mass fluxes ($-\rho\overline{\varphi.u}$ and $-\rho\overline{\varphi.v}$), respectively in the $x-$ and $y-$ directions
- ν and λ respectively kinematic molecular viscosity and molecular diffusivity
- ν_t and Γ_t respectively turbulent eddy viscosity and eddy diffusivity
- σ_t turbulent Prandtl or Schmidt number
- $\widehat{V}$ turbulence velocity scale
- L turbulence length scale
- l_m Prandtl mixing length
- κ Von Karman constant
- k kinetic turbulent energy per unit mass
- ϵ dissipation rate of k
- C_μ, $C_{1\epsilon}$, $C_{2\epsilon}$, $C_{3\epsilon}$, σ_k, σ_ϵ empirical constants related to the $k-\epsilon$ model
- C_k, C_ϵ, $C_{\Gamma\epsilon}$ empirical constants in the depth-averaged $k-\epsilon$ model
- B river width

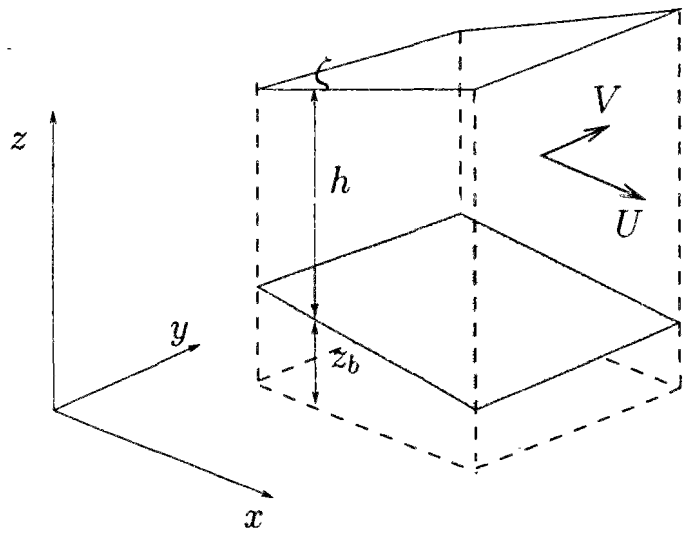


Figure 3.9: Main flow variables

Part II

Numerical solution of advection – dispersion equation

Foreword to part II

As discussed in chapter 2, scalar transport is generally described by the advection-dispersion equation which includes a hyperbolic term describing transport by the mean flow and a parabolic term which is a diffusion operator accounting for turbulent mixing, dispersion related to vertical or cross-sectional non-uniformities, etc ...This equation can alternatively be expressed in conservative form or developed into a non-conservative form by taking advantage of the flow field properties (i.e. its continuity equation).

One of the most challenging tasks in the field of numerical analysis is to solve properly this scalar transport equation in convection dominated flows, i.e. when the hyperbolic term is great with respect to the diffusive term.

In our quest for a convenient method which combines accuracy and cost-effectiveness, we have been going through the following steps :

1. Several approaches to the problem of solving the transport equation have been suggested : their principles are summarized in chapter 4.
2. We have been selecting *a priori* a few algorithms derived from one or another approach. They are described in chapter 4 and the reasons for their selection are indicated in 4.3.
3. Then, we have been assessing and comparing the performance of these schemes :
   (a) The Fourier analysis of each selected algorithm is undertaken and commented in chapter 5.
   (b) Then, the algorithms have been applied to recommended test cases, both in one-dimensional (chapter 6) and two-dimensional situations (chapter 7).
4. This led us to a choice which is discussed in section 7.5.

(nb : We have not included in this review the study of purely lagrangian schemes.
These proceed as follows : a pollutant cloud is divided in a number of cells, each one including some mass of pollutant; the displacement of these cells in the flow is computed in a manner similar to the computation of trajectories in backward characteristics methods (sec. 4.1.2); dispersion is modelled by random displacements which add to the displacement induced by the main flow. Whenever one needs to estimate the concentration at a given location, it is necessary to look first for the cells surrounding this location; then, concentration is obtained through a spatial averaging. Examples of application are given in (Mouchel, 1990; Crockett *et al.*, 1989) (one-dimensional case) and (Józsa, 1989; Leclerc & Boudreault, 1993) (two-dimensional case).
Use of such methods theoretically limits numerical diffusion as interpolation is applied only when it is

really needed. However, while it is sufficient to divide the pollutant cloud in a few hundreds cells only in order to forecast the trend of pollutant displacement, it becomes necessary to use several thousands cells to obtain detailed concentration fields (Józsa, personal communication). This can make the algorithm expensive. Besides, the solution of the flow equations requires anyway the use of a fixed computational grid. Consequently, it is not adequate to solve the advection terms of the flow equations with a purely Lagrangian algorithm. At the contrary, advection algorithms such as backward characteristics methods can be applied both to hydraulic and bioegeochemical variables in a manner which implies computational savings.

For the above reasons (computational cost, unsuitability for dealing with flow variables), we discarded the study of Lagrangian algorithms.

Avant-propos pour la partie II : "Résolution numérique de l'équation d'advection-dispersion"

Comme indiqué en conclusion du chapitre 2 (section 2.4.5), nous avons choisi, dans l'optique notamment d'applications à la simulation des écoulements et pollutions en Seine, de baser notre modèle d'hydraulique et transport sur les équations bidimensionnelles de St-Venant et d'advection-dispersion. Nous avons également choisi de rechercher des techniques de résolution optimales de ces équations dans la famille des méthodes aux différences finies à pas fractionnaires (cf chapitre 3, section 3.8).

Les équations de St-Venant et de transport de soluté ont des termes communs, ceux d'advection et dispersion. C'est pourquoi, quoique dans la pratique le calcul des écoulements précède celui des concentrations, nous nous intéresserons en premier lieu, dans cette partie II, à la résolution de l'advection-dispersion.

A l'advection correspond un opérateur mathématique hyperbolique, à la dispersion un opérateur parabolique. Il est de longue date reconnu qu'un des problèmes d'analyse numérique les plus délicats est celui de la résolution de l'advection-dispersion quand la convection - le terme hyperbolique - est dominante.

Dans notre quête d'une méthode qui combine précision et efficacité informatique, nous avons procédé comme suit :

1. Le chapitre initial (ch. 4) est consacré à une bibliographie des démarches développées à ce jour.

2. En nous basant sur cette étude bibliographique, nous avons sélectionné *a priori* quelques algorithmes relevant de l'une ou l'autre de ces démarches (voir ch. 4 et plus particulièrement la section 4.3).

3. Nous avons ensuite comparé les performances de ces différents algorithmes :

   (a) à l'aide d'une analyse de Fourier classique (ch. 5);

   (b) à l'aide de cas tests recommandés dans la littérature, en mono- (ch. 6) et bi-dimensionnel (ch. 7).

4. D'où le choix, les stratégies de résolution, proposés en section 7.5.

Chapter 4

Review of available methods

We recall first the different forms of the transport equation. Basically the conservative form reads (in tensor notations) :

$$\frac{\partial \psi}{\partial t} + \frac{\partial u_j \psi}{\partial x_j} = S \tag{4.1}$$

where u_j denotes the flow velocity along the coordinate direction j, S is a diffusion term and ψ is respectively equal to C, hC and AC in three, two and one-dimensional cases, where C stands for the scalar concentration, h for the water-depth and A for the wet section. On the other hand, the non-conservative formulation reduces to :

$$\frac{\partial C}{\partial t} + u_j \frac{\partial C}{\partial x_j} = S' \tag{4.2}$$

For the sake of simplicity, we shall introduce the different approaches with respect to the case of pure advection in a one-dimensional situation and indicate but briefly how they extend to multi-dimensional cases. As regards the treatment of the diffusion operator, this is indicated in Appendix E, section E.2.

The domain under study is discretised as is usual with FDM (cf 3.2.1). x_i denotes the coordinate of node number i. Δt is the time step. f_i^n refers to the value of dependent variable f at time $t^n = t_0 + n\Delta t$ (t_0 being the initial time).

First, we shall distinguish between methods applying to the solution of the non-conservative form of the transport equation and methods dealing with its conservative form.

4.1 Non - conservative form of the transport equation

Methods based on the non-conservative form of the transport equation mostly follow a lagrangian approach. In fact, it is straightforward to check that the pure advection equation

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} = 0 \tag{4.3}$$

is equivalent to

$$\frac{\mathrm{D}C}{\mathrm{D}t} = 0 \tag{4.4}$$

where $\frac{\mathrm{D}}{\mathrm{D}t}$ represents the total derivative along the trajectory defined by:

$$\frac{\mathrm{D}\vec{X}}{\mathrm{D}t} = \vec{U} \tag{4.5}$$

with $\vec{X}$ and $\vec{U}$ denoting respectively the coordinates and flow velocity vectors. The trajectories satisfying eq. 4.5 are termed the flow "characteristics". From eq. 4.4, it follows that *a solution of the pure - advection equation is constant along every characteristic.*

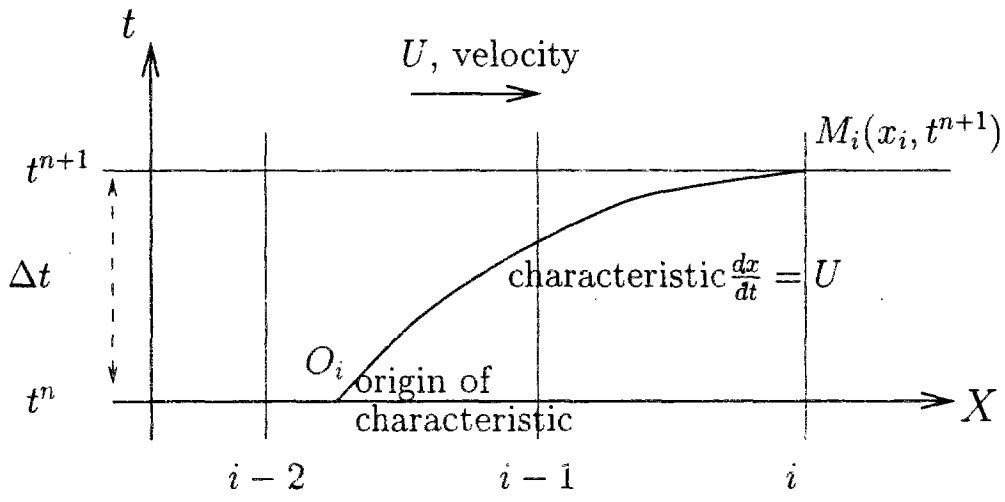


Figure 4.1: Characteristic lines & related definitions

Knowing the concentration field at time t^n, the above property can be used to predict the concentration values at a future time instant. Indeed,

$$C\left(x_i,\, t^{n+1}\right) = C\left(M_i,\, t^{n+1}\right) = C\left(O_i,\, t^n\right) = C\left(x_i - dl\,,\, t^n\right) \text{ with } dl = \int_{t^n}^{t^{n+1}} U(x,t)\,dt \tag{4.6}$$

If the flow field is uniform, dl is independent from node i and reduces to $U\,\Delta t$. The feet of characteristics (i.e. the O_i-points) do not generally coincide with grid nodes so that $C\left(O_i,\, t^n\right)$ must be interpolated from the C_i^n.

4.1.1 Lagrangian polynomial fitting

These methods were originally developed for uniform velocity and grid spacing (Crowley, 1968), with the condition $U\Delta t \leq \Delta x$, so that, assuming the velocity is positively oriented along the $x-$ coordinate axis, O_i belongs to interval $[x_{i-1}, x_i]$. In order to allow for the determination of concentration at O_i points, the concentration field is locally approximated by polynomial fitting. A set of $m+1$ data points is needed for constructing a polynomial of order m. If the polynomial is even-ordered, $m/2$ points will be located upstream of node i and another $m/2$ points downstream. If the polynomial is odd-ordered, the extra-point is added on the upstream side of i. Let P_i denote the Lagrangian polynomial built around node i. The advection scheme becomes $C_i^{n+1} = P_i[x_i - U\Delta t]$ which we would like to express as

$$C_i^{n+1} = \sum_j a_j C_{i+j}^n \qquad \text{with } j = 0, \pm 1, \pm 2, \ldots \tag{4.7}$$

The coefficients a_j depend on the grid spacing and on the velocity. They can be determined from a Taylor series expansion of relation 4.7 which yields :

$$\sum_{k=0}^{\infty} \frac{(\Delta t)^k}{k!} \frac{\partial^k C}{\partial t^k}_{|i} = \sum_j a_j \left[\sum_{k=0}^{\infty} \frac{(x_{i+j} - x_i)^k}{k!} \frac{\partial^k C}{\partial x^k}_{|i} \right] \tag{4.8}$$

Reorganizing 4.8 gives :

$$C_i^n \left(1 - \sum_j a_j\right) + \sum_{k=0}^{\infty} \frac{1}{k!} \left[(\Delta t)^k \frac{\partial^k C}{\partial t^k}_{|i} - \sum_j a_j (x_{i+j} - x_i)^k \frac{\partial^k C}{\partial x^k}_{|i} \right] = 0 \tag{4.9}$$

Using eq. 4.3 and assuming U constant, we obtain that $\frac{\partial^k C}{\partial t^k} = (-U)^k \frac{\partial^k C}{\partial x^k}$, which allows to replace the temporal derivatives in eq. 4.9 : it appears that, in order to fulfill 4.9, the a_j must satisfy the set of equations

$$\sum_j (x_{i+j} - x_i)^k a_j = (-U\,\Delta t)^k \qquad \text{for } k = 0, 1, \ldots \tag{4.10}$$

When looking for a $m-$order polynomial, we shall use eq. 4.10 for $k = 0, \ldots m$ only. For the special case of a uniform and steady flow field, the resulting approximation is $m-$order accurate in space and time, i.e. by construction, all truncation errors of order up to and including m are zero. If the velocity varies in space or time, the scheme is only first-order accurate in time (Crowley, 1968).

With uniform velocity and grid spacing, the a_j turn out to be polynomial functions of the **Courant number** $c_r = U\Delta t/\Delta x$. The complete derivation of schemes of order 1 through 10 in that case is given in (Tremback *et al.*, 1987). If the grid spacing is variable, the a_j values are dependent on the node i. If the velocity field is unsteady, the a_j values must be updated at each time instant.

A Von Neumann stability analysis (cf C.4) of these schemes can be found in (Takacs, 1985) (up to fourth-order) and (Tremback *et al.*, 1987) (up to tenth-order). As could be expected from their construction, all schemes have the so-called Courant-Friedrich-Levy stability criterion, namely $c_r \leq 1$. Both authors conclude that even-order schemes are most effective in reducing dissipation errors while odd-order schemes reduce phase errors more effectively . As summed up in (Tremback *et al.*, 1987), *an even-order scheme decreases the amplitude error of the next lower-order odd scheme, but phase errors are slightly increased. Conversely, an odd-order scheme will improve phase accuracy but increase amplitude errors over the next lower-order even scheme.*

(Takacs, 1985) introduces polynomial schemes according to a slightly different approach and studies their performance up to fourth-order. He suggests that phase errors are more troubling than amplitude errors and consequently advises the use of the third-order scheme, whose expression for uniform velocity and grid spacing is :

$$\begin{aligned} C_i^{n+1} &= C_i^n - \frac{c_r}{2}\left(C_{i+1}^n - C_{i-1}^n\right) + \frac{(c_r)^2}{2}\left(C_{i+1}^n - 2C_i^n + C_{i-1}^n\right) \\ &+ \frac{c_r}{6}\left(1-c_r\right)\left(1+c_r\right)\left(C_{i+1}^n - 3C_i^n + 3C_{i-1}^n - C_{i-2}^n\right) \end{aligned} \tag{4.11}$$

Takacs's distinctive contribution is to devise a two-step predictor-corrector sequence which boils down to 4.11 for constant U and Δx while lending itself to a generalization to non-uniform and non-positive flows more easily than the method which would consist of using 4.7 in combination with solving system 4.10. This algorithm has a flux form (see sec. 4.2 below) and reads as follows :

1. **Predictor step**

$$C_i^* = C_i^n - \left[\, F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} \,\right] \tag{4.12}$$

$$\text{where} \quad F_{i+\frac{1}{2}} = (c_r)^+_{i+\frac{1}{2}}\, C_i^n + (c_r)^-_{i+\frac{1}{2}}\, C_{i+1}^n \tag{4.13}$$

$$(c_r)_{i+\frac{1}{2}} = \frac{(U_{i+1}+U_i)\,\Delta t}{2\,(x_{i+1}-x_i)}$$

$$(c_r)^+_{i+\frac{1}{2}} = \left(\frac{c_r + |c_r|}{2}\right)_{i+\frac{1}{2}} \quad \text{and} \quad (c_r)^-_{i+\frac{1}{2}} = \left(\frac{c_r - |c_r|}{2}\right)_{i+\frac{1}{2}}$$

2. **Corrector step**

$$C_i^{n+1} = C_i^n - \frac{1}{2}\left[\, P_{i+\frac{1}{2}} - P_{i-\frac{1}{2}} \,\right] + \left[\, \alpha_{i+\frac{1}{2}} R_{i+\frac{1}{2}} - \alpha_{i-\frac{1}{2}} R_{i-\frac{1}{2}} \,\right] \tag{4.14}$$

$$\text{where } P_{i+\frac{1}{2}} = (c_r)^+_{i+\frac{1}{2}}\left(C_{i+1}^* + C_i^n\right) + (c_r)^-_{i+\frac{1}{2}}\left(C_i^* + C_{i+1}^n\right) \tag{4.15}$$

$$\begin{aligned} R_{i+\frac{1}{2}} &= \left[(c_r)^+_{i+\frac{1}{2}}\left(C_{i+1}^* - C_i^n\right) - \widehat{(c_r)}^+_{i+\frac{1}{2}}\widehat{(c_r)}^+_{i-\frac{1}{2}}\left(C_i^* - C_{i-1}^n\right)\right] \\ &- \left[(c_r)^-_{i+\frac{1}{2}}\left(C_{i+1}^n - C_i^*\right) - \widehat{(c_r)}^-_{i+\frac{1}{2}}\widehat{(c_r)}^-_{i+\frac{3}{2}}\left(C_{i+2}^n - C_{i+1}^*\right)\right] \end{aligned} \tag{4.16}$$

$$\text{with} \quad \alpha_{i+\frac{1}{2}} = \left(\frac{1+|c_r|}{6}\right)_{i+\frac{1}{2}} \quad \text{and} \quad \widehat{(c_r)}^{\pm}_{i+\frac{1}{2}} = \sqrt{|c_r|^{\pm}_{i+\frac{1}{2}}}$$

The introduction of geometric means of the velocity field $\widehat{(c_r)}$ in the $R_{i+\frac{1}{2}}$ term of the corrector step is driven by the need to guarantee stability of the scheme for non-uniform flow fields.

Takacs points out that there is no unique generalization to multi-dimensional flows for the above scheme. He nevertheless favours the construction of two-dimensional schemes by applying successively one-dimensional operators along each coordinate direction. He furthermore points out that, in all two-dimensional tests he performed, the overall damping nature of the scheme allows a simpler expression of $R_{i+\frac{1}{2}}$ to be used, namely

$$R_{i+\frac{1}{2}} = (c_r)^{+}_{i+\frac{1}{2}} \left[(C^*_{i+1} - C^n_i) - (C^*_i - C^n_{i-1})\right] - (c_r)^{-}_{i+\frac{1}{2}} \left[(C^n_{i+1} - C^*_i) - (C^n_{i+2} - C^*_{i+1})\right] \quad (4.17)$$

(Tremback *et al.*, 1987) restrict their explicit development and trials of the polynomial schemes to constant grid spacing cases. They investigate how these schemes can be written in flux form (see below) and similarly use space-splitting to adapt the schemes to multi-dimensional problems. Two kinds of 2D tests are performed, respectively with a rotational (cf chapter 7) and deformational flow field. Besides accuracy, (Tremback *et al.*, 1987) also assess the relative computational cost of each scheme (nb : all schemes have been vectorised and run on a corresponding vector-based computer). They conclude that a sixth-order scheme achieves the best balance between accuracy and efficiency.

4.1.2 Backward characteristics methods

The extension of polynomial fitting to non-uniform flows is not straightforward since, if the flow field is fast-varying, the computation of distance dl (cf eq. 4.6) by some formula like $dl = U_i \Delta t$ or $dl = U_{i-1/2} \Delta t$ can become quite erroneous. Besides, as we shall see later on, the solution of 4.10 for variable grid spacing may become expensive. Although they have been developed independently, backward characteristics methods can be viewed as a flexible extension of lagrangian polynomial methods.

The location of the foot O_i of the characteristic line is determined by integrating properly eq. 4.5 backwards in time. This is usually done using a classical Runge-Kutta method. Detailed presentations of this approach may be found in (Wang *et al.*, 1988) (single - step fourth-order Runge Kutta method) or (Baptista *et al.*, 1984; Hervouet, 1986) (multiple-step algorithms using respectively fourth and second-order Runge Kutta methods within each step). The specific algorithm used in all our tests is given in Appendix E (section E.1.1). An error in the position of O_i acts like an extra dispersion term.

The Runge-Kutta integration method extends straightforwardly to multi-dimensional cases. Although it may look simpler, splitting the calculation of the trajectory along the different coordinate lines (i.e. computing first the displacement induced by velocities parallel to the $x-$ direction, then moving along the $y-$ direction) is not advisable as it may lead quickly to serious position errors unless the Courant number is kept very small (Williamson & Rasch, 1989).

What really differs from one characteristic method to another is the chosen interpolator. Using a linear one induces an unacceptable amount of dissipation. It has been proposed (e.g. (Schohl & Holly, 1991)) to use spline functions in order to achieve the interpolation. However, defining the spline functions usually requires solving some linear system involving all nodes of the computational domain. While this may be feasible and moderately costly in one-dimensional situations, it can prove really cumbersome in multidimensional cases, especially when the studied domain is irregular. Consequently, while spline methods can yield accurate results, we have been focusing on locally defined interpolators.

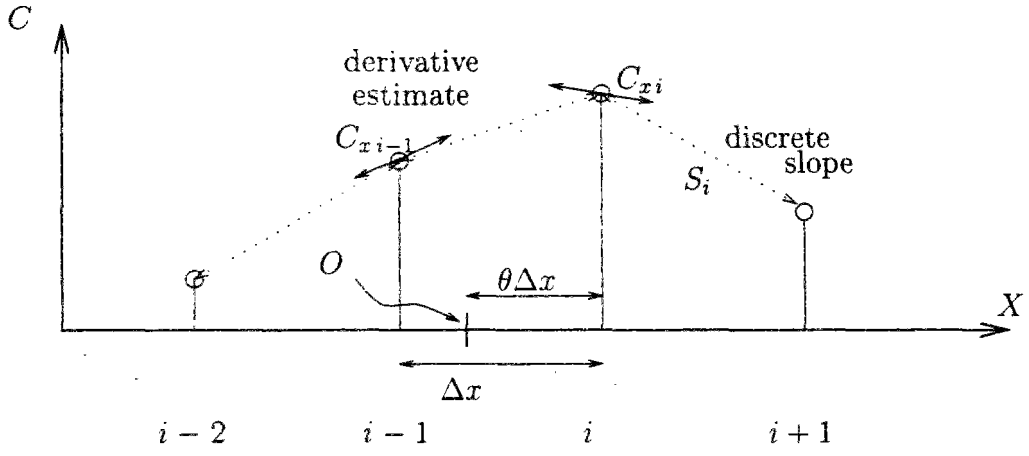


Figure 4.2: Notation for interpolating forms

4.1.2.1 Hermite cubic interpolant

One favorite interpolating form is the Hermite cubic interpolant, which reads on interval $[\, x_{i-1},\, x_i \,]$ as (cf fig. 4.2) :

$$P_i(O) = a\,C_{i-1} + b\,C_i + d\,C_{xi-1} + e\,C_{xi} \qquad (4.18)$$

$$\text{with } \Delta x = x_i - x_{i-1} \quad , \quad \theta = (\, x_i - x(O)\,)/\Delta x$$

$$\text{and } \quad a = \theta^2\,(3 - 2\theta) \quad , \quad b = 1 - a$$

$$d = +\theta^2\,(1-\theta)\,\Delta x \quad , \quad e = -\theta\,(1-\theta)^2\,\Delta x$$

If velocity and grid spacing are uniform, θ coincides with the decimal part of the Courant number c_r defined above. In 4.18, we see that the interpolant is defined in terms of the data C_i and the derivative estimates C_{xi} at the endpoints of the interval. Consequently, it depends on how the C_{xi} are evaluated.

Holly-Preissman scheme Differentiating equation 4.3 in space yields :

$$\frac{\partial C_x}{\partial t} + U\frac{\partial C_x}{\partial x} = -\frac{\partial U}{\partial x}.C_x \qquad \text{where } C_x = \frac{\partial C}{\partial x} \tag{4.19}$$

It appears that if C is a solution of 4.3, its derivative satisfies a similar transport equation with an additional source term $-\frac{\partial U}{\partial x}.C_x$. The (Holly & Preissmann, 1977) algorithm relies on this property : $C_x(x,t)$ is computed by solving 4.19 with a backward characteristics method as for C. The interpolating form is obtained by differentiating the Hermite cubic interpolant :

$$\begin{aligned} C_x(O) &= f\,(C_i - C_{i-1}) + g\,C_{xi-1} + h\,C_{xi} \qquad (4.20)\\ \text{with } f &= 6\theta\,(1-\theta)\,/\,\Delta x\ ,\ g = \theta(3\theta-2) \text{ and } h = (1-\theta)\,(1-3\theta) \end{aligned}$$

The Runge-Kutta methods mentioned above break the total trajectory from O_i to M_i into a series of segments. The right-hand side corrective term in eq. 4.19 is integrated along the trajectory by applying the trapezium rule to each segment (Holly & Usseglio-Polatera, 1984).

Whenever C obeys some transport equation with non-zero dispersion, the evaluation of C_x is further corrected by applying to it the same amount of dispersion (Holly & Usseglio-Polatera, 1984).

The major drawback of this method, whose remarkable accuracy is demonstrated in (Holly & Preissmann, 1977; Holly & Usseglio-Polatera, 1984; Usseglio-Polatera, 1988) and subsequently in this report, is that *it adds explicitly all first-order derivatives of each transported scalar to the set of studied dependent variables.* Specific questions arise as regards the adequate prescription of derivative values at the inflow boundaries.

Rasch-Williamson scheme (Williamson & Rasch, 1989; Rasch & Williamson, 1990) studied a large number of interpolation schemes obtained by combining interpolating forms (three types, among them the Hermite cubic interpolant), derivative estimates (seven methods) and modification of these estimates in order to obtain *shape-preserving interpolants* i.e. which maintain the monotonicity and/or convexity implied in the discrete data set. The selection of derivative estimates and shape-preserving criteria is based notably on previous work by (Fritsch & Carlson, 1980; Hyman, 1983).

The relative performance of each scheme is assessed with respect to interpolation of test shapes (Williamson & Rasch, 1989), one-dimensional (Williamson & Rasch, 1989) and

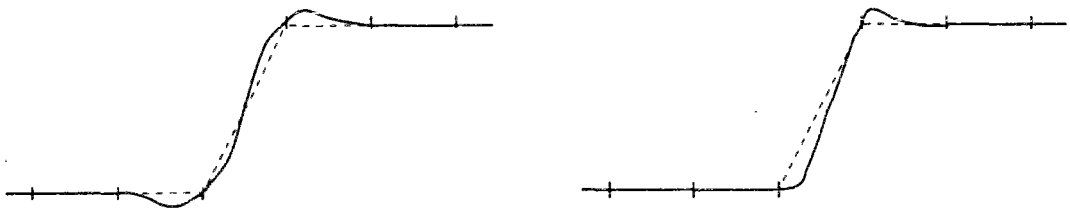
two-dimensional (Rasch & Williamson, 1990) transport of narrow distributions displaying strong gradients. From this thorough analysis, it follows that the best derivative estimate to combine with the Hermite interpolant is the **Akima estimate** (cf details in (Akima, 1970)) which reads :

$$C_{xi} = \begin{cases} (\alpha S_{i-1} + \beta S_i) / (\alpha + \beta) & \text{if } \alpha + \beta \neq 0 \\ (S_{i-1} + S_i) / 2 & \text{if } \alpha + \beta = 0 \end{cases} \tag{4.21}$$

S_i denotes the discrete slope (first-order one-sided approximation of the derivative) $S_i = (C_{i+1} - C_i) \ / \ (x_{i+1} - x_i)$ and the weights $\alpha/(\alpha+\beta)$ and $\beta/(\alpha+\beta)$ are inversely proportional to an approximation to the curvature on each side of node i:

$$\left.\begin{array}{rcl} \alpha & = & |S_{i+1} - S_i| \\ \beta & = & |S_{i-1} - S_{i-2}| \end{array}\right\} \tag{4.22}$$

Figure 4.3 illustrates the difference between the Akima estimate and a simple arithmetic estimate (namely $C_{xi} = (S_i + S_{i-1})/2$) : it represents the Hermite cubic interpolant which would correspond to each estimate while the data set defines a sharp front. The Akima estimate is seen to induce slightly more overshoot but much less undershoot (i.e. negative concentrations) at the base of the front.



Arithmetic estimate Akima estimate

Figure 4.3: Hermite cubic approximation of a sharp front

The different alternatives for extending the Hermite cubic interpolant (and, incidentally, the Holly-Preissmann and Rasch-Williamson schemes) to two-dimensional cases are detailed in E.1.2.

4.1.2.2 Shape-preserving cubic interpolant

Figure 4.3 illustrates also a typical drawback linked to the use of high-order polynomial interpolants. While they are supposed to be accurate, they do not systematically preserve features of the scalar distribution such as monotonicity and positiveness and can raise spurious extrema

(minima or maxima as seen on fig. 4.3). The Hermite cubic interpolant has been thoroughly studied (Fritsch & Carlson, 1980; Hyman, 1983). It follows that, in order for the interpolant P_i (cf eq. 4.18) to reflect the features of the data set, the derivative estimates C_{xi} at grid nodes must fulfill some conditions.

We recall that the scalar distribution C is defined at discrete locations $\{x_i\}$, $i = 1, \mathcal{N}$ and that the discrete slope S_i is $(C_{i+1} - C_i)$ / $(x_{i+1} - x_i)$. On the double grid interval $[\, x_{i-1}, x_{i+1} \,]$, the data are locally monotonic if $S_{i-1}\, S_i > 0$, locally convex if $S_{i-1} > S_i$, locally concave if $S_{i-1} < S_i$. The corresponding conditions on the derivative estimates are (Rasch & Williamson, 1990) :

Necessary Condition for Convexity/Concavity,

$$\text{on } [\, x_{i-1}, x_i \,] \;:\; (C_{x\,i-1} - S_{i-1})\,(S_{i-1} - C_{xi}) > 0 \qquad (4.23)$$

$$\text{on } [\, x_{i-1}, x_{i+1} \,] \;:\; (C_{xi} - S_{i-1})\,(S_i - C_{xi}) > 0 \qquad (4.24)$$

Necessary Condition for Monotonicity,

$$\text{on } [\, x_{i-1}, x_i \,],\ \text{if } S_{i-1} \neq 0 \quad \text{sign}(C_{x\,i-1}) = \text{sign}(S_{i-1}) = \text{sign}(C_{xi})$$

$$\text{if } S_{i-1} = 0 \quad C_{x\,i-1} = C_{xi} = 0 \qquad (4.25)$$

$$\text{on } [\, x_{i-1}, x_{i+1} \,],\ \text{if } S_{i-1}\, S_i > 0 \quad \text{sign}(S_{i-1}) = \text{sign}(C_{xi}) = \text{sign}(S_i)$$

$$\text{if } S_{i-1}\, S_i \leq 0 \quad C_{xi} = 0 \qquad (4.26)$$

The *necessary and sufficient* condition for monotonicity (Fritsch & Carlson, 1980) is rather complex and one may prefer to use a simpler, sufficient albeit not necessary, condition (Rasch & Williamson, 1990) which reads :

$$\text{for monotonicity on } [\, x_{i-1}, x_i \,], \quad 0 \leq C_{xj}/S_{i-1} \leq 3 \ \text{ for } j = i,\, i-1 \qquad (4.27)$$

$$\text{on } [\, x_{i-1}, x_{i+1} \,], \quad 0 \leq C_{xi}/S_j \leq 3 \ \text{ for } j = i,\, i-1 \qquad (4.28)$$

It appears that the constraints the C_x need to satisfy are functions of the discrete slopes. Each node belongs to two intervals. Prior to using P_i, one may modify C_{xi} and $C_{x\,i-1}$ so that they respect the data set features on $[\, x_{i-1}, x_i \,]$ uniquely (e.g obey conditions such as 4.23, 4.25, 4.27). Then, independent corrections may be applied to C_{xi} and $C_{x\,i-1}$ when working respectively with P_{i+1} and P_{i-1}. The resulting piecewise interpolant is only continuous (we say it is C^0). Alternatively, a derivative estimate C_{xi} may be forced to satisfy *simultaneously* the constraints associated with the surrounding intervals and the same value of the estimate is used for interpolation on adjacent intervals (conditions 4.24, 4.26, 4.28). In that case, both the interpolant and its derivative are continous (we say it is C^1). The latter choice is obviously more restrictive.

In particular, at an extremum where the data are not monotonic over the surrounding double interval, limiting according to 4.26 provides a severe restriction : C_{xi} is constrained to be zero and thus, the interpolant puts the extremum on grid node i. Physical extrema may occur between nodes : as they disregard that fact, restrictions 4.26 and 4.28 can induce serious damping (also termed clipping). In order to avoid this undesirable behaviour, Hyman (Hyman, 1983) suggested relaxing 4.26 and 4.28 in the vicinity of extrema : in such case, C_{xi} is allowed to keep its original sign but its amplitude is constrained in order to provide control of possible overshoots on the neighbouring intervals. The Hyman limiter (to apply instead of 4.26 and 4.28) reads :

$$C_{xi} \leftarrow \text{sign}(C_{xi}) \min\left(|C_{xi}|, 3|S_{i-1}|, 3|S_i| \right) \tag{4.29}$$

Extensions of the above conditions and limiters to the two-dimensional case can be found in (Williamson & Rasch, 1989). The various trials reported in (Williamson & Rasch, 1989; Rasch & Williamson, 1990) lead to the conclusion that the relevance of shape-preserving interpolants depends on the problem studied. If ensuring positiveness and mass conservation of the transported scalar are essential to the modeller, he may indeed find it beneficial to limit derivatives. On the other hand, if one is mainly interested in forecasting accurately extreme values, it seems preferable to relax any constraint.

4.1.2.3 Other methods

Taylor Series Combination (Dan N'Guyen, 1988) proposed to evaluate $C(O)$ by a weighted average of quadratic Taylor series expansion written for each endpoint of interval $[\, x_{i-1}, x_i \,]$. The first and second-order derivatives intervening in the Taylor series are estimated with centred differences. Consequently, for uniform grid spacing, the interpolating formula reduces to :

$$\begin{aligned} C(O) &= \Upsilon_i + a\,(\Upsilon_{i-1} - \Upsilon_i) \qquad (4.30) \\ \text{with} \quad \Upsilon_{i-1} &= C_{i-1} + \frac{(1-\theta)}{2}(C_i - C_{i-2}) + \frac{(1-\theta)^2}{2}(C_i - 2C_{i-1} + C_{i-2}) \\ \Upsilon_i &= C_i - \frac{\theta}{2}(C_{i+1} - C_{i-1}) + \frac{(\theta)^2}{2}(C_{i+1} - 2C_i + C_{i-1}) \end{aligned}$$

Intuitively, it appears that more weight should be given to Υ_i, the series developed around node i, when O is closest to this node (i.e. θ close to 0). Conversely, more weight should be given to Υ_{i-1} when θ approaches 1. Besides, the interpolation should be symmetric, which implies that $a(\theta) + a(1-\theta) = 1$. The weight function suggested in (Dan N'Guyen, 1988) is

$$a(\theta) = \theta^2(3 - 2\theta) \tag{4.31}$$

which ensures notably that the interpolating form defined by 4.30 satisfies $\frac{\partial C}{\partial x}(\theta = 0) = d_i$ and $\frac{\partial C}{\partial x}(\theta = 1) = d_{i-1}$ where the d_j, $j = i, i-1$ denote the second-order accurate centred approximations to C first-order spatial derivative at each endpoint.

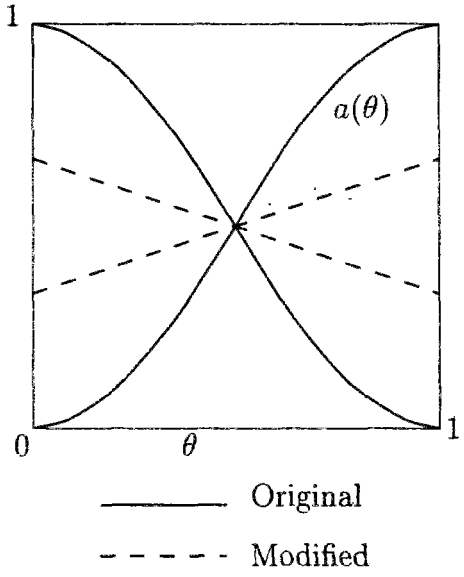


Figure 4.4: Weights for Dan N'Guyen interpolant

It is easily checked that the truncation error (TE) of eq. 4.30 together with 4.31 is third-order in space. For uniform grid spacing, a fourth-order TE could be obtained with the modified weight $a(\theta) = (1+\theta)/3$: in such cases, 4.30 is equivalent to relation 4.11.

Minimax method (Li, 1990) suggested a quadratic interpolant based on the four nodes surrounding the foot of the characteristic which, according to the Von Neumann analysis and the tests presented (Li, 1990), has attractive features, performing better than schemes stemming from quadratic (three nodes) and cubic (four nodes) polynomial fitting (the latter one being given by 4.11). The scheme appears to be less accurate than Holly-Preissmann's but avoids using scalar derivatives as dependent variables and requires consequently less computational effort.

Li introduces first his algorithm in the finite element framework. For the sake of simplicity, we shall consider a constant velocity U. The studied domain $\mathcal{D}$ is divided into uniform linear elements of length Δx : the scalar distribution at time $n\Delta t$ is given by $C(x, n\Delta t) = \sum_i C_i^n \phi_i(x)$ where ϕ_i denotes the linear basis function associated with node i. The exact solution at the next time level is $C^*(x, [n+1]\,\Delta t) = \sum_i C_i^n \phi_i(x - U\Delta t)$ (see fig. 4.5) while the numerical solution reads $C(x, [n+1]\,\Delta t) = \sum_i C_i^{n+1} \phi_i(x)$. This solution is bound to minimize the error $(C - C^*)$ according to some norm. The choice of the norm determines the solution : e.g. for the euclidian norm, C corresponds to the usual least squares approximation. Li (Li, 1990) proposes to carry

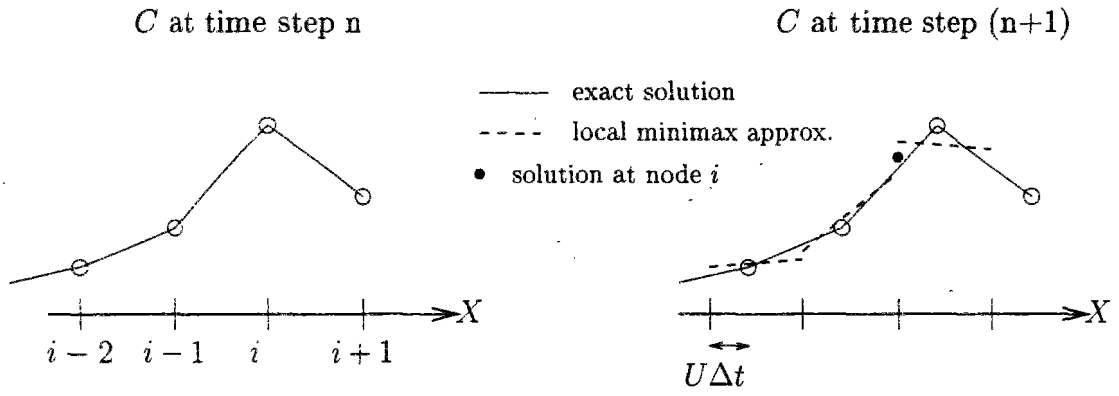


Figure 4.5: Derivation of the minimax interpolant

out the minimization individually on each element and chooses to measure the error with

$$\max_{[x_{i-1},x_i]} \mid C(x) - C^*(x) \mid$$

The corresponding best local approximation on each element is plotted in figure 4.5. It is termed the **minimax approximation**. Since a node connects two elements, the approximation is discontinuous there. To resolve this problem, Li suggests taking the nodal value C_i^{n+1} as the average of the left-element and right-element values. The resulting interpolant reads :

$$C(O) = \theta C_{i-1} + (1-\theta)\, C_i + \frac{1}{4}\theta\,(1-\theta)\,(-C_{i+1} + C_i + C_{i-1} - C_{i-2}) \qquad (4.32)$$

where θ still denotes the decimal part of the Courant number $c_r = U\Delta t/\Delta x$.

The approach leading to relation 4.32 may not seem quite rigorous. However, each of the C_j, $j = i-2, \ldots, i+1$ can be developed as a Taylor series around point $i-\frac{1}{2}$ (i.e. with coordinate $0.5\,(x_{i-1} + x_i)$). Then, it appears that formula 4.32 is equivalent to the second-order Taylor Series expansion of distribution C around point $i - \frac{1}{2}$. Relying on centred differences, we may interprete $-C_{i+1} + C_i + C_{i-1} - C_{i-2}$ as :

$$(C_i - C_{i-2}) - (C_{i+1} - C_i) = 2\Delta x\,(d_{i-1} - d_i) = -2\,\Delta x^2 \left.\frac{\partial^2 C}{\partial x^2}\right|_{i-\frac{1}{2}} + 0\left(\Delta x^4\right) \qquad (4.33)$$

(nb : for the definition of d_i, see above. We recall also that $0\,(\Delta x^m)$ denotes an expression which lower-order term is proportional to Δx^m).
Applying 4.33 together with the second-order derivative estimates d_i, it is straighforward to generalize formula 4.32 to uneven grid spacing (and variable velocities).

We must say that neither the minimax method nor the Dan N'Guyen one has been developed with an eye to preventing the occurrence of negative concentrations and ensuring monotone

interpolations. The only avowed goal was to achieve a good spatial resolution. We are not aware of any work similar to what was done with respect to the Hermite cubic interpolant, i.e. dealing with the development of appropriate shape-preserving modifications to these two interpolants.

4.2 Conservative form of the transport equation

4.2.1 Writing the algorithms in a flux conservative form

In these algorithms, the computational domain is split into small cells defined around each node i by $x_{i-1/2} \leq x \leq x_{i+1/2}$ where $x_{i+1/2} = 0.5\,(x_i + x_{i+1})$. The mass in box i is assumed to be equal to $\psi_i \Delta x_i$ with $\Delta x_i = x_{i+1/2} - x_{i-1/2}$. The variation of mass between times t^n and $t^{n+1} = t^n + \Delta t$ is computed by integrating 4.1 : it depends on the advective fluxes at the right and left boundaries of the box, respectively $F_{i+1/2}$ and $F_{i-1/2}$. The scalar obeys :

$$\psi_i^{n+1} = \psi_i^n - \frac{1}{\Delta x_i}\left(F_{i+1/2} - F_{i-1/2}\right) \tag{4.34}$$

$$\text{where} \quad F_{i+1/2} = \int_{t^n}^{t^{n+1}} U(x_{i+1/2},\tau)\psi(x_{i+1/2},\tau)\,d\tau \tag{4.35}$$

The fluxes $F_{i+1/2}$ can be approximated by different methods :

1. First, we can use the mean-value theorem which states that there exists some time s, $t^n \leq s \leq t^{n+1}$, so that 4.35 is tantamount to

$$F_{i+1/2} = \Delta t\, U(x_{i+1/2},s)\,\psi(x_{i+1/2},s) \tag{4.36}$$

(nb : we shall use from here on the notation $U_{i+1/2}$ and $\psi_{i+1/2}$ for quantities $U(x_{i+1/2})$ and $\psi(x_{i+1/2})$ respectively)

The velocity field is an input of the transport equation. If the velocities are known, or computed, on the same grid as that used for scalar transport, the $U_{i+1/2}$ are usually determined by linear interpolation, both in time and space, between surrounding nodes. Some combined models for flow and scalar transport may eventually use staggered grids where the $x_{i+1/2}$ coincide with the locations where velocities are computed : this simplifies the evaluation of the $U_{i+1/2}$.

The main problem lies in evaluating the $\psi_{i+1/2}$. Explicit methods compute it by performing some interpolation over the concentration field at time t^n. The interpolating forms used at that stage are most frequently polynomials like those introduced above.
$\psi_{i+1/2}$ can alternatively be expressed as a combination of the concentrations at both time

instants n and $n+1$. This leads to implicit schemes requiring the inversion of some linear system.

This approach is illustrated in sections 4.2.2 and 4.2.3.

2. Alternatively, we can notice that only the particles which lie within some distance dl of $x_{i+1/2}$ happen to cross the cell boundary during the time step, so that instead of 4.35, we can consider

$$F_{i+1/2} = \int_{x_{i+1/2}-dl}^{x_{i+1/2}} \psi(x, t^n)\, dx \tag{4.37}$$

To apply 4.37 requires first assessing dl. The favourite formula appears to be $dl = U_{i+1/2}\,\Delta t$. Then, the scalar distribution ψ at time t^n needs to be approximated on each cell by some integrable shape function ϕ_i. The shape functions in use are polynomial ones.

Schemes belonging to this category are naturally explicit.

They are discussed in sections 4.2.4 and 4.2.5.

4.2.2 Schemes based on the determination of a mean interface value

We present hereafter a few explicit algorithms. Their description is brief. Indeed, we intend to use them mainly for introducing the dilemna faced in the treatment of the transport equation, namely how to combine, on one hand, accurate resolution in the vicinity of sharp gradients of the transported scalar and, on the other hand, preservation of properties of the scalar field such as positiveness, monotonicity ... (nb : this point has already been tackled in the case of backward characteristics methods, cf sec. 4.1.2.2).

The most straightforward approximation to $\psi_{i+1/2}$ is certainly $\psi_{i+1/2} = 0.5\,(\psi_i + \psi_{i+1})$. This is tantamount to approximating the partial derivative $\frac{\partial U\psi}{\partial x}$ by centred differences (cf eq. 3.9). Unfortunately, this method leads to the appearance of spurious oscillations, or "wiggles", in the numerical solution (Roache, 1985). A cure for the unwanted wiggles is provided by the use of the **donor cell** approach :

$$F_{i+1/2} = U_{i+1/2}\,\psi_i \quad \text{if } U_{i+1/2} \geq 0 \text{ and } F_{i+1/2} = U_{i+1/2}\,\psi_{i+1} \text{ otherwise}$$

which can be summed up by the single formula :

$$F_{i+1/2} = \frac{1}{2}\left[\left(U_{i+1/2} + \left|U_{i+1/2}\right|\right)\psi_i + \left(U_{i+1/2} - \left|U_{i+1/2}\right|\right)\psi_{i+1}\right] \tag{4.38}$$

This amounts to replacing $\frac{\partial U\psi}{\partial x}$ by so-called upstream or upwind differences. Such a method is devoid of wiggles but is unfortunately plagued by excessive damping. Indeed, it can be shown (see for instance (Smolarkiewicz, 1983)) that the combination of relations 4.34 and 4.38 approximates in fact the following :

$$\frac{\partial \psi}{\partial t} + \frac{\partial U\psi}{\partial x} = \frac{\partial}{\partial x}\left(K_{\text{impl}}\frac{\partial \psi}{\partial x}\right) \quad \text{with } K_{\text{impl}} = \frac{1}{2}\left(|U|\,\Delta x - \Delta t\, U^2\right) \tag{4.39}$$

A first strategy for avoiding excessive damping consists in "removing" explicitly the numerical diffusion. For instance, (Smolarkiewicz, 1983) rewrites the numerical diffusion term as :

$$\frac{\partial}{\partial x}\left(K_{\text{impl}}\frac{\partial \psi}{\partial x}\right) = -\frac{\partial U_d \psi}{\partial x} \tag{4.40}$$

$$\text{with } U_d = -\frac{K_{\text{impl}}}{\psi}\frac{\partial \psi}{\partial x} \quad \text{if } \psi \neq 0 \tag{4.41}$$

$$U_d = 0 \quad \text{otherwise} \tag{4.42}$$

Now, defining an "anti-diffusion" velocity $\tilde{u} = -U_d$, he suggests solving 4.1 in two steps. First, an intermediate solution $\psi^\star$ is determined by application of the donor-cell approximation. Secondly, $\psi^\star$ is corrected by applying to it the advection equation with the anti-diffusion velocity $\tilde{u}$. This step is also conducted with the donor-cell method. The corrective step may be repeated several times, in order to remove more and more artificial damping. Due to its inherent simplicity, this algorithm extends straightforwardly to multidimensional applications (Smolarkiewicz, 1984; Smolarkiewicz & Clark, 1986; Smolarkiewicz & Grabowski, 1990).

A second strategy is to use higher-order approximations of $\psi_{i+1/2}$. For instance, (Leonard, 1979) proposed to approximate the face value $\psi_{i+1/2}$ ($= \psi_f$) with the help of a quadratic polynomial based on ψ values at the surrounding nodes i and $i+1$, plus the next upstream node, namely $i-1$ when $U_{i+1/2}$ is positive, $i+1$ when it is negative. In the case of uniform grid spacing, the resulting third-order accurate QUICK scheme reads :

$$\psi_f = \frac{1}{2}\left(\psi_C + \psi_D\right) - \frac{1}{8}\left(\psi_D - 2\psi_C + \psi_U\right) \tag{4.43}$$

$$\text{with } \psi_C = \psi_i \quad , \quad \psi_D = \psi_{i+1} \, , \; \psi_U = \psi_{i-1} \quad \text{if } U_{i+1/2} \geq 0$$

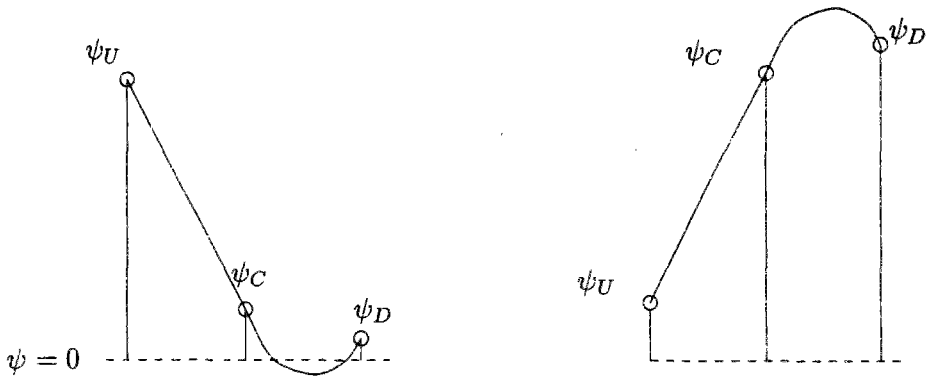
$$\psi_C = \psi_{i+1} \quad , \quad \psi_D = \psi_i \, , \quad \psi_U = \psi_{i+2} \quad \text{if } U_{i+1/2} < 0$$

Subscripts U, C, D refer respectively to the Upstream, Central and Downstream nodes used to construct the interpolation. According to these notations, it appears that ψ_f denotes the scalar value at the *outflow* face of the cell centred around node C. From now on, ψ_u will denote the value at the *inflow* face, i.e. midpoint between nodes U and C.

As discussed above in the case of the Hermite cubic interpolant (sec. 4.1.2.1), high-order approximations have unfortunately some shortcomings. For strong gradients of the concentration field, the quadratic approximation may raise locally non-physical or dubious values for ψ_f, as illustrated on figure 4.6. More precisely, the behaviour of the quadratic form can be characterized according to the normalized concentration

$$\tilde{\psi}_C = \frac{\psi_C - \psi_U}{\psi_D - \psi_U}$$

(nb : $\tilde{\psi}_U = 0$, $\tilde{\psi}_D = 1$; quadratic approximation yields $\tilde{\psi}_f = \frac{3}{4}\tilde{\psi}_C + \frac{3}{8}$ and $\tilde{\psi}_u = \frac{3}{4}\tilde{\psi}_C - \frac{1}{8}$)
Local monotonicity (i.e. on interval [x_U, x_D]) of the scalar discrete data set is expressed by the



(a) : undershoot (b) : overshoot

Figure 4.6: Non-physical quadratic interpolations

condition $0 \leq \tilde{\psi}_C \leq 1$. A monotonicity preserving scheme should be such that the underlying continuous approximation to the discrete scalar distribution on [x_U, x_D] yields :

$$0 \leq \tilde{\psi}_u \leq \tilde{\psi}_C \leq \tilde{\psi}_f \leq 1 \tag{4.44}$$

However, for $\tilde{\psi}_C \leq 1/4$ and $\tilde{\psi}_C \geq 3/4$, the quadratic approximation is not monotone : it displays extrema respectively in intervals [x_U, x_u] and [x_f, x_D]. Notably, for $\tilde{\psi}_C \geq 5/6$, ψ_f lies out of the range [$\min(\psi_U, \psi_D)$, $\max(\psi_U, \psi_D)$] and for $\tilde{\psi}_C \leq 1/6$ the corresponding $\tilde{\psi}_u$ is negative. Methods relying on higher-order approximation polynomials are similarly plagued by the possible occurrence of undershoots (fig. 4.6, sketch (a)) or overshoots (fig. 4.6, sketch (b)).

Different methods have been designed in order to overcome this difficulty. For instance, (Leonard, 1981; Leonard, 1988) proposed a refinement of the QUICK method (first termed the EXQUISITE, then the SHARP algorithm) : it consists of replacing in monotonic regions polynomial fitting by exponential fitting, which best accommodates sharp gradients while ensuring the interpolation remains monotone. However, the resulting formula for $\tilde{\psi}_f$ (Leonard, 1981) is rather complicated. Fortunately, it turns out that QUICK and exponential fitting are close in the range $\left|1 - 2\tilde{\psi}_C\right| \leq 0.3$, namely, in terms of non-normalized variables,

$$| \psi_D - 2\psi_C + \psi_U | \leq 0.3 | \psi_D - \psi_U |$$

which can be interpreted as reflecting a "smooth" (small-curvature) region. Besides, in the remaining strong-curvature regions, the exponential interpolant can be adequately piecewise approximated by simpler relations (Leonard, 1988). In non-monotonic regions, the algorithm is extended by resorting either to QUICK or to the donor cell method ($\tilde{\psi}_f = \tilde{\psi}_C$).

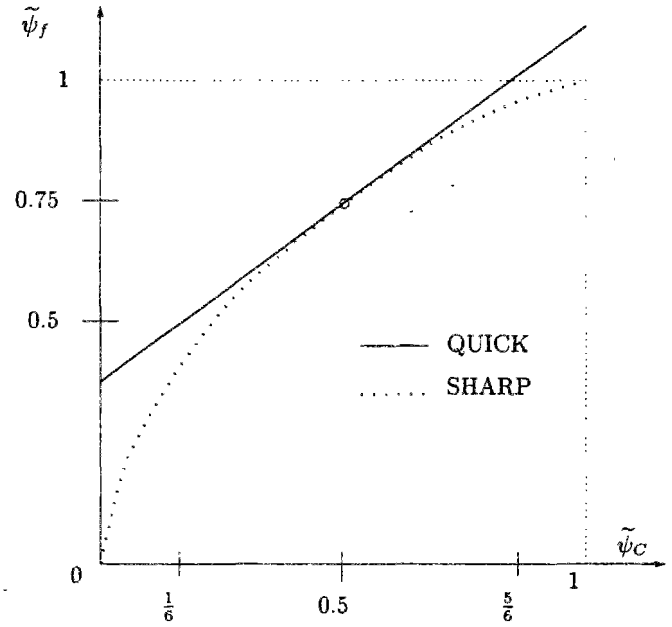
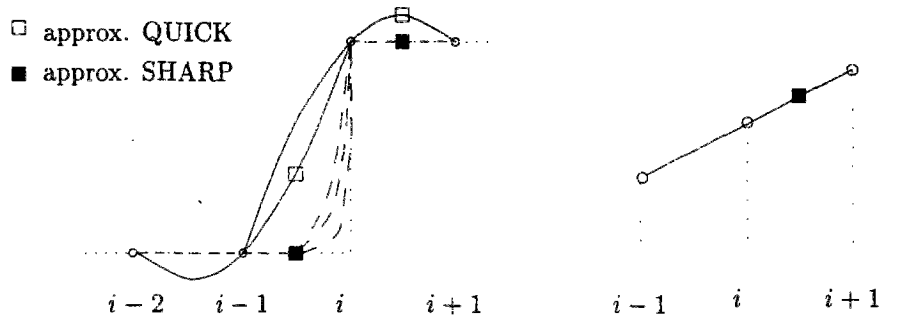


Figure 4.7: QUICK & SHARP face value approximations

Figure 4.7 allows to compare the QUICK and SHARP $\tilde{\psi}_f$ estimations. With monotonic SHARP, we observe naturally that, for all $\tilde{\psi}_C$, $0 \leq \tilde{\psi}_f \leq 1$. It also appears that $\tilde{\psi}_f$ reduces to zero when $\tilde{\psi}_C$ is null (i.e. $\psi_C = \psi_U$), so that the SHARP interpolant assumes implicitly in that case that ψ remains constant some distance downstream of the central node. The difference between both interpolants is perhaps better understood when considering the example of a sharp front propagation (cf figure 4.8, sketch (a)). By drawing a line through the set of points $\left\{ (x_{i+j}, \psi_{i+j}),\ j = 0, \pm\frac{1}{2}, \pm 1 \right\}$, we may visualize how the algorithm "interprets" the discrete data distribution, i.e. what continuous data distribution it implies. The quadratic forms used by QUICK to derive each face value are indicated by the solid line. The continuous "reconstitutions" by SHARP (dashed lines) are but tentative in area $\left[x_{i-1/2}, x_i \right]$: yet we may notice they all correspond to a sharper gradient at the base of the front. Besides, there is no overshoot at the upper edge of the front. For smooth-curvature regions, there is not much difference between both schemes. In the special case when the discrete data are aligned (cf sketch (b) fig. 4.8), i.e. $\tilde{\psi}_C = 1/2$, the most logical estimation of $\tilde{\psi}_f$ is obtained trough linear interpolation, which yields $\tilde{\psi}_C = 3/4$. Both QUICK and SHARP raise this estimation (cf fig. 4.7).

The development of the SHARP algorithm has been based on the assumption of a uniform grid. Analogous formulae which incorporate the effect of grid variations could be set up. However, Leonard (Leonard, 1988) suggests that simply using the constant-grid spacing formulae on a variable grid results in negligible errors provided the adjacent mesh width ratios lie within the range $0.8 \sim 1.25$.



(a) : discontinuity (b) : smooth curve

Figure 4.8: QUICK & SHARP applications to strong and smooth-curvature shapes

The SHARP algorithm yields excellent results when applied to the propagation of fronts (Leonard, 1988). Its relevance as regards the transport of other shapes need to be assessed. Besides, as it is derived from the QUICK scheme, it suffers rather stringent stability requirements : a formal Von-Neumann stability analysis (cf C.3.1) concludes indeed that the QUICK scheme is unstable for pure convection (Leonard, 1979). However, as pointed out in (Leonard, 1988), the instability occurs at the long-wavelength end of the Fourier spectrum ($\lambda \to \infty$). The use of a finite grid imposes some cut-off on the Fourier spectrum ($\lambda \leq N\Delta x$ with N total number of nodes) and the resulting stability criterion becomes $c_\tau \leq \pi^2/2N^2$ (c_τ Courant number). According to (Leonard, 1988), even this may be violated when using time-marching towards a steady-state solution, because any unstable modes tend to be suppressed by the steady-state boundary conditions. On the basis of numerical experiments, Leonard suggests that in such case the maximum allowable Courant number is about 0.2 ... which remains small ! In order to relax the stability requirements, it is necessary to apply some degree of implicitation. For the QUICK scheme, it would result in something like :

$$\begin{aligned} \psi_f \;=\; & \theta \left[\frac{1}{2}\left(\psi_C + \psi_D\right) - \frac{1}{8}\left(\psi_D - 2\psi_C + \psi_U\right) \right]^{n+1} \\ + \; & (1-\theta) \left[\frac{1}{2}\left(\psi_C + \psi_D\right) - \frac{1}{8}\left(\psi_D - 2\psi_C + \psi_U\right) \right]^{n} \qquad (\theta \in]0,1]) \end{aligned}$$

The performance of such approximations with different values of *theta* (denoted the implicitation parameter) has been investigated by (Falconer & Liu, 1988; Chen & Falconer, 1992). Contrary to these authors'conclusion, we do not find the overall results particularly impressive. The implicit QUICK formulation is perhaps suitable when embedded within a hydraulic flow model and applied to compute such variables as velocity and elevation. In the performed tests, we suspect nevertheless that the advective terms in the momentum equations were not essential. On the other hand, when dealing with the convection-dominated propagation of scalar field with narrow shapes and strong gradients, the resolution is insufficient, as was demonstrated in

(Simon, 1990a).
It is not straightforward to formulate an implicit version of the SHARP scheme as, because of exponential fitting, ψ_f turns to be a non-linear function of ψ values at neighbouring nodes. Yet, one could contemplate to make SHARP implicit through an iterative process applied at each time instant, as is done in other algorithms derived from QUICK (Gaskell & Lau, 1988).

4.2.3 Limiting the interface value estimation

Monotonicity-preserving algorithms are usually built as follows : first, an interface value is estimated by some unconstrained scheme; secondly, this estimate is corrected in order to ensure monotonicity. We shall now examine which limitations are enforced, which criteria are used for that purpose.

(Gaskell & Lau, 1988) examine the case of the steady-state convection-diffusion equation :

$$\frac{\partial u_j \psi}{\partial x_j} = S \tag{4.45}$$

(where source-term S accounts for the combined effect of diffusion, external sources and biogeochemical reactions). They discuss extensively the constraints that the face value estimates ψ_f and ψ_u need to satisfy in order for the corresponding time-marching resolution of eq. 4.45 to be physically meaningful.

If the discrete data set is monotone, the face value approximations need to satisfy inequalities 4.44. We recall that these inequalities read

$$0 \leq \tilde{\psi}_u \leq \tilde{\psi}_C \leq \tilde{\psi}_f \leq 1$$

The various cases corresponding to a non-monotone discrete data set are sketched on figure 4.9.

- Let ψ_C be a local maximum (cf fig 4.9, cases (a) and (c)). This is physically possible if and only if $S^{(C)}$ (source term in cell centred around C) is positive. If the velocity is uniform, integrating 4.45 demonstrates that $S^{(C)}$ is proportional to and has the same sign as the difference ($\psi_f - \psi_u$). The ψ_u and ψ_f evaluations must both reflect $S^{(C)}$ positiveness and ensure there are no further extrema lying between the nodes. Consequently, we should have $\psi_f \geq \psi_u$, $\psi_U \leq \psi_u \leq \psi_C$ and $\psi_D \leq \psi_f \leq \psi_C$.
- Now, if ψ_C is a local minimum (cf fig 4.9, cases (b) and (d)), which implies a negative $S^{(C)}$, ψ_f and ψ_u must abide by the following inequalities :
  $\psi_f \leq \psi_u$, $\psi_C \leq \psi_u \leq \psi_U$ and $\psi_C \leq \psi_f \leq \psi_D$.

Hence, it follows that in non-monotone regions the normalized face value $\tilde{\psi}_f$ needs to satisfy :

$$\text{for } \tilde{\psi}_C \geq 1, \quad 1 \leq \tilde{\psi}_f \leq \tilde{\psi}_C \tag{4.46}$$

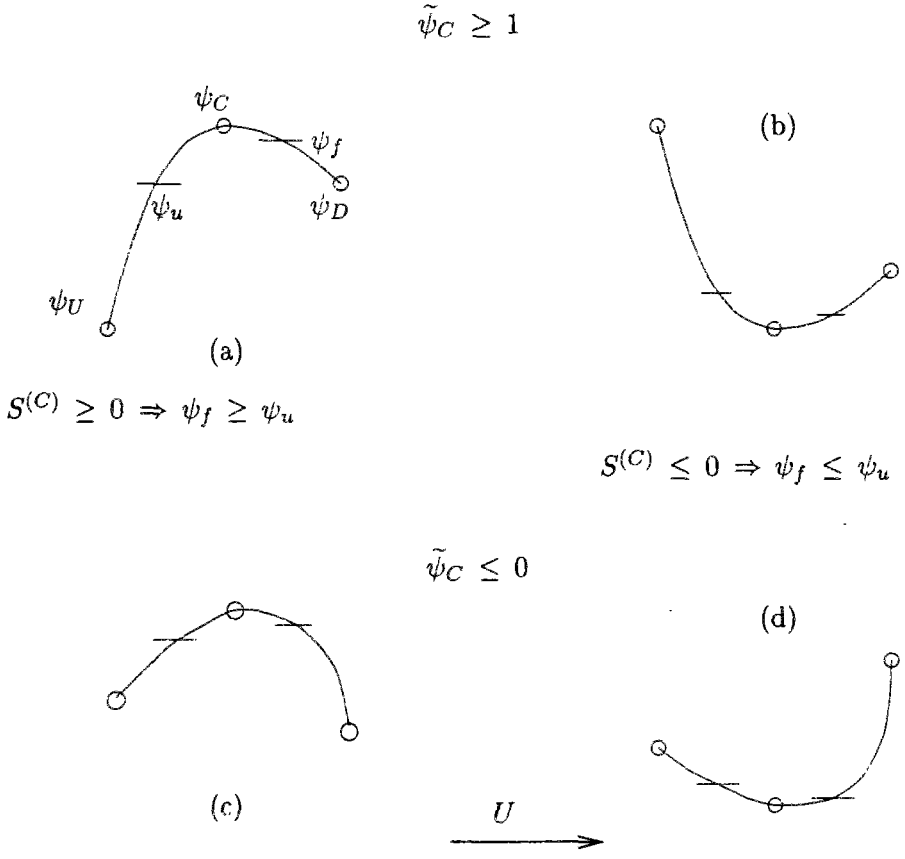


Figure 4.9: Non-monotone discrete data set : the possible configurations

$$\text{for } \tilde{\psi}_C \leq 0, \quad \tilde{\psi}_C \leq \tilde{\psi}_f \leq 0 \tag{4.47}$$

The set of inequalities 4.44, 4.46 and 4.47 ensures both the preservation of monotonicity when needed and the satisfaction of physical bounds in non-monotone areas. (Gaskell & Lau, 1988) term the algorithms which satisfy these constraints *boundedness-preserving* algorithms. The functions $\tilde{\psi}_f = f\left(\tilde{\psi}_C\right)$ which characterize such schemes pass trough the shaded areas in figure 4.10. Notably, the ψ_f function must pass through points (0,0) and (1,1).

An example of a boundedness-preserving algorithm is supplied in (Gaskell & Lau, 1988) : it consists of a variation on QUICK but simpler than SHARP. The SMART algorithm (as sketched on figure 4.10) relies on a piecewise-linear estimation for $\tilde{\psi}_f$. In non-monotone regions, it makes use simply of $\tilde{\psi}_f = \tilde{\psi}_C$. It uses QUICK estimation, thus being third-order accurate, wherever QUICK yields consistent results. In the remaining areas (respectively $0 \leq \tilde{\psi}_C \leq 1/6$ and $5/6 \leq \tilde{\psi}_C \leq 1$) the ψ_f function ensures continuity with respect to the QUICK approximation and yields the suitable value for $\tilde{\psi}_C = 0$ and 1. The SMART algorithm is implemented through an iterative process, applied at each time instant (Gaskell & Lau, 1988). It appears to perform as well as SHARP ... and is plagued by similar stability constraints.

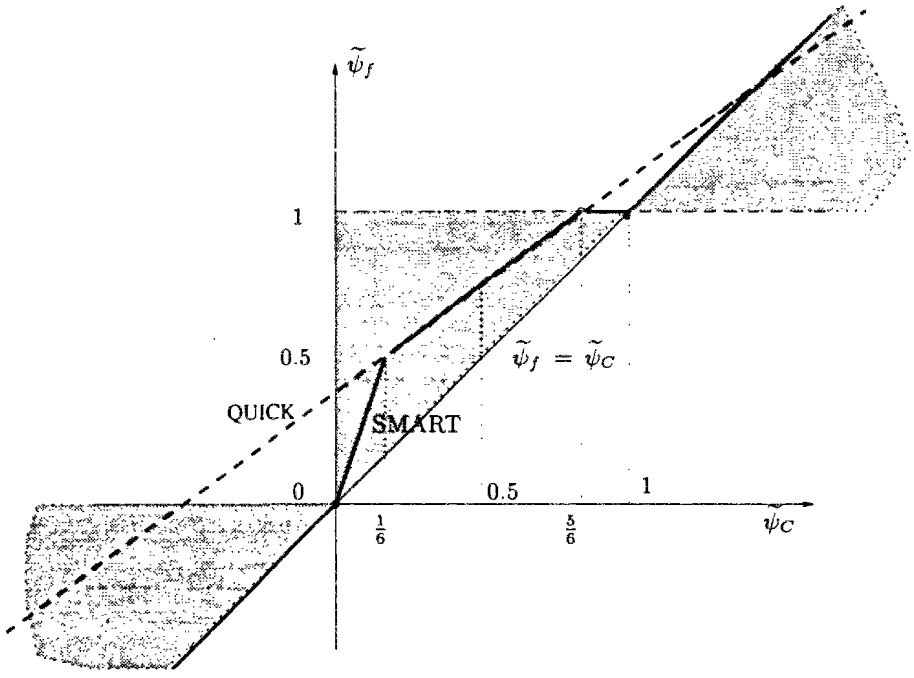


Figure 4.10: Steady-state convection : constraints for ensuring boundedness

(Leonard & Niknafs, 1991) further correct the criterion 4.44 for unsteady situations. Assuming that the velocity is uniform and denoting by c_r the corresponding Courant number, it

follows that $\tilde{\psi}_C$ at the next time instant obeys

$$\tilde{\psi}_C^{n+1} = \tilde{\psi}_C^n - c_r \left(\tilde{\psi}_f - \tilde{\psi}_u \right)$$

First, ψ_C^{n+1} must be positive. Besides, considering the downstream displacement of the scalar profile during the time step, with $c_r \leq 1$, it follows that ψ_C^{n+1} should range between ψ_C^n and ψ_U^n. These conditions can be expressed in terms of the normalized concentration : in short, a coherent scheme would enforce

$$0 \leq \tilde{\psi}_C^{n+1} \leq \tilde{\psi}_C^n \quad \text{if } 0 \leq \tilde{\psi}_C^n \leq 1 \tag{4.48}$$

Taking worst case conditions, this yields requirements somewhat more stringent than for steady-state applications (Leonard & Niknafs, 1991), namely

$$\tilde{\psi}_C \leq \tilde{\psi}_f \quad \text{and} \quad \tilde{\psi}_f \leq \frac{1}{c_r} \tilde{\psi}_C \tag{4.49}$$

which are sketched in figure 4.11. The set of inequalities 4.49 is termed the *Universal Limiter* by Leonard who points out it can be applied in combination with any arbitrary high-order explicit scheme : whenever, the $\tilde{\psi}_f$ approximation yielded by such a scheme lies outside the shaded area on fig. 4.11, it is simply replaced by the nearest allowable $\tilde{\psi}_f$ at the same $\tilde{\psi}_C^n$.

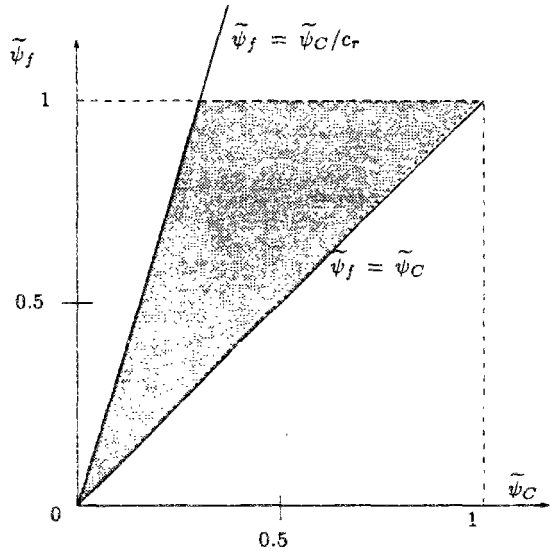


Figure 4.11: Unsteady convection : monotonicity constraints

The performance of the Universal Limiter with respect to other limiters is favourably demonstrated in (Leonard & Niknafs, 1991). However, the case of transporting narrow shapes underscores the following drawback. As the shape travels, the physical extrema may lie indeed *between* grid nodes. As the limiter uses only grid node values, the exact value of the extrema may then

be "forgotten" and excessive damping occurs, even when high-order (seventh and ninth-order) schemes are used. This "clipping" phenomenon has already been mentioned in sec. 4.1.2.2. (Leonard & Niknafs, 1991) propose a method to discriminate between spurious and physical extrema. It relies on the analysis of differences $\{ (\Delta_j = (\psi_{i+j} - \psi_{i+j-1}), j = -2 \ldots 3 \}$ around each node i. According to (Leonard & Niknafs, 1991), a physical maximum (resp. minimum) occurs at i if we have : Δ_1 and Δ_2 negative (resp. positive), Δ_0 and Δ_{-1} positive (resp. negative), $|\Delta_1| \leq |\Delta_2|$ and $|\Delta_0| \leq |\Delta_{-1}|$. The last two inequalities reflect the plausible concavity (maximum) or convexity (minimum) observed in the scalar distribution in the vicinity of a physical extremum. Conversely, a spurious extremum is generally associated to short-wavelength oscillations in the numerical solution, with rapidly changing value, gradient and curvature. The limiter is relaxed at every node which appears to be either a physical extremum or its immediate neighbour. A significant improvement in the solutions is henceforth obtained.

4.2.4 Schemes based on spatial integration of an approximate distribution

A large number of schemes may be built according to the choice of function ϕ_i which approximates the concentration field on cell i.
In the following, we shall not dwell on the problem which amounts to finding the correct bounds of the flux integrals, i.e. determining where the material flowing into or out of cell i comes from. As mentioned above, it is generally assumed that this region is an area of length $U_{i+1/2}\Delta t$. While this approximation proves accurate enough when the velocity field is varying smoothly, its validity may break down in the case of strong flow gradients and discontinuities (Rigaudière, 1992) leading to the need of special "bound-searching" algorithms (Even & Poulin, 1993) whose cost may eventually be similar to those of characteristics backtracking schemes.

The development of approximation ϕ_i depends notably on what significance we give to discrete value ψ_i. Till now, ψ_i has been considered as the discrete value of the scalar distribution ψ at location x_i. It follows that the product $\psi_i \Delta x_i$ is but a first-order approximation to the mass contained in cell i. Alternatively, ψ_i can be considered as the average of ψ over the cell, determined by

$$\psi_i = \frac{1}{\Delta x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} \psi(x)\, dx$$

Such a definition is more consistent with the use of finite difference approximation 4.34 to equation 4.1. Yet, it does not necessarily imply that $\psi_i = \psi(x_i)$.

1. The approximations can be built by fitting a polynomial of arbitrary degree through the set of data points $\{(x_{i+j}, \psi_{i+j}), j = j_{min}, \cdots j_{max} \}$.

   The first possibility is to centre the polynomial ϕ_i around point $x_{i+1/2}$ which corresponds to the interface with the next cell. Let m be ϕ_i order. If m is odd, ϕ_i is based on nodes

$i-(m-1)/2$ to $i+1+(m-1)/2$, which is symmetric with respect to $x_{i+1/2}$. If m is even, the extra point is added on the upstream side of $i+1/2$. The corresponding expressions (on uniform grid) for fluxes $F_{i+1/2}$ obtained for $m = 1$ through 10 can be found in (Tremback *et al.*, 1987).

The second possibility is to centre ϕ_i around node i (Bott, 1989a). This will lead to slightly different expressions. Indeed, a cubic approximation to ψ centred around $x_{i+1/2}$ use nodes $i-1$ to $i+2$ while, if ϕ_i is centred around x_i and the velocity is positive, the rule of upstream weighting leads us to use nodes $i-2$ to $i+1$ in order to build ϕ_i.

The well-known QUICKEST scheme takes for ϕ_i the quadratic approximation based on nodes $i-1$, i, $i+1$. It is the generalization of QUICK to unsteady situations (Leonard, 1979). When velocity and grid spacing are uniform, the resulting scheme reads as eq. 4.11. However, the generalization suggested in (Leonard, 1979) for non-uniform flows and grids is different from that proposed by Takacs (cf relations 4.12 to 4.17). It is written :

$$\psi_i^{n+1} = \psi_i^n - \frac{\Delta t}{\Delta x_i}\left(U_{i+\frac{1}{2}}\psi^*_{i+\frac{1}{2}} - U_{i-\frac{1}{2}}\psi^*_{i-\frac{1}{2}}\right) \tag{4.50}$$

$$\psi^*_{i+\frac{1}{2}} = \frac{\psi_i+\psi_{i+1}}{2} - \frac{U_{i+\frac{1}{2}}\Delta t}{2}\,\mathrm{GRAD}_{i+\frac{1}{2}} - \frac{(\Delta x_{i+\frac{1}{2}})^2}{6}\left[1-c^2_{i+\frac{1}{2}}\right]\mathrm{CURV}_{i+\frac{1}{2}} \tag{4.51}$$

and

$$\Delta x_{i+\frac{1}{2}} = x_{i+1}-x_i\,, \quad c_{i+\frac{1}{2}} = U_{i+\frac{1}{2}}\Delta t/\Delta x_{i+\frac{1}{2}}\,, \tag{4.52}$$

$$\mathrm{GRAD}_{i+\frac{1}{2}} = (\psi_{i+1}-\psi_i)/\Delta x_{i+\frac{1}{2}} \tag{4.53}$$

$$\mathrm{CURV}_{i+\frac{1}{2}} = \left(\mathrm{GRAD}_{i+\frac{1}{2}} - \mathrm{GRAD}_{i-\frac{1}{2}}\right)/\Delta x_i \quad \text{if } U_{i+1/2} \geq 0 \tag{4.54}$$

$$= \left(\mathrm{GRAD}_{i+\frac{3}{2}} - \mathrm{GRAD}_{i+\frac{1}{2}}\right)/\Delta x_{i+1} \quad \text{if } U_{i+1/2} < 0 \tag{4.55}$$

2. Another alternative is to obtain the ϕ_i by expanding Taylor series around node i (Simon, 1990b). Beyond order 2, the corresponding polynomials no longer coincide with polynomials issued from data fitting, as they do not satisfy $\phi_i(x_{i+j}) = \psi_{i+j}$ for $|j| > 1$.

3. The third option is to use polynomials so that, on their definition interval, their spatial moments (up to some order) equate those of the scalar distribution.

   For instance, Bott (Bott, 1989b) proposes 2^{nd} and 4^{th}-order area-preserving polynomials defined so that their integral over cells centred around nodes i, $i\pm1$ ($i\pm2$) coincide with the mass in these cells (respectively $\Delta x_{i+j}\psi_{i+j}$, $j = 0, \pm1, \pm2$). The same polynomials (up to order 10) are introduced in (Tremback *et al.*, 1987) following a slightly different line of thought.

   Let $\Phi^p_{i,j}$ and Ψ^p_j denote respectively the pth-order moment of function ϕ_i and scalar distribution ψ over cell j, defined as (nb: the coordinates origin has been shifted to x_i) :

$$\Phi^p_{i,j} = \int_{x_{j-1/2}}^{x_{j+1/2}} (x-x_i)^p\,\phi_i(x)\,dx$$

$$\Psi^p_j = \int_{x_{j-1/2}}^{x_{j+1/2}} (x-x_i)^p\,\psi(x)\,dx$$

Instead of ensuring *mass preservation over different cells* (i.e. $\Phi^0_{i,j} = \Psi^0_j$ for $j = i, i\pm1\ldots$), we may set the order m of polynomial ϕ_i and then impose *higher-order moments equality over the sole cell i*, i.e. $\Phi^p_{i,i} = \Psi^p_i$ for $p = 0, 1, \ldots P$ (Boris & Book, 1973; Van Leer, 1977; Vatvani & Montazeri, 1989). If P is strictly less than m, there are degrees of freedom to define the shape functions. These need only to be continuous. For instance, a shape function may be only piecewise linear over cell i (changing of slope within the cell is allowed) (Boris & Book, 1973).

Figure 4.12 illustrates different ways to construct the shape function. The fine solid line is the quadratic upstream-weighted approximation to ψ (cf QUICKEST above) while the dashed line above is the area-preserving (over cells $i-1$, i, $i+1$) second-order approximation. (The two differ by a constant which reduces to $(-\psi_{i+1} + 2\psi_i - \psi_{i-1})/24$ on uniform grid; consequently, they are close on smooth-curvature regions). The thick lines represent linear approximations $\phi_i = \psi_i + S_i(x - x_i)$ based on the sole restriction that they are mass-preserving over cell i : S_i may thus be chosen freely. If we pick simply $S_i = 0$ (solid line), we find the donor cell method. If we pick for instance $S_i = (\psi_{i+1} - \psi_i)/(x_{i+1} - x_i)$ (dashed line), it boils down to the classic second-order Lax-Wendroff scheme (Vatvani & Montazeri, 1989).

4. Finally, different approaches may be combined. For instance, (Bott, 1989a) suggests building the ϕ_i in two steps, first by polynomial fitting, then by normalizing the resulting functions so that they become mass-preserving on cell i only. The final approximation keeps thus something of the shape of the discrete data set.

In short, there is a lot to choose from and it is worth noting that the same scheme can be derived and justified following quite different lines of thought (e.g. the QUICKEST/TAKACS one).

The behaviour of fluxes corresponding to integrals of more or less complicated functions is certainly more difficult to grasp and to illustrate than the behaviour of point evaluations. However, as the literature review reveals, we face the same troubles than with the algorithms presented in paragraph 4.2.2 : when shape functions are high-order polynomials, the resulting $F_{i+1/2}$ may induce non-physical overshoots or undershoots in the numerical solution; on the other hand, low-order approximations (e.g. by step functions) yield too dissipative schemes unable to resolve accurately short wavelengths related to narrow extrema or sharp gradients. Once again, the solution lies in correcting the evaluations computed by unbounded, high-order, schemes : this time, we shall constrain or weight *fluxes* instead of bounding *point evaluations* of the dependent transported scalar. A summary of the different approaches is given in the next section.

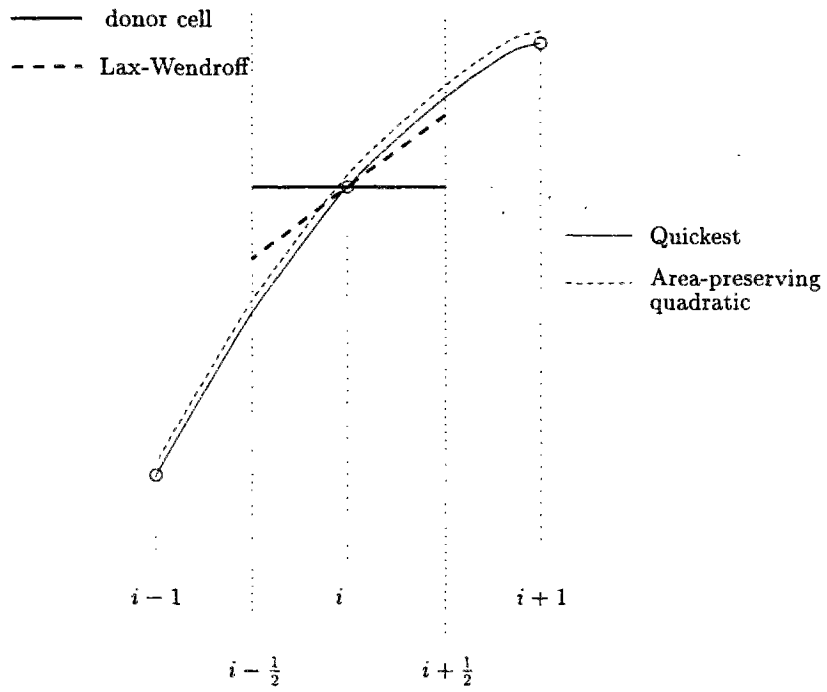


Figure 4.12: Examples of shape functions for flux computation

4.2.5 Limiting the fluxes

The first significant development in that field is ascribed to Boris & Book (Boris & Book, 1973; Book *et al.*, 1975; Boris & Book, 1976). Such algorithms are termed **Flux-Corrected Transport methods**. FCT have since been extensively studied, discussed and applied. To present their principles, we shall follow the guidelines provided in (Zalesak, 1979).

At each time step, FCT algorithms proceed as follows :

1. Compute a first estimation $F^L_{i+1/2}$ of the advective flux by applying some low-order, monotonicity-preserving, scheme
2. Compute a second estimation $F^H_{i+1/2}$ by some high-order scheme
3. Estimate the numerical diffusion induced by the low-order scheme by subtracting $F^H_{i+1/2}$ from $F^L_{i+1/2}$. Define the *anti-diffusive flux* according to

$$A_{i+1/2} = F^H_{i+1/2} - F^L_{i+1/2} \tag{4.56}$$

4. Compute an intermediate solution ψ^* by applying the low-order scheme

$$\psi^*_i = \psi^n_i - \frac{1}{\Delta x_i}\left(F^L_{i+1/2} - F^L_{i-1/2}\right) \tag{4.57}$$

5. *Limit* the anti-diffusive fluxes so that the final solution computed in step 6 below is free from undesirable behaviour

$$A^C_{i+1/2} = \lambda_{i+1/2} A_{i+1/2} \tag{4.58}$$

6. Correct the intermediate solution

$$\psi_i^{n+1} = \psi_i^* - \frac{1}{\Delta x_i}\left(A^C_{i+1/2} - A^C_{i-1/2}\right) \tag{4.59}$$

The critical step is of course step 5. Without limitation, ψ^{n+1} would simply be the high-order scheme solution. The above definition of FCT is quite general, with a considerable amount of flexibility available in the choice of the respective high-order and low-order schemes.

The limiting step can be achieved according to different criteria :

1. The original limiter (Boris & Book, 1973) is such that anti-diffusive fluxes should neither create new extrema nor accentuate existing extrema in the intermediate solution ψ^*. It reads :

$$\begin{aligned} A^C_{i+1/2} &= \sigma_{i+1/2} \max\Big\{ 0, \min\Big[\left|A_{i+1/2}\right|, \\ &\qquad \sigma_{i+1/2}(\psi^*_{i+2} - \psi^*_{i+1})\Delta x_{i+1}, \sigma_{i+1/2}(\psi^*_i - \psi^*_{i-1})\Delta x_i \Big] \Big\} \end{aligned} \tag{4.60}$$

$$\sigma_{i+1/2} = \begin{cases} +1 & \text{if} \quad A_{i+1/2} \geq 0 \\ -1 & \text{if} \quad A_{i+1/2} < 0 \end{cases}$$

The effectiveness of 4.60 can easily be checked, notably with the help of some appropriate graphics (e.g. (Book *et al.*, 1975; Zalesak, 1979)).

2. However, as ψ^* is produced by a dissipative scheme, its extrema may be significantly lower than those of the scalar distribution ψ^n at the preceding time. Limiting with respect to a diffused distribution may bring too much damping, in spite of anti-diffusive step 6.

   A first modification suggested by Zalesak is to limit the fluxes with respect to the extrema of the previous distribution ψ^n. Thus, a previous extremum can no longer be forgotten by applying the too brutal step 3.

3. (Zalesak, 1979) experienced better numerical results with the new limiter. However, damping of narrow shapes (a gaussian of approximate width $8\Delta x$) was still important. The reason is that, as the shape travels, the physical extrema may lie *between* grid nodes and thus be forgotten if one refers only to node values (cf section 4.2.3).

   Consequently, Zalesak proposed to "reconstruct", when necessary, the possible extrema of the scalar field (cf fig. 4.13) and to apply limitation with respect to these. Considerable improvement of the numerical results followed.

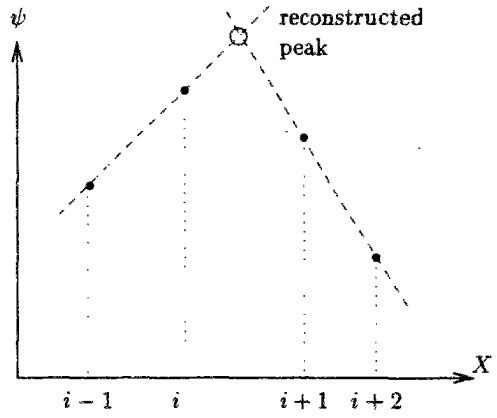


Figure 4.13: Reconstruction of extrema

The detailed formulae for the versions (2) and (3) of limiting factors $\lambda_{i+1/2}$ can be found in (Zalesak, 1979), as well as the generalization to multidimensional problems (to see also in (Book *et al.*, 1975)).

FCT methods have experienced considerable success in the treatment of discontinuities (sharp fronts), even with the original limiter (1), as demonstrated in (Boris & Book, 1973; Book *et al.*, 1975; Boris & Book, 1976). Yet, the accurate treatment of narrow shapes requires both the use of a modified limiter (3) and the use of a definitely high-order scheme (an eighth-order one) (Zalesak, 1979).

Another approach to flux limiting is to constrain the fluxes **uniquely in order to ensure positiveness** (e.g. (Smolarkiewicz, 1983; Bott, 1989a)). In such methods, the step 1, which consists of resorting to a lower-order scheme, may be omitted (e.g. (Bott, 1989a)). Hence, the appropriate limitation is applied directly to the high-order fluxes rather than to anti-diffusive fluxes. The possible flow configurations are sketched in figure 4.14, with the shaded areas representing the areas contributing to the fluxes. The reader will easily check that, for ψ_i^{n+1} to stay positive, sufficient and necessary conditions are respectively :

(a) $F_{i-1/2} \geq 0$ and $F_{i+1/2} \leq \psi_i^n \Delta x_i$

(b) $F_{i-1/2} \geq 0$ and $F_{i+1/2} \geq 0$

(c) $F_{i+1/2} \geq 0$ and $F_{i-1/2} \leq \psi_i^n \Delta x_i$

(d) $F_{i-1/2} + F_{i+1/2} \leq \psi_i^n \Delta x_i$

In order to obtain a "compact" definition of the desirable limiter, let us first introduce the

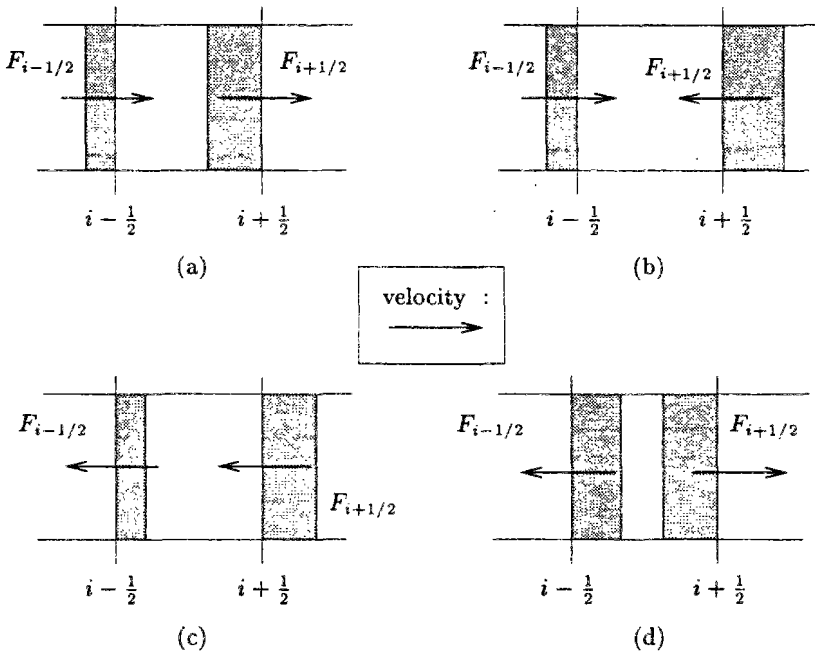


Figure 4.14: Possible orientations of fluxes according to velocity

following quantities :

$$l^{\pm}_{i+1/2} = \Delta t \frac{U_{i+1/2} \pm \left|U_{i+1/2}\right|}{2} \quad ,$$

$$F^{+}_{i+1/2} = \int_{x_{i+1/2}-l^{+}_{i+1/2}}^{x_{i+1/2}} \phi_i(x)\,dx \quad , \quad F^{-}_{i+1/2} = \int_{x_{i+1/2}}^{x_{i+1/2}-l^{-}_{i+1/2}} \phi_{i+1}(x)\,dx$$

$$I^{\pm}_{i+1/2} = \max\left(0,\, F^{\pm}_{i+1/2}\right) \quad \text{and} \quad I_i = \int_{x_{i-1/2}}^{x_{i+1/2}} \phi_i(x)\,dx$$

Considering cell i, the incoming fluxes are $I^{+}_{i-1/2}$ and $I^{-}_{i+1/2}$ while the outgoing fluxes are $I^{-}_{i-1/2}$ and $I^{+}_{i+1/2}$. $U_{i+1/2} > 0$ implies naturally that $F^{-}_{i+1/2}$ is null, conversely $U_{i+1/2} < 0 \Rightarrow F^{+}_{i+1/2} = 0$.

The general procedure for obtaining a positive definite scheme stands as follows :

- Compute all unconstrained fluxes $F^{\pm}_{i+1/2}$, their positive values $I^{\pm}_{i+1/2}$ and the total integrals I_i, by some (preferably high-order) scheme
- Limit fluxes according to

$$F_{i+1/2} = \frac{I^{+}_{i+1/2}}{F^{\max}_{i}} \cdot \left(\psi^{n}_{i}\Delta x_i\right) - \frac{I^{-}_{i+1/2}}{F^{\max}_{i+1}} \cdot \left(\psi^{n}_{i+1}\Delta x_{i+1}\right) \tag{4.61}$$

$$\text{with} \quad F^{\max}_{i} = \max\left(I_i,\, I^{+}_{i+1/2} + I^{-}_{i-1/2} + \epsilon\right) \tag{4.62}$$

(nb : ϵ is a small scalar introduced to avoid division by zero).
The limiter is designed so that neither the individual corrected fluxes nor the sum of the outgoing corrected fluxes can exceed the total mass contained in cell i.

- Advance the solution to next time level by :

$$\psi_i^{n+1} = \psi_i^n - \frac{1}{\Delta x_i}\left(F_{i+1/2} - F_{i-1/2}\right)$$

This procedure can be extended to multidimensional situations either in one single step, either by splitting the resolution according to the different coordinate directions. In the latter case, the limiter remains unchanged. In the former case, when computing $F_{i,j}^{\max}$ (for cell centred around node (i,j)), one needs to replace the second term of the parenthesis in eq. 4.62 by the sum of all outgoing fluxes, considered along every coordinate direction (Smolarkiewicz, 1989).

While the procedure outlined above applies to any high-order scheme, it is possible to simplify the limiter if the "raw" fluxes $F_{i+1/2}^{\pm}$ already satisfy some constraint. Such is the case for instance with the area-preserving shape functions of Bott (Bott, 1989b) which satisfy $I_i = \psi_i \Delta x_i$ by construction.

The positiveness-preserving limiter prevents undershoots and controls to some extent the growth of possible overshoots, while not completely forbidding them as does the first limiter (more exactly as do versions (1) and (2) of the first limiter). However, it applies only to scalars whose initial distribution is positive definite. If some negative values are present in the initial distribution, or introduced during the computation (e.g. by some ill-conditioned biogeochemical sink term), the scheme becomes unstable (Smolarkiewicz & Clark, 1986). Extension of such algorithms to scalar distributions with variable sign may nevertheless be considered at the expense of a slight modification. Indeed, as pointed out in (Smolarkiewicz & Clark, 1986), the transport equation is invariant with respect to the addition of a constant. Consequently, the problem may be solved by applying the scheme to the modified field $\psi' = \psi - A$, A being some constant satisfying $A \leq \min_{[x,t]} \psi$. Unfortunately the process of limiting the fluxes yields non-linear algorithms : hence, the numerical solution is not insensitive to the choice of A. Smolarkiewicz noted a considerable improvement in his scheme accuracy when choosing a large negative value as A (Smolarkiewicz & Clark, 1986) : in that case, his formally second-order scheme mimicked a third-order one, this resulting in an improved resolution of narrow shapes. However, this particular result may not be taken for granted for other positive-definite schemes.

Finally, we may mention a class of schemes frequently used in computational fluid dynamics (engineering applications such as in aeronautics), the **Total Variation Diminishing** (TVD)

schemes. TVD apply to the resolution of hyperbolic conservation laws

$$\frac{\partial w}{\partial t} + \frac{\partial f(w)}{\partial x} = 0$$

of which the scalar transport equation is one of the simplest examples. These are essentially second-order, explicit or implicit, algorithms where the limiter is designed so that the quantity $\sum_i |w_{i+1} - w_i|$ is not allowed to grow from one time level to the next. A good introduction to these methods is supplied for instance in (Yee, 1987). TVD schemes appear to treat successfully step discontinuities, e.g. in compressible fluids problems involving shocks. Yet, their only second-order spatial accuracy does not recommend them as regards the transport of distributions involving narrow extrema (Leonard & Niknafs, 1991).

4.3 A selection of schemes

What can be concluded from sections 4.1 and 4.2 is that many numerical methods indeed have been proposed for the solution of the transport equation. What are we looking for ? An ideal scheme would be accurate, easy to implement and cost-effective. It seems that till now no scheme has proved fully satisfactory with respect to these three points.

We cannot rely only on the literature review to select an appropriate method. Indeed, what we find generally in papers is the assessment of one particular scheme performance through its Fourier analysis (when it is possible, i.e. when the scheme yields a linear dependence between advected scalar values and its values at the previous time level) and through one or two test cases at most. For instance, the sensitivity of the numerical forecasts with respect to a broad range of Courant numbers is most often not investigated (apart from exceptionally detailed analysis as provided in (Williamson & Rasch, 1989; Rasch & Williamson, 1990)). Besides, the test cases used for different schemes are frequently distinct so that an objective comparison is not feasible. Finally, the discussion of CPU requirements is often omitted. *What we need thus is to achieve a systematic comparison within a common reference framework, made of relevant test cases.* Fortunately, the setting-up of such a reference framework has been undertaken in the context of the Convection-Diffusion Forum (held in connection with the International Conferences on Computational Methods in Water Resources) (Baptista *et al.*, 1988). In the next chapter, we shall rely on these problems to analyze and compare the behaviour of several methods.

It would be of course a gigantic task to compare every available method : prior to applying the tests, it is essential to extract a handful of relevant algorithms, representative of different approaches to the solution of convection-diffusion problems.

- The category of backward characteristics methods needs obviously to be explored.

  A popular interpolating form is the Hermite cubic interpolant (cf sec 4.1.2.1) :

- The Holly-Preissmann method (cf sec. 4.1.2.1) has long been known for its impressive accuracy. Yet, it is has been blamed for its computational cost.
- From the works of Rasch & Williamson it follows that the combined use of the cubic interpolant with an Akima derivative estimate is an interesting alternative, notably as regards the transport of narrow shapes. From now one, we shall term this method the "Rasch algorithm".

The recently proposed Dan N'Guyen and minimax methods (sec. 4.1.2.3) are representative of methods using interpolators distinct from the Hermite one.

The table below sums up some features of these four schemes (in 1D case), namely the stencil of points they use for interpolating the advected value at the foot of the characteristic and their formal order of accuracy, monitored as the order of the leading term in the scheme truncation error.

Table 4.1: Features of tested backward characteristic methods

Scheme	Stability condition	Nodes involved in interpolation (foot between nodes j and $j-1$)	Error Order	Comments
Holly-Preissmann	N	$j-1$ to j	4	$\frac{\partial C}{\partial x}$ as new dependent var.
Rasch	O	$j-3$ to $j+2$	4	
Dan N'Guyen	N	$j-2$ to $j+1$	3	
Minimax	E	$j-2$ to $j+1$	3	

The use of a wider stencil in the Rasch method is due to the estimation of the local curvatures, needed in order to weight the slope estimates. The more nodes a scheme uses, the more adaptation is needed in the vicinity of boundaries.

- Now, let us consider the schemes developed to deal with the conservative form of the transport equation.

  We have been focusing mainly on algorithms where fluxes are computed from an approximate distribution of the scalar (cf sec. 4.2.4). We felt it provided a sounder estimation of fluxes than determining the fluxes from a mean interface value (sec. 4.2.2) : in the latter case, if the interface value is computed explicitly this leads to stringent stability requirements (cf QUICK, SHARP and SMART for instance), whereas the choice of some degree of implicitation in order to stabilize the scheme is always somewhat arbitrary.

The shortcomings of donor-cell and second-order methods such as the Lax-Wendroff are well known (Roache, 1985). Consequently, we have been looking at higher-order schemes.

From a literature review, it stems that the unlimited QUICKEST scheme is quite popular and widely used in numerous water surface flow and transport models. It thus seemed natural to include it in our sample.

Higher-order schemes relying on polynomial approximations have been investigated by (Tremback *et al.*, 1987; Bott, 1989a; Bott, 1989b). We have restricted our analysis to schemes whose stencil is not wider than the stencil used by the most demanding backward characteristic methods, i.e. that of the Rasch one. Consequently, we have been excluding the study of FCT methods (cf sec. 4.2.5), which, while they perform quite well with respect to steep front propagation, seem unable to deal correctly with narrow shape transport unless the higher-order method used is a eighth-order one (Zalesak, 1979; Vatvani & Montazeri, 1989), involving thus nine grid nodes in the computation of each advected node value.

As stated above, polynomial approximations used for flux evaluation may be built according to different methods : Taylor Series expansion (Simon, 1990b), polynomial fitting (Tremback *et al.*, 1987; Bott, 1989a), mass preservation criterion (Bott, 1989b) ... We realized a preliminary assessment of the performance of each of these approaches, which results are summarized in F.1. This led us to test two algorithms where the approximations are respectively a third-order expansion of Taylor series (termed BOTT3 from now on) and a fourth-order mass preserving polynomial (termed BOTT4 hereafter). Both schemes are applied with a limiter enforcing positiveness.

Table 4.2: Features of tested flux-form methods

Scheme	Stability condition	Nodes involved in computation advected value at node j	Error Order
QUICKEST	$c_r \leq 1$	$j-2$ to $j+1$	4
BOTT3	"	$j-3$ to $j+2$	5
BOTT4	"	$j-3$ to $j+2$	6

We have finally been working with this sample of seven schemes, whose performances are evaluated in the next chapter.

4.4 Résumé français : "Étude bibliographique des méthodes disponibles"

L'équation de transport peut s'écrire sous forme conservative (notation tensorielle)

$$\frac{\partial \psi}{\partial t} + \frac{\partial u_j \psi}{\partial x_j} = S$$

où u_j est la vitesse de l'écoulement suivant la direction j, S un terme de diffusion et ψ respectivement égal à C, hC, AC en dimension 1, 2 et 3 (C concentration, h hauteur d'eau, A section mouillée). **Compte tenu des propriétés de l'écoulement** (conservation de la masse d'eau), **on peut également la développer sous une forme non-conservative** :

$$\frac{\partial C}{\partial t} + u_j \frac{\partial C}{\partial x_j} = S'$$

Dans ce chapitre nous nous intéressons au premier chef au traitement de l'opérateur d'advection (pour la diffusion, voir annexe E, section E.2), les différents algorithmes étant introduits, pour rester simple, dans un cadre monodimensionnel. **On distingue les méthodes s'appliquant à la forme non-conservative (section 4.1) et à la forme conservative (section 4.2).**

Les méthodes appliquées à la forme non-conservative reposent sur une approche lagrangienne. On vérifie en effet que l'équation d'advection pure

$$\frac{\partial C}{\partial t} + U \frac{\partial C}{\partial x} = 0$$

est équivalente à $\frac{\mathrm{D}C}{\mathrm{D}t} = 0$ où $\frac{\mathrm{D}}{\mathrm{D}t}$ représente la dérivée totale le long de courbes définies par $\frac{\mathrm{D}\vec{X}}{\mathrm{D}t} = \vec{U}$ ($\vec{X}$ et $\vec{U}$ respectivement vecteur des coordonnées et vecteur vitesse du flot), lesquelles sont appelées **courbes caractéristiques de l'écoulement. Les solutions de l'équation d'advection pure sont constantes le long des caractéristiques.**

Si le champ de concentrations C est connu à l'instant de calcul t^n, on peut en déduire son évolution à l'instant $t^{n+1} = t^n + \Delta t$. Soit M_i, d'abscisse x_i, le noeud de calcul No i :

$$C_i^{n+1} = C\left(x_i, t^{n+1}\right) = C\left(x_i - dl, t^n\right) \text{ avec } dl = \int_{t^n}^{t^{n+1}} U(x,t)\, dt$$

Par interpolation, on peut déterminer $C(x_i - dl, t^n)$. (nb : on notera O_i le point d'abscisse $x_i - dl$) Si les vitesses sont uniformes dl est indépendant du noeud et égal à $U\Delta t$.

Les **méthodes lagrangiennes d'ajustement polynômial** (section 4.1.1) ont été initialement développées dans le cas de vitesses et grille de calcul uniformes, vérifiant de plus la condition de Courant-Friedrich-Levy, c'est à dire $c_r = U\Delta t/\Delta x \leq 1$ (c_r est dit nombre de Courant).

- **Le champ de concentration au temps t^n est localement approximé, au voisinage de chaque noeud i, par un polynôme P_i, tel que $P_i(x_{i+j}) = C_{i+j}$ pour $j = 0, \pm 1, \ldots$.** c_r étant plus petit que 1, O_i se trouve dans la maille voisine de M_i. **L'algorithme de résolution se réduit donc à $C_i^{n+1} = P_i[x_i - U\Delta t]$, que l'on réécrit sous la forme $C_i^{n+1} = \sum_j a_j C_{i+j}^n$, les noeuds $i+j$ étant les voisins de i utilisés pour construire P_i.**

  Les a_j sont déterminés en s'appuyant sur un développement en série de Taylor et en tirant parti des relations entre dérivées temporelles et spatiales des solutions de l'équation d'advection. On en déduit que :

$$\sum_j (x_{i+j} - x_i)^k a_j = (-U\,\Delta t)^k \qquad \text{pour } k = 0, 1, \ldots$$

  Pour obtenir un polynôme d'ordre m, on utilisera $m+1$ relations pour $k \leq m$. Pour que les a_j soient définis de façon unique, on devra utiliser m voisins en sus du noeud i. L'approximation résultante est telle que toutes les erreurs de troncature sont d'ordre strictement supérieur à m si U est uniforme et constant. Dans le cas contraire, le schéma n'est que d'ordre 1 en temps (i.e. l'erreur de troncature contient un terme proportionnel à Δt).

- Pour vitesse et pas d'espace uniformes, les a_j sont des fonctions polynômiales de c_r. Si vitesse ou pas d'espace sont variables, les a_j dépendent du noeud i considéré. Si les vitesses sont instationnaires, le calcul des a_j doit être actualisé à chaque pas de temps. Tout ceci alourdit la mise en oeuvre du schéma.

- Ce type de schéma a été étudié, notamment jusqu'à l'ordre 10 inclus par (Tremback *et al.*, 1987). L'analyse de Fourier indique que les schémas d'ordre pair et impair sont plus efficaces respectivement pour limiter les erreurs d'amortissement (norme du facteur d'amplification s'approchant de 1) et de déphasage (argument du facteur d'amplification tendant vers 0). **Après tests numériques et informatiques limités à des grilles de calcul uniformes, (Tremback *et al.*, 1987) conseille l'utilisation d'un schéma d'ordre 6.**

  (Takacs, 1985) se contente de l'étude des algorithmes d'ordre 4 au plus et **recommande celui d'ordre 3, dont il propose une mise en oeuvre en deux temps, sous forme d'un schéma prédicteur-correcteur qui a pour avantage de se prêter efficacement à une généralisation à des vitesses et pas d'espace variables.**

Un défaut des méthodes polynômiales est qu'elles reposent sur une estimation assez sommaire de *dl*. Ce point fait par contre l'objet d'une étape de calcul à part entière dans les méthodes aux caractéristiques (section 4.1.2). *dl* est évalué noeud par noeud en intégrant l'équation $\frac{\mathrm{D}\vec{X}}{\mathrm{D}t} = \vec{U}$ par un algorithme de Runge-Kutta plus ou moins raffiné. Cette fois-ci, le point origine O_i (dit "pied" de la caractéristique) peut être situé partout, sans qu'il soit besoin de vérifier la condition de Courant-Friedrich-Levy. **Pour l'étape d'interpolation, on a de nouveau recours à des approximations polynômiales locales.**

- Un interpolateur couramment utilisé est l'**interpolateur cubique de Hermite**, qui est construit, entre 2 noeuds, à partir de la valeur de la fonction et de ses dérivées premières aux noeuds. Selon l'estimation que l'on utilise pour les dérivées nodales, on aboutit à différents schémas.

Dans la **méthode d'Holly-Preissmann**, on exploite la propriété suivante : si une fonction satisfait l'équation d'advection, il apparaît que sa dérivée spatiale d'ordre 1 vérifie une équation de transport similaire, moyennant l'ajout d'un terme source fonction du gradient de vitesse. On résoud donc l'équation de transport également pour la dérivée, l'interpolateur pour celle-ci étant obtenu par dérivation du polynôme de Hermite.

Rasch et Williamson ont comparé les performances de diverses méthodes d'estimation des dérivées, basées sur des fonctions polynômiales ou rationnelles. Pour les cas d'advection de champs de concentration étroits, aux forts gradients, le meilleur estimateur semble celui d'**Akima** : la dérivée en un point est une combinaison linéaire de ses estimations au 1^{er} ordre, à droite et à gauche, les poids dépendant de la courbure de la fonction de part et d'autre du noeud.

- D'autres interpolateurs sont utilisés. Par exemple, **(Dan N'Guyen, 1988)** propose d'utiliser la combinaison de **séries de Taylor quadratiques** développées à partir des noeuds de calcul encadrant le pied de la caractéristique O_i. Dans la méthode **minimax** de **(Li, 1990)**, on utilise également une série de Taylor quadratique, mais cette fois-ci développée autour du milieu de la maille de calcul contenant O_i.

- L'utilisation de formes polynômiales de degré élévé permet théoriquement d'améliorer la précision de l'interpolation. (nb : on a employé le terme "théoriquement" pour la raison suivante. En passant d'un ordre m à un ordre $m+1$, on passe d'une erreur proportionnelle à Δx^{m+1} à une erreur proportionnelle à Δx^{m+2}, laquelle tend plus vite vers 0, *si* Δx *est petit.* Dans les cas pratiques, si Δx est mal dimensionné, trop grand, on peut observer l'inverse : l'erreur engendrée par une approximation d'ordre $m+1$ devient plus importante que l'erreur engendrée par des interpolateurs de moindre degré). Cependant, en contrepartie, au dessus du degré 1 (interpolation linéaire), ces formes n'assurent pas la conservation de propriétés comme la monotonicité ou la positivité du champ de concentrations. Ceci peut conduire à l'apparition d'extrema parasites (concentrations négatives ou maxima artificiels) dans les régions de fort gradient.

  L'interpolateur d'Hermite a été largement étudié : d'où l'identification des conditions que doivent satisfaire les estimées des dérivées nodales afin que l'interpolateur préserve le caractère convexe ou concave, ou bien la monotonicité des concentrations. Ces conditions peuvent être utilisées pour borner les estimées, quelque soit la méthode employée pour les calculer (Rasch-Williamson, Holly-Preissmann, etc ...).

  Rasch et Williamson ont comparé les performances de versions non-bornées et bornées de l'interpolateur hermitien. Il apparaît que ces dernières améliorent effectivement conservation de la masse et de la positivité mais ceci au prix d'un certain amortissement des pics de concentration. Selon le problème traité, il n'est donc pas systématiquement pertinent d'utiliser une version bornée.

Dans les **méthodes appliquées à la forme conservative**, on partitionne le domaine de calcul en petites cellules définies autour de chaque noeud i par $x_{i-1/2} \leq x \leq x_{i+1/2}$ où $x_{i+1/2} = 0.5\,(x_i + x_{i+1})$. On suppose la masse dans la boîte i égale au produit de la longueur Δx_i de la boîte par la valeur nodale de ψ. **La variation de masse au cours d'un pas de temps** est calculée en intégrant l'équation de transport : elle **est égale à la différence entre**

les flux advectifs sortant et entrant par ses frontières, soit

$$\psi_i^{n+1} = \psi_i^n - \frac{1}{\Delta x_i}\left(F_{i+1/2} - F_{i-1/2}\right) \quad \text{avec} \quad F_{i+1/2} = \int_{t^n}^{t^{n+1}} U(x_{i+1/2},\tau)\psi(x_{i+1/2},\tau)\,d\tau$$

Il y a différentes méthodes d'évaluation des flux $F_{i+1/2}$. La première repose sur l'utilisation du théorème de la moyenne, selon lequel il existe un instant s ($t^n \leq s \leq t^{n+1}$) tel que $F_{i+1/2} = \Delta t\, U(x_{i+1/2},s)\,\psi(x_{i+1/2},s)$. Le champ de vitesses est généralement une donnée du problème, on se concentre alors sur l'estimation de la **valeur à l'interface** $\psi_{i+1/2}$.

- Dans les **méthodes explicites** (section 4.2.2), on suppose tout simplement que $s = t^n$.

  Parmi les méthodes de calcul de $\psi_{i+1/2}$ les plus connues on trouve la méthode UPWIND (dite aussi, schéma d'Euler décentré amont) où l'on affecte à $\psi_{i+1/2}$ sa valeur dans la cellule "amont", d'où vient le flot, à savoir $\psi_{i+1/2} = \psi_i$ si $U_{i+1/2} \geq 0$, $\psi_{i+1/2} = \psi_{i+1}$ sinon. Cette méthode est trés stable mais induit malheureusement un degré d'amortissement inacceptable dans le cas de pics de concentration marqués.

  Cet amortissement peut éventuellement être corrigé explicitement (application d'une "anti-diffusion", cf (Smolarkiewicz, 1983)). Une autre stratégie est d'avoir recours à des approximations d'ordre plus élevé. Par exemple, dans la populaire méthode QUICK (Leonard, 1979), $\psi_{i+1/2}$ est issue d'une approximation quadratique construit sur les noeuds i et $i+1$ ainsi que sur le noeud "amont" le plus proche (i.e. $i-1$ si $U_{i+1/2} \geq 0$, $i+2$ sinon).

  On se heurte de nouveau au problème du non-respect de la positivité et de la monotonicité. Pour pallier cette difficulté, on peut *limiter* les estimées des valeurs de ψ aux interfaces (cf section 4.2.3), de même que l'on pouvait contraindre les estimées des dérivées nodales avant d'appliquer l'interpolateur hermitien. Une autre approche (Leonard, 1981; Leonard, 1988) est de substituer à l'approximation polynômiale, dans les zones de fort gradient, une approximation par des morceaux d'exponentielle (algorithme SHARP, sec. 4.2.2). Cet algorithme performant a malheureusement un domaine de stabilité restreint ($c_r \leq 0.2$).

- La **transcription implicite** des divers schémas explicites évoqués ci-dessus pose quelques difficultés, que le schéma résultant soit peu performant (c'est le cas de QUICK) ou que la mise en oeuvre soit délicate (par exemple, SHARP reposant sur des formules d'estimation non-lineaire, le passage en implicite requiert des itérations).

La deuxième méthodologie d'estimation des $F_{i+1/2}$ repose d'abord sur une réécriture de ceux-ci. On peut en effet remarquer que **seules les particules d'eau qui sont situées en deçà d'une certaine distance de $x_{i+1/2}$ seront susceptibles de traverser la frontière correspondante de la cellule.** Par conséquent, $F_{i+1/2}$ est égal à $\int_{x_{i+1/2}-dl}^{x_{i+1/2}} \psi(x,t^n)\,dx$.
dl a une forme comparable au dl des méthodes non-conservatives. Il est le plus souvent estimé par $dl = U_{i+1/2}\Delta t$. Ensuite, pour calculer $F_{i+1/2}$, il nous reste à approximer sur chaque cellule ψ par une fonction ϕ_i intégrable et à **intégrer cette distribution approchée**. Les schémas de cette catégorie sont naturellement explicites. ϕ_i est de type polynômial. Il existe différentes façons de la construire (cf section 4.2.4) :

- **bâtir le polynôme qui passe par les points** $\{(x_{i+j},\psi_{i+j}),\, j = j_{min},\dots j_{max}\}$

On peut choisir de "centrer" le polynôme ϕ_i autour du point interface $x_{i+1/2}$. Si ϕ_i est d'ordre m impair, les noeuds concernés seront répartis symétriquement autour de $i+1/2$: il s'agira des noeuds $i-(m-1)/2$ à $i+1+(m-1)/2$. Si m est pair, on rajoute un noeud du côté "amont" de $i+1/2$. On peut alternativement centrer ϕ_i autour du noeud i lui-même (Bott, 1989a). On débouche sur des expressions légèrement différentes. Par exemple, si les vitesses sont positives, les approximations cubiques centrées autour de $i+1/2$ et i reposeront respectivement sur les noeuds $i-1$ à $i+2$ et $i-2$ à $i+1$.

Le schéma bien connu QUICKEST utilise pour ϕ_i l'approximation quadratique basée sur les noeuds $i-1$, i et $i+1$. Il peut être vu comme la version implicite de QUICK. On peut vérifier que, dans le cas de vitesse et grille de calcul uniformes, il est équivalent au schéma de Takacs d'ordre 3 mentionné plus haut. Toutefois, les généralisations de QUICKEST et TAKACS aux situations non uniformes sont distinctes.

- **utiliser un développement en série de Taylor autour du noeud i**
  Au delà du second ordre, ces polynômes diffèrent de la première famille car ils ne satisfont plus $\phi_i(x_{i+j}) = \psi_{i+j}$ pour $|j| > 1$.

- **utiliser des polynômes dont les moments, sur leur intervalle de définition, sont égaux, jusqu'à un certain ordre, à ceux de la distribution ψ.**

  Ainsi, Bott (Bott, 1989b) propose des polynômes "conservatifs" d'ordre 2 et 4 définis de façon que leur intégrale sur les cellules centrées autour des noeuds i, $i \pm 1$ ($i \pm 2$) coincide avec la masse contenue dans ces cellules (soit $\Delta x_{i+j}\psi_{i+j}$, $j = 0, \pm 1, \pm 2$).

  On peut alternativement opter pour des polynômes qui, au lieu de préserver la masse (moment d'ordre 0) sur *plusieurs* cellules, assurent l'égalité de *plusieurs* moments, d'ordre plus éléve, sur *une seule* cellule.

- **combiner plusieurs approches**
  Par exemple, (Bott, 1989a) suggère également de bâtir ϕ_i en deux étapes : d'abord par ajustement polynômial entre noeuds voisins de i; ensuite, en normalisant l'expression ainsi obtenue afin qu'elle préserve la masse sur la cellule i uniquement.

- Le comportement de flux provenant de l'intégration de fonctions plus ou moins compliquées est a priori moins évident à cerner que celui d'estimées ponctuelles. Toutefois, la bibliographie permet de conclure que l'on retrouve des problèmes similaires. Quand les flux sont basés sur des formes polynômiales de degré élevé on peut rencontrer des phénomènes d'extrema parasites (concentrations négatives ou au contraire surestimées). Quand l'approximation est d'ordre faible (fonctions constantes par morceaux ou linéaires), la diffusion numérique est trop importante. Là encore, on peut **borner ou pondérer les estimées des flux** (cf section 4.2.5), comme on l'a fait pour les estimées ponctuelles.

  Parmi les approches pour ce faire, on trouve la famille des **schémas FCT (Flux-Corrected-Transport)**, qui sont construits ainsi :

  1. On calcule une première estimation $F^L_{i+1/2}$ à partir d'une approximation d'ordre faible, mais préservant le caractère monotone, puis une seconde estimation $F^U_{i+1/2}$, issue d'un ordre élevé.
  2. On estime la diffusion numérique induite par le premier schéma en soustrayant $F^U_{i+1/2}$ à $F^L_{i+1/2}$. Plus précisément, on définit le flux anti-diffusif par $A_{i+1/2} = F^H_{i+1/2} - F^L_{i+1/2}$.

3. On calcule une solution intermédiaire à l'aide du schéma de faible ordre,
   $\psi_i^* = \psi_i^n - \left(F_{i+1/2}^L - F_{i-1/2}^L\right)/\Delta x_i$
4. On limite les flux anti-diffusifs de façon à ce que la solution finale (cf point 5) ne présente pas de comportement indésirable, soit $A_{i+1/2}^C = \lambda_{i+1/2} A_{i+1/2}$. Le "limiteur" est basé sur les propriétés de la solution intermédiaire ψ^* et éventuellement de la distribution initiale ψ^n (voir ses différentes versions en 4.2.5).
5. La solution finale est obtenue par correction de la solution intermédiaire,
   soit $\psi_i^{n+1} = \psi_i^* - \left(A_{i+1/2}^C - A_{i-1/2}^C\right)/\Delta x_i$.

Les méthodes FCT se sont révélées trés adéquates pour le traitement de discontinuités (fronts de concentration en marches d'escalier). Cependant, le transport correct de distributions étroites aux pics marqués requiert l'emploi d'un second schéma d'ordre *vraiment* élévé (ordre 8, cf (Zalesak, 1979)).

On peut également se contenter de **borner les flux afin d'assurer seulement la positivité du schéma** (e.g. (Smolarkiewicz, 1983; Bott, 1989a)). En ce cas, on peut omettre la première phase des FCT, à savoir le calcul d'une solution intermédiaire par un schéma de faible ordre, et appliquer directement les limitations aux flux estimés d'après l'approximation d'ordre élévé. Les conditions de positivité sont développées en fin de section 4.2.5. Si les approximations polynômiales vérifient par construction certaines propriétés (c'est le cas par exemple des polynômes conservatifs de Bott), les conditions s'en trouvent allégées.

En bref, il apparaît qu'il existe de nombreux chemins possibles pour résoudre l'advection. Il est difficile de se faire une idée exacte de leur potentiel d'après une simple bibliographie. En effet, les performances de ces schémas sont exposées de façon hétérogène, sur des tests et dans des conditions d'application différentes. L'indication de leur coût informatique est de plus souvent absente. Fort heureusement, dans le contexte du forum Convection-Diffusion (Baptista *et al.*, 1988) a été entrepris depuis quelques années la constitution d'une base de cas-tests de référence sur laquelle nous nous appuierons pour évaluer objectivement les schémas. Cependant, comme il ne nous est pas loisible d'approfondir l'étude de toutes les méthodes, il nous faut bien en sélectionner quelques unes *a priori*.

- Replaçons nous d'abord dans le cadre de notre première application in situ, les déversements d'orage en Seine. Nous devrons donc traiter du devenir de sources *ponctuelles* de pollution, présentant des différences de concentration élévées avec le milieu récepteur. *Les gradients au voisinage des points de rejet seront donc forts.* Par ailleurs, les études antérieures sur la Seine montrent que *ces gradients, notamment transverses, tendent à perdurer* (longueur de mélange de plusieurs kilomètres, cf 2.4.4), *l'advection étant dominante. Les phénomènes sont enfin trés transitoires.*

  On s'intéressera donc en premier lieu aux méthodes qui semblent donner de bons résultats dans le cas du transport de distributions de taille limitée, aux extrema marqués.

- Dans ce cadre d'application, les **méthodes aux caractéristiques** bénéficient d'une bonne réputation, quoiqu'on leur reproche fréquemment leur coût.

  La méthode d'**Holly-Preissmann** est souvent invoquée comme une référence. Nous la retiendrons donc, ainsi qu'une autre méthode a priori plus "économique" utilisant l'interpolateur hermitien,

à savoir la **méthode de Rasch-Williamson** basée sur l'estimateur d'Akima. Nous testerons également deux autres méthodes s'appuyant sur un interpolateur seulement quadratique, à savoir les schémas de **Dan N'Guyen** et du **minimax**.

- Faisons maintenant le tour des méthodes résolvant la forme conservative de l'équation d'advection.

  Nous n'avons pas retenu de méthode s'appuyant sur la détermination d'une valeur à l'interface. Les méthodes explicites précises de cette catégorie (e.g. SHARP) ont des contraintes de stabilité trés restrictives et leur passage à une formulation implicite est fort complexe.

  Parmi les méthodes où les flux advectifs sont calculés par intégration d'une approximation polynômiale des distributions, nous ne nous sommes intéressés qu'aux méthodes où le polynôme est au moins d'ordre 2, les autres semblant affectées par trop de diffusion numérique.

  Il semble naturel d'inclure dans notre panel le schéma **QUICKEST** qui est trés largement répandu et bénéficie d'une excellente réputation malgrè (?) sa simplicité.

  Pour des ordres plus élévés, nous nous sommes néammoins fixé une limite, à savoir de nous en tenir aux schémas où la distribution approchée n'est pas basée sur plus de noeuds que dans le cas de notre méthode aux caractéristiques la plus "gourmande" (à savoir, pour l'interpolation dans la maille i, l'utilisation de 6 valeurs nodales, des No $i-3$ à $i+2$). C'est pourquoi nous avons par exemple écarté les méthodes FCT (ordre 8, soit 9 noeuds, nécessaire pour traiter les pics de concentration). On est donc amené à faire un tri entre les méthodes où le polynôme approximant est d'ordre 3 ou 4. Ce tri préliminaire s'est appuyé sur quelques tests restreints, détaillés en annexe F.1. En conclusion, **nous avons retenu deux algorithmes**, baptisés par la suite, **BOTT3 et BOTT4, où les approximations sont respectivement un développement cubique en série de Taylor et un polynôme d'ordre 4 bâti sur la propriété de conservation de la masse. Les deux algorithmes sont appliqués avec un limiteur garantissant la positivité.**

Ces 7 schémas sont étudiés en détail dans les 3 chapitres suivants.

Chapter 5

Fourier analysis of selected schemes

Fourier analysis applies to linear partial difference equations (PDE) and their *linear* finite difference approximations (FDE). As explained in Appendix C.3.1, if a PDE is linear, any of its solutions can be described as a linear superposition of sinusoids (Fourier Series). Similarly, the corresponding FDE solutions can be written as a Fourier series. Yet, only the sinusoids whose wavelength is a multiple of the grid spacing Δx can be accounted for in the FDE solutions.

For each elementary Fourier component φ_k, with wavelength $k\Delta x$, the PDE and related FDE reduce respectively to :

$$\begin{aligned} \varphi_k^{n+1} &= \mathcal{G}_{\text{ex}}(k) \quad \varphi_k^n \\ \varphi_k^{n+1} &= \mathcal{G}_{\text{num}}(k)\, \varphi_k^n \end{aligned}$$

the superscript n referring to the φ_k value at time level n.
$\mathcal{G}_{\text{ex}}$ is termed the *exact amplification factor* of the PDE, $\mathcal{G}_{\text{num}}$ is the *numerical amplification factor* of the FDE. As the FDE is but an approximation to the PDE, each φ_k is propagated with some amplitude error and at a somewhat erroneous celerity. This is reflected respectively by discrepancies between the modulus and arguments of $\mathcal{G}_{\text{num}}$ and $\mathcal{G}_{\text{ex}}$ (cf App. C.4).

The exact amplification factor of the one-dimensional advection equation with constant velocity U, time step Δt and grid spacing Δx, is, for wavelength $k\Delta x$ (cf C.4)

$$\mathcal{G}_{\text{ex}}(k) = \exp -jc_r\theta_k \quad \text{with} \quad \theta_k = \frac{2\pi}{k} \quad \text{and} \quad c_r = \frac{U\Delta t}{\Delta x} \tag{5.1}$$

The Fourier analysis cannot be applied to each of our selected seven schemes (cf sec. 4.3). The RASCH backward characteristic method relies on a non-linear derivative estimate; the BOTT3 and BOTT4 algorithms make use of a non-linear renormalization of the advective fluxes. Consequently, the Fourier analysis is possible only for Holly-Preissmann, Dan N'Guyen, Minimax and TAKACS/QUICKEST algorithms.

The Fourier analysis of Holly-Preissmann algorithm is somewhat more complicated than the study of other schemes. Indeed, with this method we solve a transport equation both for the concentration value C and its first-order spatial derivative C_x so that the FDE reads in fact

$$\begin{pmatrix} \varphi^{n+1} \\ \varphi_{x\,k}^{n+1} \end{pmatrix} = K_{\text{num}} \begin{pmatrix} \varphi^{n} \\ \varphi_{x\,k}^{n} \end{pmatrix}$$

where φ_k and $\varphi_{x\,k}$ denote the Fourier components of wavelength $k\Delta x$ for the concentration and its derivative respectively and K_{num} is a 2×2 matrix. The exact transition matrix is :

$$K_{\text{ex}} = \begin{pmatrix} \exp -jc_r\theta_k & 0 \\ 0 & \exp -jc_r\theta_k \end{pmatrix}$$

The error introduced by the finite-difference approximation is determined by finding the eigenvalues and eigenvectors of K_{num} then comparing the former with $\exp -jc_r\theta_k$. It appears that K_{num} has two complex eigenvalues λ_1 and λ_2 : the first tends toward $\exp -jc_r\theta_k$ as θ_k tends toward zero (i.e. as the wavelength increases) while the second one is very different from the exact eigenvalue (Holly & Preissmann, 1977). The eigenvectors related to λ_1 and λ_2 are termed primary and secondary modes. According to (Holly & Preissmann, 1977), they represent respectively the proper and parasitic solutions resulting from the use of the Holly & Preissmann approximation to the transport equation.

The results of the Fourier analysis are illustrated by figures 5.1 through 5.5 :

- The solid lines refer to Courant numbers $c_r = 0.5, 1$
- The uneven dashed lines refer to Courant numbers $c_r = 0.1, 0.9$
- The dotted lines refer to Courant numbers $c_r = 0.2, 0.8$
- The short dashed lines refer to Courant numbers $c_r = 0.3, 0.7$
- The long dashed lines refer to Courant numbers $c_r = 0.4, 0.6$

1. We have not been plotting the celerity error for the secondary mode of Holly-Preissmann scheme : it reaches a few hundred percent for all wavelengths ! As pointed out in (Holly & Preissmann, 1977), the amplitude error of the secondary mode is so important for all wavelengths (see fig. 5.5) that we can expect this parasitic component to be smoothed out of the solution after a few time steps. Consequently, it is chiefly the behaviour of the primary mode which controls the accuracy of the Holly-Preissmann scheme : from now on, we shall focus only on this primary mode.

2. As exposed in the previous chapter, all schemes are exact for $c_r = 1$. Consequently, there is neither amplitude nor celerity error for this Courant number value.

3. All schemes exhibit symmetry with respect to $c_r = 0.5$. More precisely, the amplitude errors are the same for c_r and $1 - c_r$ and the phase shifts (i.e. the difference between the arguments of exact and numerical amplification factors) are opposite for c_r and $1 - c_r$.

   For $c_r = 0.5$ the amplitude error is always maximum while the phase (or celerity) error is cancelled.

   Quickest, Dan N'Guyen and Minimax algorithms have a prevailing lagging phase error for $c_r < 0.5$ and a leading phase error for $c_r > 0.5$. The Holly-Preissmann scheme exhibits the opposite behaviour.

4. For all schemes, the amplitude error appears to be negligible for wavelengths superior to $8\,\Delta x$. Considering only this amplitude error, the schemes rank as follows :

   $$\text{Holly-Preissmann} > \text{Dan N'Guyen} > \text{Quickest} > \text{Minimax}$$

   Yet, while the superior accuracy of Holly-Preissmann method is obvious, the last three schemes yield relatively close results.

5. Considering only celerity error, the schemes rank as follows :

   $$\text{Holly-Preissmann} > \text{Minimax} > \text{Quickest} > \text{Dan N'Guyen}$$

   Apart from the Dan N'Guyen scheme which has a slow decreasing phase error (see fig. 5.2), the algorithms have negligible celerity error for wavelengths superior to $8\,\Delta x$. Once again, the superiority of Holly-Preissmann algorithm is overwhelming.

6. In conclusion, apart from the possible exception of the Dan N'Guyen scheme with its lasting phase error, all schemes should perform quite well when dealing with the transport of smooth shapes, characterized with long wavelengths ($\geq 8\,\Delta x$).

   The Holly-Preissmann scheme should undoubtedly prove to be the most accurate method. The relative grading of the three other schemes appears more difficult to forecast, as they rank differently according to amplitude and celerity criteria.

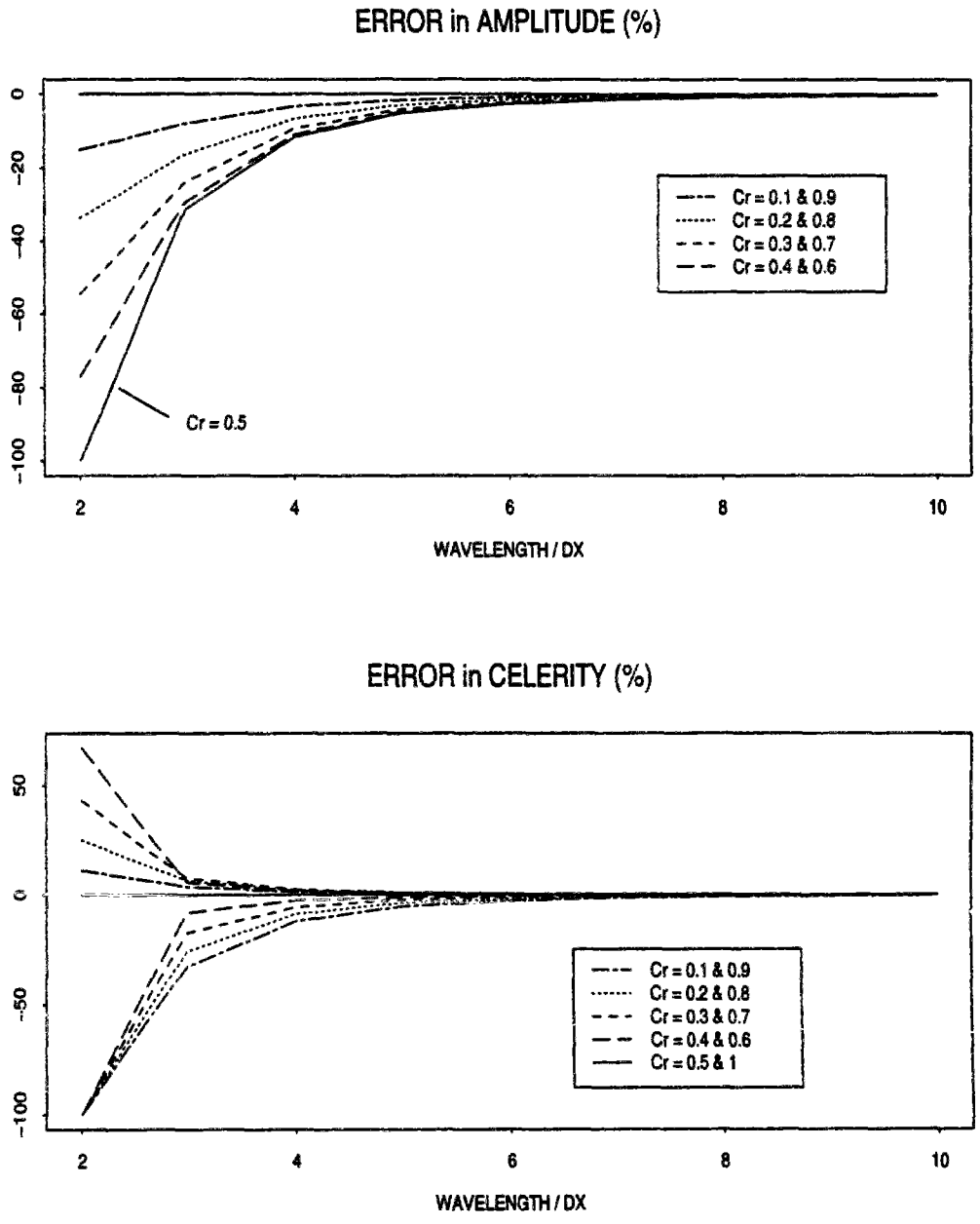


Figure 5.1: Fourier Analysis of TAKACS/QUICKEST scheme

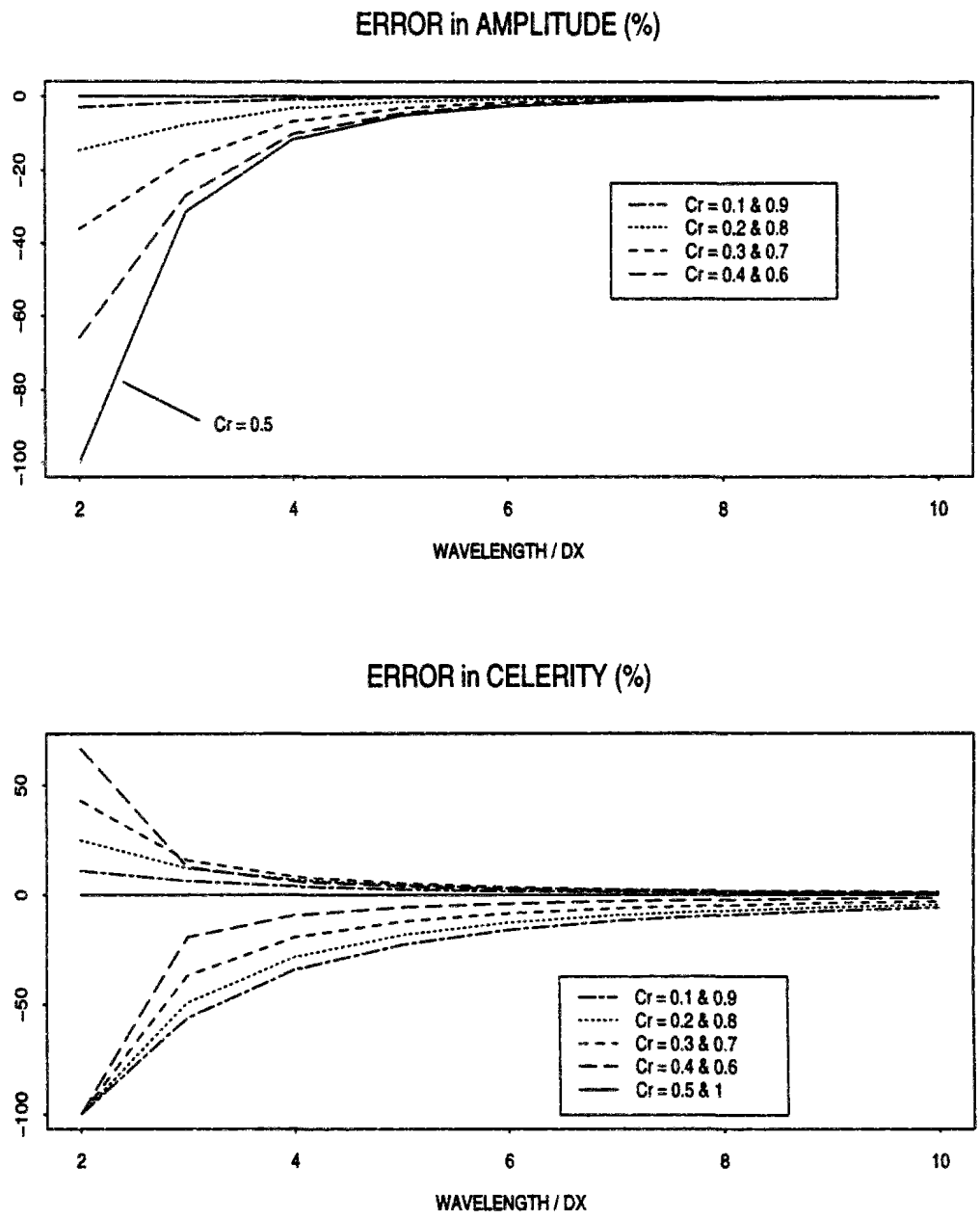


Figure 5.2: Fourier Analysis of Dan N'Guyen scheme

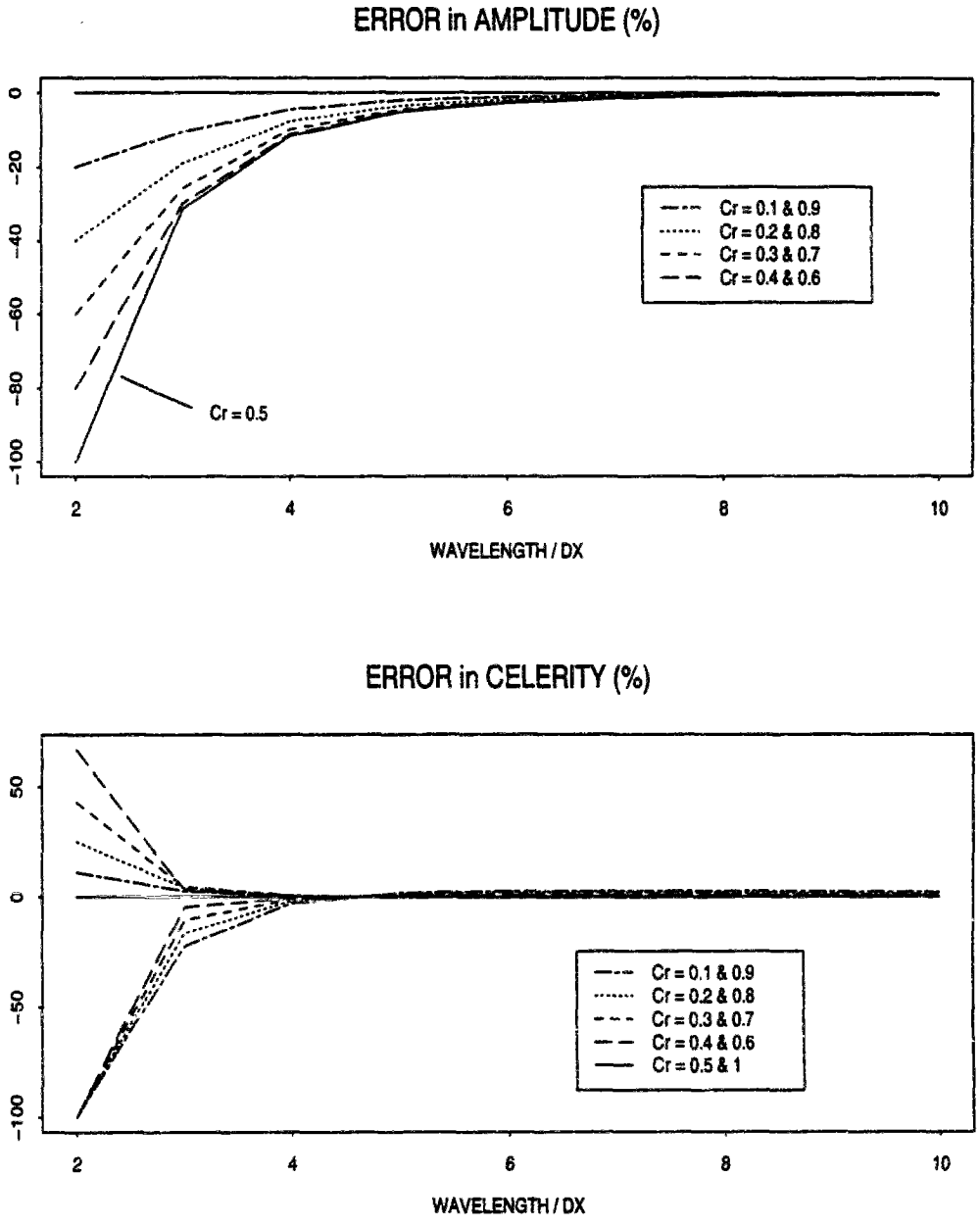


Figure 5.3: Fourier Analysis of Minimax scheme

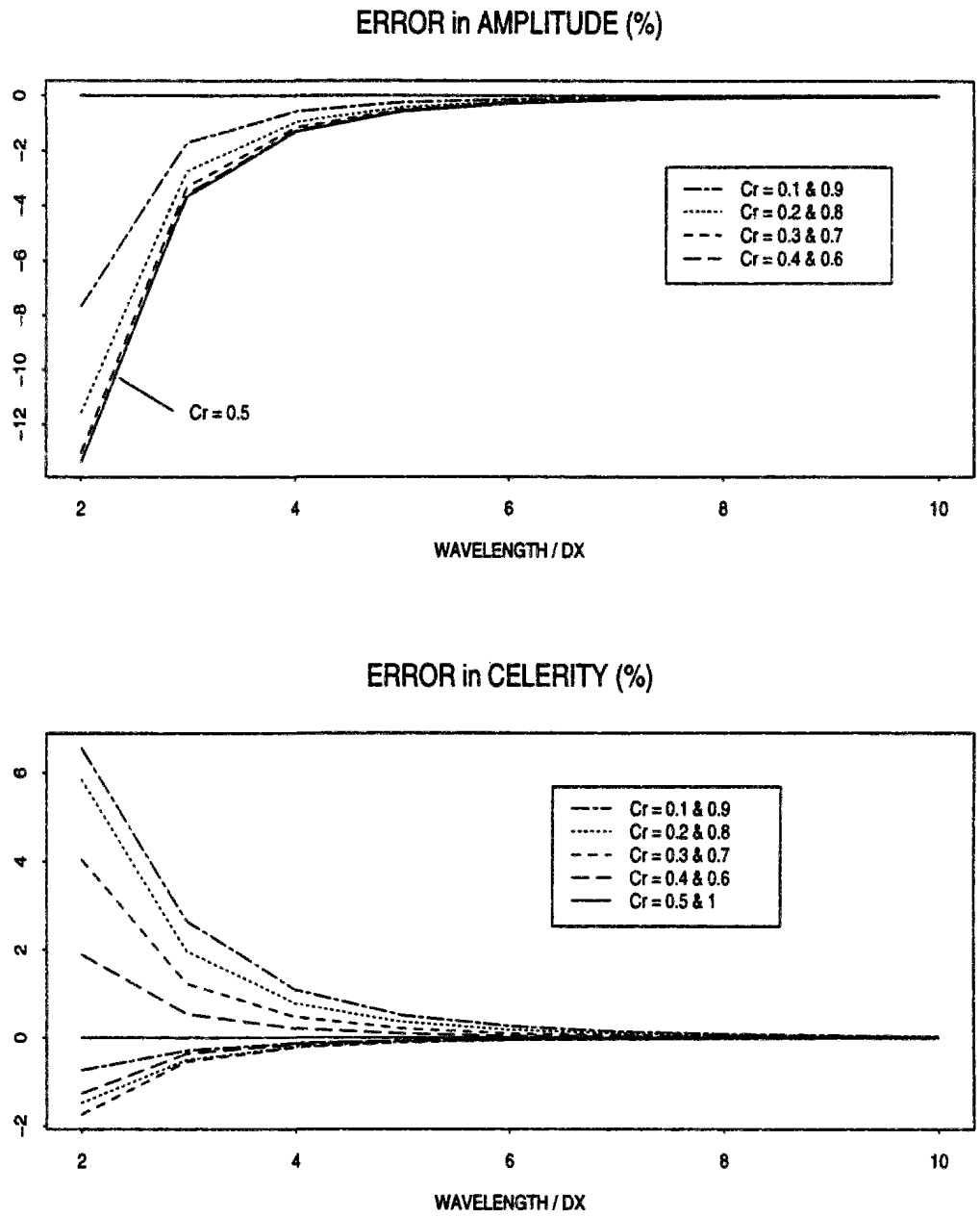


Figure 5.4: Fourier Analysis of Holly-Preissmann scheme - Primary mode

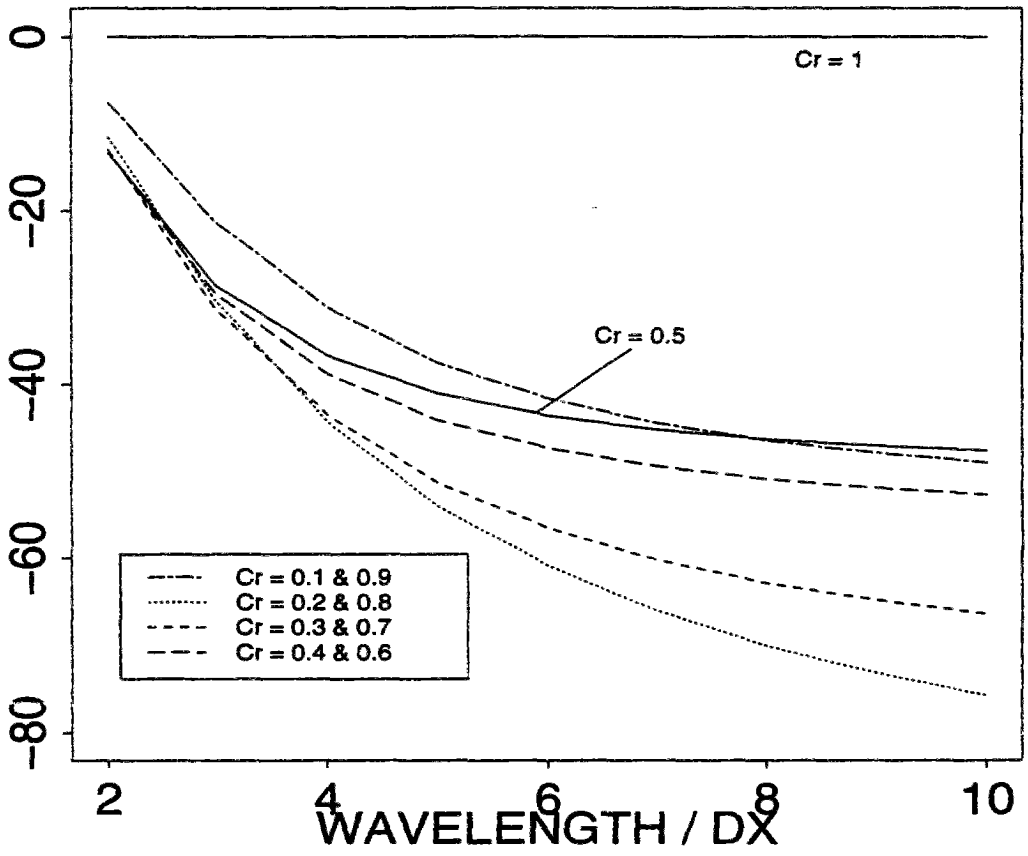


Figure 5.5: Amplitude error (%) of Holly-Preissmann scheme - Secondary mode

Résumé du Chapitre :
"Analyse de Fourier des schémas retenus"

L'analyse de Fourier est applicable aux équations linéaires aux dérivées partielles (EDP) et à leurs approximations aux différences finies (MDF) linéaires. Si une EDP est linéaire, chacune de ses solutions peut être écrite sous forme de la combinaison linéaire de sinusoides (une série de Fourier). Similairement, les solutions de la MDF appliquée peuvent être développées en série de Fourier. Cependant, seules les composantes dont la longueur d'onde est un multiple du pas d'espace peuvent être prises en compte dans la solution de la MDF.

Pour chaque composant élémentaire de la série de Fourier φ_k, de longueur d'onde associée $k\Delta x$, l'EDP et la MDF appliquée se réduisent respectivement à :

$$\begin{aligned} \varphi_k^{n+1} &= \mathcal{G}_{\text{ex}}(k) \quad \varphi_k^n \\ \varphi_k^{n+1} &= \mathcal{G}_{\text{num}}(k)\, \varphi_k^n \end{aligned}$$

(l'indice n se référant à l'instant de calcul numéro n).
$\mathcal{G}_{\text{ex}}$ et $\mathcal{G}_{\text{num}}$ sont dénommés respectivement le **facteur d'amplification exact** de l'EDP et le **facteur d'amplification numérique** de la MDF. Comme la MDF n'est qu'une approximation de l'EDP, chaque φ_k est propagé avec quelque erreur d'amplitude et de célérité, qui se reflètent respectivement dans les différences entre les modules et les phases de $\mathcal{G}_{\text{ex}}$ et $\mathcal{G}_{\text{num}}$. L'analyse de Fourier consiste dans le calcul et l'étude de ces différences pour les longueurs d'onde pertinentes. Cependant, les facteurs d'amplification ne peuvent être déterminés que pour des cas simples. Dans le cadre qui nous intéresse, à savoir l'équation d'advection, on doit se restreindre à une situation monodimensionnelle, avec une vitesse uniforme et constante, et un pas d'espace également constant.

L'analyse de Fourier ne peut être appliquée à l'ensemble des 7 schémas que nous avons sélectionnés (cf sec. 4.3). La méthode aux caractéristiques de Rasch-Williamson fait usage d'estimateurs non-linéaires des dérivées nodales; les algorithmes BOTT3 et BOTT4 incluent une renormalisation non-linéaire des flux advectifs. Par conséquent, **l'analyse de Fourier ne pourra concerner que les schémas d'Holly-Preissmann, de Dan N'Guyen, du Minimax et de QUICKEST/TAKACS.**

Les erreurs d'amplitude et déphasage sont fonction du nombre de Courant $c_r = U\Delta t/\Delta x$ ou de sa partie décimale. Tous les schémas étudiés ont des erreurs symétriques par rapport à $cr = 0.5$, c'est à dire qu'on observe pour c_r et $1 - c_r$ respectivement même erreur d'amplitude et déphasage opposé.

L'erreur d'amplitude des 4 schémas est négligeable pour les longueurs d'onde supérieures à $8\Delta x$. Holly-Preissmann est largement supérieur aux 3 autres, dont les résultats sont proches

(Dan légèrement meilleur que Quickest, lui-même un peu plus performant que Minimax).

En ce qui concerne l'erreur de phase, Holly-Preissmann est encore une fois bien meilleur. Les résultats de Minimax et Quickest sont quasi-équivalents, avec un léger avantage au premier schéma. Dan se comporte beaucoup moins bien. Notamment, son erreur de phase décroit lentement quand la longueur d'onde s'accroît, contrairement à ce qui se passe pour les 3 autres schéma pour lesquels elle est négligeable au delà de $8\Delta x$.

En résumé, l'analyse de Fourier conclut à la supériorité indiscutable du schéma d'Holly-Preissmann. Le rang respectif des 3 autres schémas est plus difficile à établir, puisqu'ils se classent différemment selon les critères de préservation d'amplitude et de phase. On peut se faire quelque souci quant au déphasage que semble exhiber le schéma de Dan.

Chapter 6

One-dimensional test cases

6.1 Design of test cases

As mentioned in the conclusion of chapter 4, literature review is not sufficient to assess fairly numerical schemes. Indeed, papers include generally more talk than testing, as the theoretical presentation of the method already requires some space and as the papers length is usually quite limited. Besides, applied tests are not always relevant with respect to the needs of prospective users. They are not the same for all schemes. Finally, they are not always well documented so that it is difficult to reproduce them.

Ideal tests should allow one to evaluate unambiguously the level of performance of different algorithms. Thus, ideally, these are situations where some reference solution is available. Unfortunately, due to the complexity of flow and transport equations, these cases are few. As regards the transport equation, most of them have been compiled by the organizers of the Convection-Diffusion Forum (Baptista *et al.*, 1988) who intend to provide a common framework for evaluating advection-diffusion solutions. We judged it was fair to support their effort, in our own limited way. Our test cases are thus inspired by those proposed in (Baptista *et al.*, 1988; Rasch & Williamson, 1990) : they constitute somehow an extension of the test problems suggested in both papers. Indeed, we tried for instance more time steps and, for cases including dispersion, more values of the diffusivities, in order to get a more complete picture of the behaviour of our selected schemes.

Each of the performed tests is designed to investigate some practical problem faced in "real life" :

- Tests about pure advection of concentration hills (section 6.3.1) and of advection & dispersion of gauss-hills (section 6.4) mimic what occurs in the immediate vicinity of a point

source. There, we find a concentration field defined by few points and characterized by sharp gradients.

- In section 6.3.1, the problem of what is the required number of points to define properly initial conditions is briefly tackled. Indeed, we shall notice a significant improvement when dealing with wider pollutographs.

- The advection combined with dispersion tests (section 6.4) helps one to figure out what are the acceptable domains of application of various schemes. This is obtained by drawing error isolines as a function of both Courant and Peclet numbers (the latter one is a kind of measure of the respective strengths of advection and diffusion phenomena). For instance, we shall see that if one wants less than 10 % error on the peak concentration, use of flux-form (QUICKEST, BOTT3 & BOTT4) methods should be restricted to Peclet numbers less than 10.

- Tests with an irregularly-spaced grid (sec 6.3.2) or an unsteady flow (sec 6.3.3) are useful because they oblige one to program the complete version of the algorithms, without simplification, so that their computational requirements are more equitably judged. Besides, dealing with uneven grid spacing or unsteady flow is commonplace in real life applications.

- Finally, tests presented in section 6.5 interestingly combine two practical difficulties : a correct definition of the pollutant inputs into the studied domain and the transport of rather narrow sources.

Analysis of the tests results provide much more information regarding the relevance of each algorithm than can yield a theoretical accuracy study like the Fourier analysis performed in the previous chapter. A quantitative basis for assessing the performance of the schemes is supplied by all or part of the set of error measures defined in table 6.1 (extracted from (Baptista *et al.*, 1988)). These measures cover a variety of features of the numerical solution. Such a broad range of error measures is felt to be necessary, as different numerical methods may tend to yield different dominant kinds of errors and as the different errors may prove more or less relevant with respect to the planned application.
(nb : in table 6.1 the superscript e refers to the exact solution; $m(t)$ denotes the total mass of exact solution C^e).

Table 6.1: Error measures for 1D convection-diffusion tests

Sym--bol	Description	Definition	Comments & Exact Value	
Φ	L2 error norm normalized by total mass	$\Phi(t) = \frac{1}{m(t)} \left[\int_\Omega (C(x,t) - C^{ex}(x,t))^2 \, dx \right]^{1/2}$	Integral measure overall error of numerical solution	0
ϵ	Relative error on peak concentration	$\epsilon(t) = \frac{C^{ex}_{\max}(t) - C_{\max}(t)}{C^{ex}_{\max}(t)}$	Point measure of numerical damping	0
Ψ	Absolute value maximal neg. concentration, normalized by ex. peak val.	$\Psi(t) = \left\lvert \frac{C_{\max,\mathrm{neg}}(t)}{C^{ex}_{\max}(t)} \right\rvert$	Point measure of spurious oscillations (wiggles) in numerical solution	0
μ_0	0^{th}-moment of conc. field, normalized by exact value	$\mu_0(t) = \frac{1}{m(t)} \int_\Omega C(x,t) \, dx$	Integral measure of mass preservation	1
ξ	Error in peak position, normalized by exact travel distance	$\xi(t) = \frac{x^{ex}_{\max}(t) - x_{\max}(t)}{Ut}$	Point measure of phase shift introduced by numerical solution	0
μ_x	Error in center of mass position normalized by exact travel distance	$\mu_x(t) = \frac{E^{ex}(t) - E(t)}{Ut}$ with $E(t) = \frac{\int_\Omega xC(x,t) \, dx}{\int_\Omega C(x,t) \, dx}$	Integral measure of phase shift introduced by numerical solution	0
μ_{xx}	Centered 2^{th}-moment of conc. field, normalized by exact value	$\mu_{xx}(t) = \frac{\int_\Omega [\, x - E(t) \,]^2 \, C(x,t) \, dx}{\int_\Omega [\, x - E^{ex}(t) \,]^2 \, C^{ex}(x,t) \, dx}$	Integral measure of numerical spreading	1

6.2 Tests presentation

The set of test problems concerns only uniform flows. It is governed by the equation :

$$\frac{\partial C}{\partial t} + U.\frac{\partial C}{\partial x} = \Gamma \frac{\partial^2 C}{\partial x^2} \qquad (-\infty < x < \infty) \tag{6.1}$$

where U (m/s) and Γ (m^2/s) denote respectively the velocity and the longitudinal dispersion coefficient.

1. **Flow conditions**
   Two kinds of uniform flows have been applied :

$$\begin{aligned} \text{steady flow} \quad U &= 0.5\,\text{m.s}^{-1} \\ \text{unsteady sinusoidal flow} \quad U &= 1.5 \sin\left(\frac{2\pi t}{10800}\right) \text{m.s}^{-1} \end{aligned}$$

   The studied values of dispersion are indicated hereafter, case by case.

2. **Initial conditions**
   Tests concern both the transport of various kinds of concentration-hills and the propagation of a steep front. The corresponding initial concentration fields $C(x,0) = C^0(x)$ are defined as follows :

$$\begin{aligned} \text{Gauss-hill,} \quad C^0(x) &= \exp\left\{-\frac{(x-x_0)^2}{2.\sigma_0^2}\right\} \\ \text{Triangle-hill,} \quad C^0(x) &= \begin{cases} 1 - \dfrac{|x-x_0|}{l_0} & \text{if } |x-x_0| \le l_0 \\ 0 & \text{otherwise} \end{cases} \\ \text{Square-wave,} \quad C^0(x) &= \begin{cases} 1 & \text{if } |x-x_0| \le l_0 \\ 0 & \text{otherwise} \end{cases} \\ \text{Steep front,} \quad C^0(x) &= 0 \end{aligned}$$

   x_0 corresponds to the center of mass of the initial concentration-hills.
   σ_0 denotes the standard deviation of the gauss-hill, l_0 the half-length of the triangle or square-wave distribution. For gauss-hills, if $|x - x_0| \ge 3\sigma_0$, $C^0(x)$ is less than 1 % of the peak concentration $C^0(x_0)$, so that the length of the gauss-hill is approximately 6 σ_0.

   Running the Holly-Preissmann algorithm requires furthermore to initialize the derivatives. We have been testing two options : setting all derivatives to zero or computing them from $C^0(x)$ with the help of an Akima derivative estimate. The second choice yields systematically better results, as illustrated in Appendix F.2.

3. **Space discretization**
   The domain under study is defined by $0 \le x \le 13000$m.

Two alternative kinds of discretization have been used.
$x(1) = 0$ for both grids and, for $2 \leq i \leq 66$:

$$\begin{aligned} \text{Grid 1} \quad x(i) - x(i-1) &= 200. \\ \text{Grid 2} \quad x(i) - x(i-1) &= 200. - 75.\cos\frac{2.\pi\,(i-1)}{65} \end{aligned}$$

In grid 2, the grid spacing varies between $\pm$ 40 % of its average value of 200 m. However, the grid varies smoothly : adjacent mesh width ratios lie in the range 0.96 to 1.04 .

4. **Temporal discretization**

Numerical and exact solutions have been compared at $t = 10800$s for all trials.

For steady flow and grid 1, the simulations have been performed for the time steps indicated in table 6.2. In this table, we have also mentioned the number of iterations required to reach the final computational time ($\mathcal{N}$) and the related Courant number (c_r). Backward characteristics methods do not need to comply with the stability requirement, $c_r \leq 1$ all over the computational domain, and can thus be tested for a wider range of time steps.

The utilisation of non-uniform grid 2 in combination with steady flow conditions lowers the maximum allowable time step for flux-form methods. The local Courant number varies now all over the computational domain : $c_r \leq 1$ everywhere if and only if $\Delta t \leq 240$s. Table 6.3 indicates the spatial average of c_r for each tested time step.

The time steps tested for unsteady flows are given in table 6.4. In this table we have indicated the maximum instantaneous Courant number reached during the simulation c_r(max) and the average Courant number, c_r(moy), whose estimation is based on the average velocity modulus, approximately 0.955 m/s for the sinusoidal flow above.

Table 6.2: Steady flow & Grid 1 : Time steps tested

All methods								
Δt (s)	48	100	150	200	240	300	360	400
c_r	0.12	0.25	0.375	0.5	0.6	0.75	0.9	1
$\mathcal{N}$	225	108	72	54	45	36	30	27
Backward characteristic methods only								
Δt (s)	450	540	600	720	900	1080	1200	
c_r	1.125	1.35	1.5	1.8	2.25	2.7	3	
$\mathcal{N}$	24	20	18	15	12	10	9	

Table 6.3: Steady flow & Grid 2 : Time steps tested

	All methods					Backward characteristic methods only		
Δt (s)	48	100	150	200	240	300	360	400
c_r(moy)	0.13	0.27	0.40	0.54	0.64	0.80	0.97	1.07
$\mathcal{N}$	225	108	72	54	45	36	30	27

Table 6.4: Unsteady sinusoidal flow & Grid 1 : Time steps tested

	All methods				Characteristics	
Δt (s)	54	72	108	133.334	144	200
c_r(moy)	0.26	0.34	0.52	0.64	0.69	0.96
c_r(max)	0.40	0.54	0.81	1.00	1.08	1.5
$\mathcal{N}$	200	150	100	81	75	54

5. **Boundary conditions**

Inflow boundary conditions will be indicated later on, case by case. Outflow boundary conditions should be such that $\lim_{x\to\infty} = 0$.

As explained in previous chapters, advection and dispersion are solved in different steps. Only inflow boundary conditions are needed for the solution of advection. Dispersion is then simply neglected at boundaries.

With the Holly-Preissmann algorithm, it is necessary to prescribe the derivative also at each inflow boundary. For the sake of simplicity, let us assume such a boundary is located at node 1. The equation solved during the advection stage may be developed there at time level $n+1$ as follows :

$$\frac{C_1^{n+1} - C_1^n}{\Delta t} + U_1^{n+1} \left(\frac{\partial C}{\partial x}\right)_1^{n+1} = 0$$

Consequently, knowing the prescribed concentration values C_1^n, it is possible to calculate an estimate of the derivative at the boundary.

6.3 Pure advection of concentration-hills

6.3.1 Steady flow and Uniform grid

For all trials, the initial center of mass x_0 has been set to 2000 m. The gauss-hill standard deviation is 264 m, the triangle and square wave half-lengths respectively 800 and 1200 m.

The dimensionless length L of each source is defined as its length divided by the grid spacing : the gauss-hill, triangle-hill and square-wave dimensionless lengths are respectively 8, 8 and 12. The sources appear to be chiefly defined with the help of $L+1$ grid nodes.

At inflow boundary, we prescribe $C(0,t) = 0 \quad \forall t > 0$.

At time t, the **exact solution** is given by :

$$C(x,t) = C^0\left(x - \int_0^t u(\tau)\, d\tau \right)$$

Gauss-hill The dependency of error measures with respect to the Courant number is illustrated in figures 6.1 to 6.4. We indicate hereafter the legend used in the plots and the nickname we shall use from now on to refer to each scheme :

- Results for backward characteristic methods are plotted in solid lines (Holly-Preismann scheme - "HOLLY"), short dashed lines (Rasch-Williamson algorithm - "RASCH"), long dashed lines (Minimax/Li method - "LI"), small dotted lines (Dan N'Guyen method - "DAN").
- As regards the flux-form methods, they have all been plotted with un-even dashed lines but bear different marks : no marks for the 4^{th}-order area-preserving scheme ("BOTT4"), black triangles for the algorithm corresponding to the 3^{rd}-order Taylor Series expansion ("BOTT3"), squares for the Takacs/Quickest scheme ("QUICKEST").

We can compare all seven schemes only in the range $c_r \leq 1$.

1. **Damping** (cf fig. 6.1)
   As regards the preservation of the peak value, one scheme appears far superior to the others : RASCH never attenuates the peak value more than 5% ; yet, it introduces eventually some minor (less than 2 %) overshoot. Then come HOLLY and BOTT4 algorithms, which yield fairly close results for $c_r \leq 1$, and BOTT3. QUICKEST and the two last backward characteristics methods (DAN and LI) have nearly indistinguishable results. For BOTT3, QUICKEST, LI and DAN, damping at small Courant numbers is quite severe.

Except for RASCH, the damping decreases steadily as the Courant number increases from 0.12 to 1, a value for which all schemes yield an exact solution. For $c_r \geq 1$, the backward characteristics methods results appear to depend both on the decimal part of the Courant number and on the total number of iterations (the less iterations we perform, the less interpolation errors we add up).

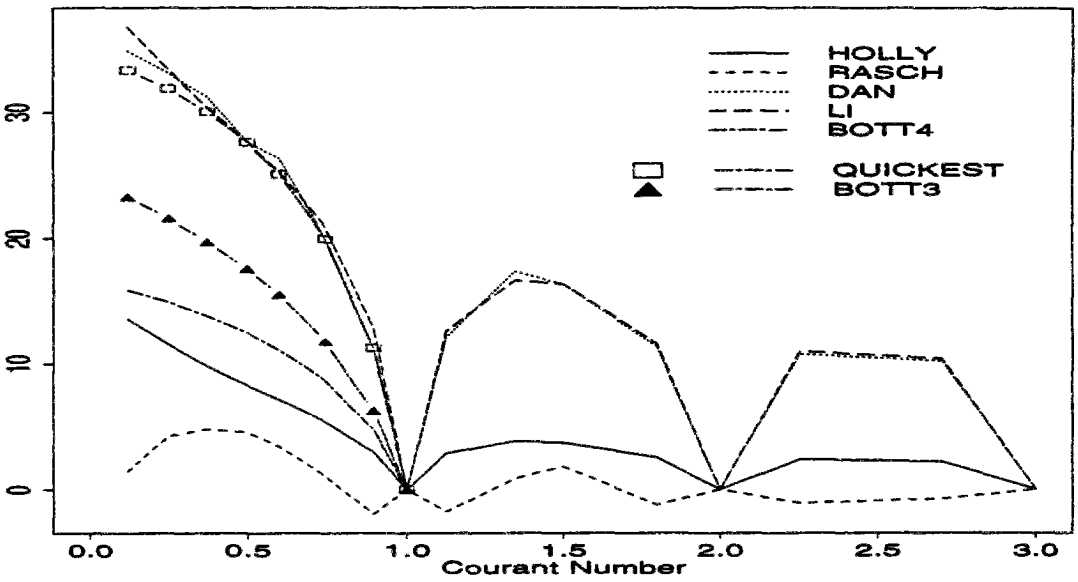


Figure 6.1: Relative error on peak value (%) - Gauss-hill L=8

2. **Spurious undershoots** (cf fig. 6.2)
   BOTT3 and BOTT4 are positive definite schemes and consequently their results cannot be plagued by negative concentrations.

   RASCH especially and HOLLY give excellent results, with undershoots less than 0.2 and 2 % of the peak value respectively. The performance of QUICKEST and LI schemes is also satisfying (undershoots always inferior to 6.5 and 5.3 % respectively). In contrast, DAN behaviour is quite catastrophic.

3. **L2 error norm** (cf fig. 6.3)
   The L2 error norm is quite small for RASCH, HOLLY, BOTT3 and BOTT4. The error yielded by the RASCH scheme is nearly constant, whereas for other methods it decreases from $c_r = 0.12$ to 1.

   QUICKEST and LI perform similarly. The worst scheme according to this criterion is DAN.

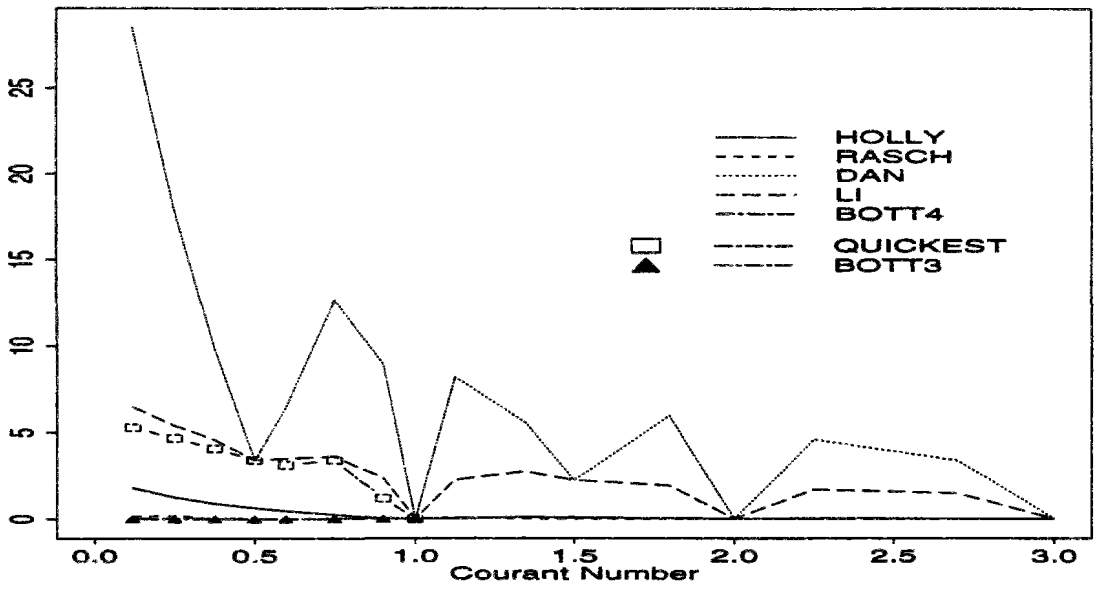


Figure 6.2: Max. Neg. concentration (in % of peak value) - Gauss-hill L=8

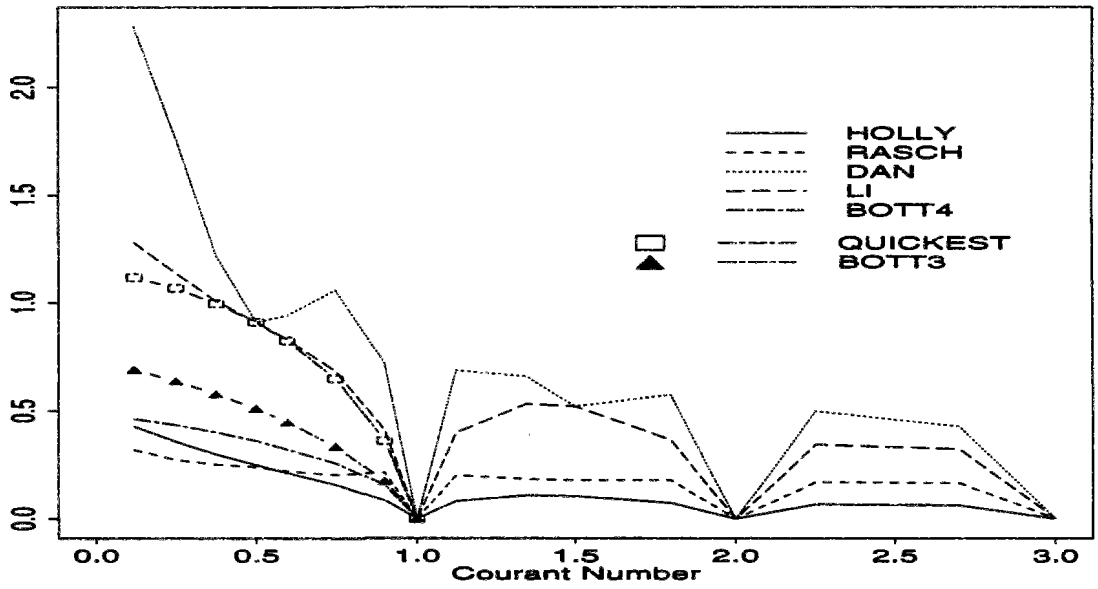


Figure 6.3: L2 error norm (in % of exact mass) - Gauss-hill L=8

4. **Mass preservation**

   By construction, QUICKEST, BOTT3 and BOTT4 are conservative, while the backward characteristics methods are not. Yet, HOLLY, DAN and LI yield negligible relative mass error (approximately 10^{-10}). For the last two schemes, this indicates that the existing negative concentrations are somehow compensated, most probably because of spurious spreading of the advected shapes.

   The only scheme which has some (moderate) trouble with mass preservation is the RASCH one. The average relative mass error is 0.4 % in the range $0.5 \leq c_r \leq 3$ (with no error when the decimal part of the Courant number is either 0 or 0.5). It is significantly bigger for the smallest Courant numbers (3.6, 2.1 and 1.1 % for $c_r = 0.12, 0.25, 0.375$ respectively).

5. **Phase shift**

   All schemes locate properly the peak except DAN : for $c_r < 0.5$, it places the peak upstream of its exact location (one cell error for $cr = 0.25, 0.375$, two cells error for $cr = 0.12$); for $0.5 < c_r < 1$, the peak is one cell downstream; there is no error for $c_r \geq 1.35$.

   The integral measure μ_x of the phase shift (cf figure 6.4) is somehow deceptive. Indeed, for some schemes (e.g. DAN), the spurious negative concentrations may somehow cancel out the integral error : while the computed pollutograms are obviously shifted with respect to the exact solution, the position of their center of mass is exact. This is illustrated in figure 6.5, which displays the result of all backwards characteristics methods for $c_r = 0.25$ (the exact solution corresponds to cross marks) : while the related μ_x is significantly bigger than for LI and DAN pollutograms, the RASCH pollutogram is much less shifted than these two distributions !

   On the whole, the error is quite small (less than 0.1 % for all schemes, except the RASCH one for which it remains generally smaller than 0.4 %). The schemes all have a lagging phase error when the decimal part of the Courant number is less than 0.5, a leading phase error otherwise. (It confirms for some of them the results of the Fourier analysis).

6. **Numerical spreading** (cf fig. 6.6)

   This error measure is also somewhat inadequate : for schemes plagued by the occurrence of non-negligible negative concentrations (DAN, LI, QUICKEST), μ_{xx} has the exact value 1, while the corresponding pollutograms are obviously too spread. This is illustrated once again by the results of the backwards characteristics methods, for $c_r = 0.5$, figure 6.7 (on which DAN and LI results cannot be distinguished).

   We may also note on this figure that, while the computed numerical spreading of the RASCH pollutogram is rather important (1.19 for this case), this pollutogram and the exact solution are in fact quite close. It appears that, like the integral phase shift, the

variances ratio is not the kind of error measure one can blindly trust to decide upon the accuracy of a scheme nor upon its relative ranking with respect to other methods.

In summary, the best scheme for this test case appears to be RASCH, in spite of the fact that it tends to deform slightly the transported shape, making it more triangular. Not far behind we find the HOLLY and BOTT4 algorithms which yield fairly close results; yet, the former has no stability restriction. The BOTT3 scheme also performs correctly, except for the smallest Courant numbers ($c_r \leq 0.25$), for which it exhibits too much damping. This too strong attenuation of the peak values is even more marked for the LI and QUICKEST schemes, whose results are generally not distinguishable. Finally, use of the DAN scheme is not advisable, as it raises unacceptable undershoots. The results of the four best schemes are furthermore illustrated on figure 6.8.

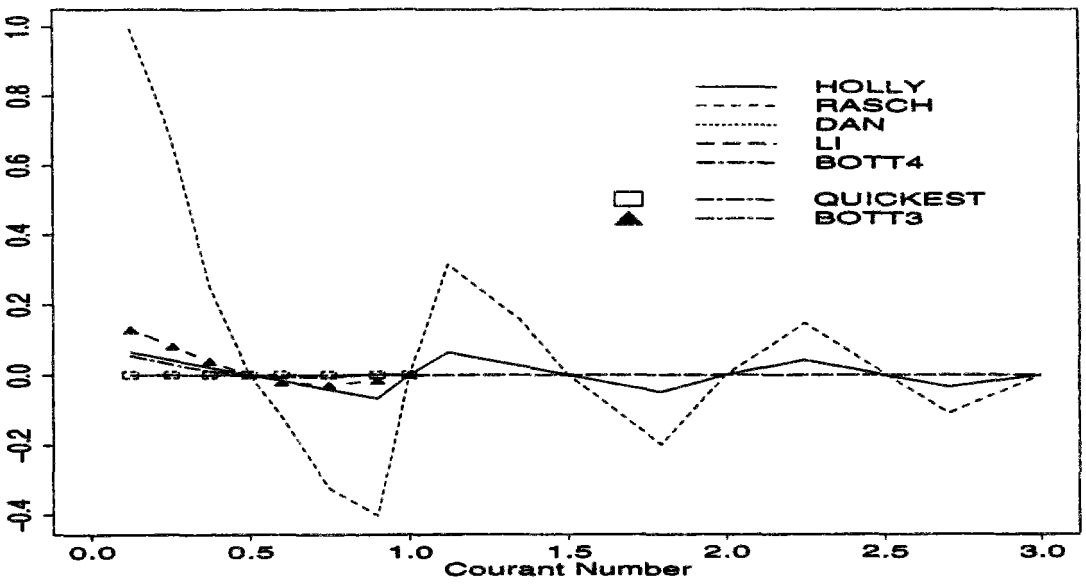


Figure 6.4: Global phase shift (in % travel distance) - Gauss-hill L=8

Triangle-hill All schemes perform slightly better for this somewhat wider shape. They rank as for the previous test.

This time, the RASCH algorithm is undoubtedly the best method. We have noticed it tended to make the gauss-hill more triangular; here it seems to be perfectly suited to the transport of an initially triangular shape. Its mass error is no longer significant and its global phase shift is also reduced (nb : for all schemes the absolute value of the global phase shift μ_x remains less than 0.06 % of the total travel distance). It introduces no more spurious overshoot (cf fig. 6.9) but slightly more undershoot than with the gauss-hill (cf fig. 6.10). The RASCH scheme (instead of HOLLY for the gauss-hill) is also the best one according to the L2 error norm criterion (cf fig. 6.11).

Once again, RASCH, HOLLY and BOTT4 and, to a lesser extent, BOTT3 prove to behave satisfactorily. The only scheme which has serious problems is DAN (which is, for instance, the only one misplacing the peak).

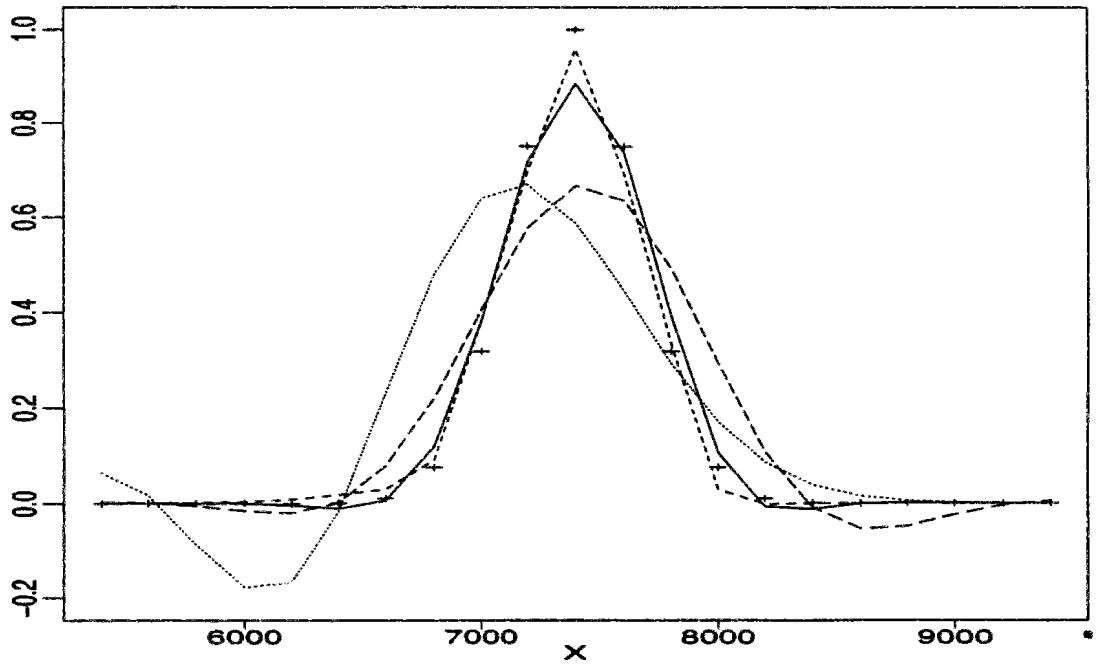


Figure 6.5: Gauss-hill L=8. Backward characteristic methods $c_r = 0.25$

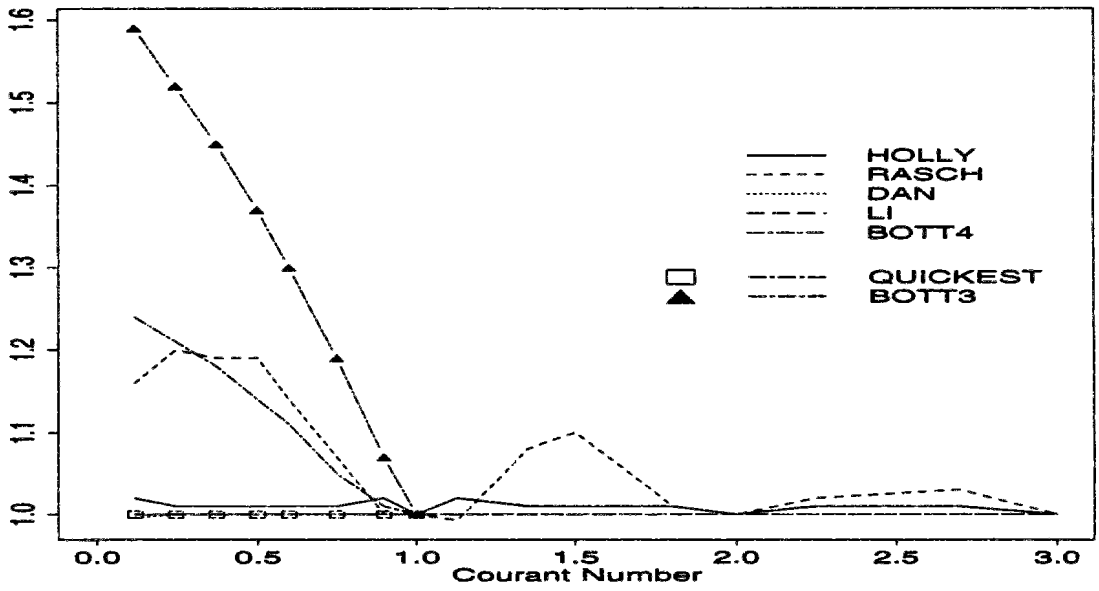


Figure 6.6: Ratio of numerical to exact variance - Gauss-hill L=8

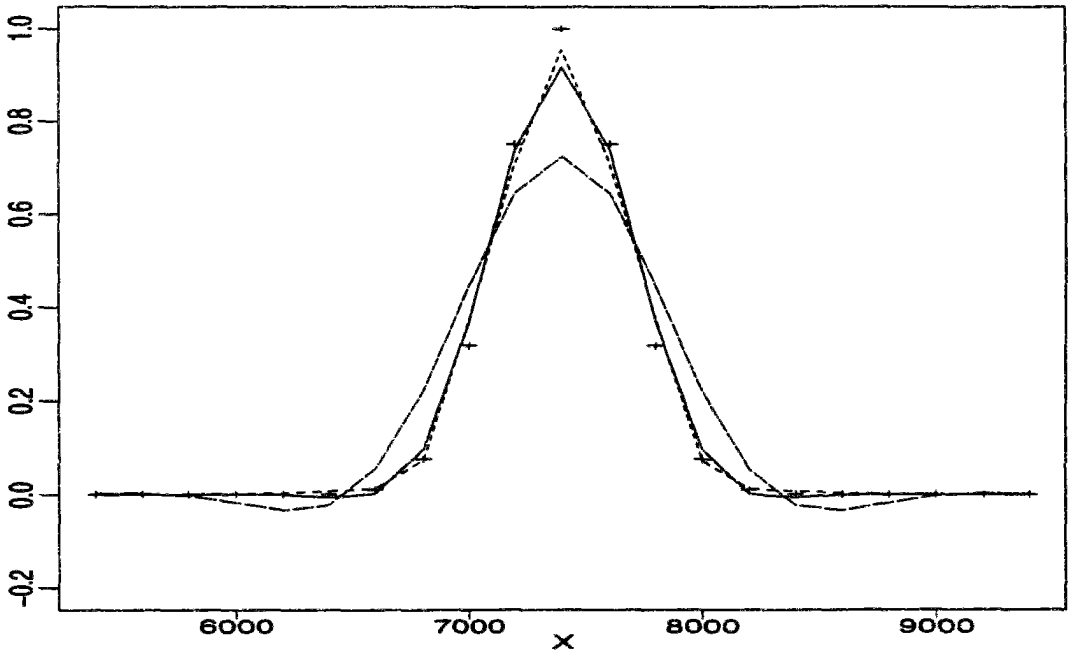


Figure 6.7: Gauss-hill L=8. Backward characteristic methods $c_r = 0.5$

Square wave We shall discuss the different errors in a slightly different order than above. It is worth noting that for this test case the discrete error on peak location ξ is obviously not relevant.

1. This time, the transported shape no longer undergoes damping but rather systematic **overshoots** at its upper edges and **undershoots** at its base (except for the positive definite schemes BOTT3 and BOTT4). The behaviour of both errors in the range $c_r \leq 1$ is more erratic than previously (see figures 6.13 and 6.14) : they no longer decrease steadily as the Courant number approaches 1. Once again, the best performing scheme according to these two criteria is RASCH, followed by HOLLY. As regards the magnitude of overshoots, LI, QUICKEST, BOTT3, BOTT4 raise similar results, the latter being no longer the best of these four schemes. The worst performance is DAN's.

2. The mass error is negligible, except for DAN scheme (1.2 % and 0.2 % of relative mass loss respectively for $c_r = 0.12$ and 0.25).

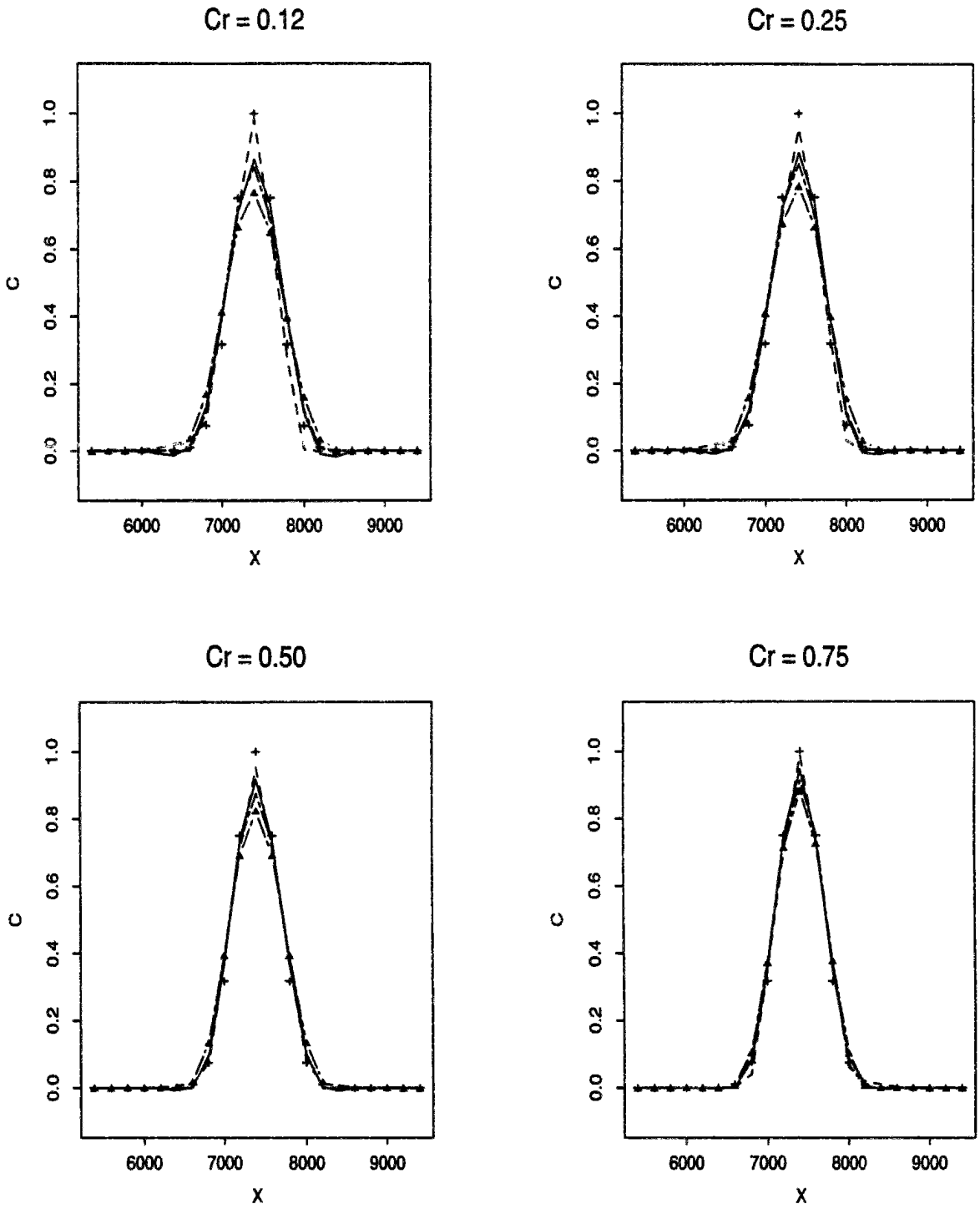


Figure 6.8: Gauss-hill L=8. Overview of the performance of the best schemes

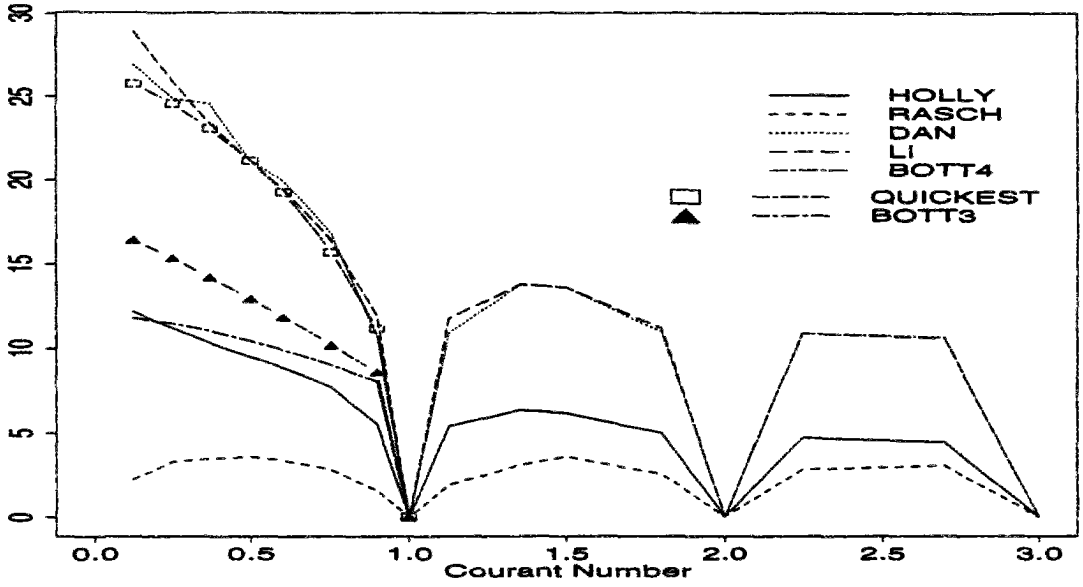


Figure 6.9: Relative error on peak value (%) - Triangle-hill

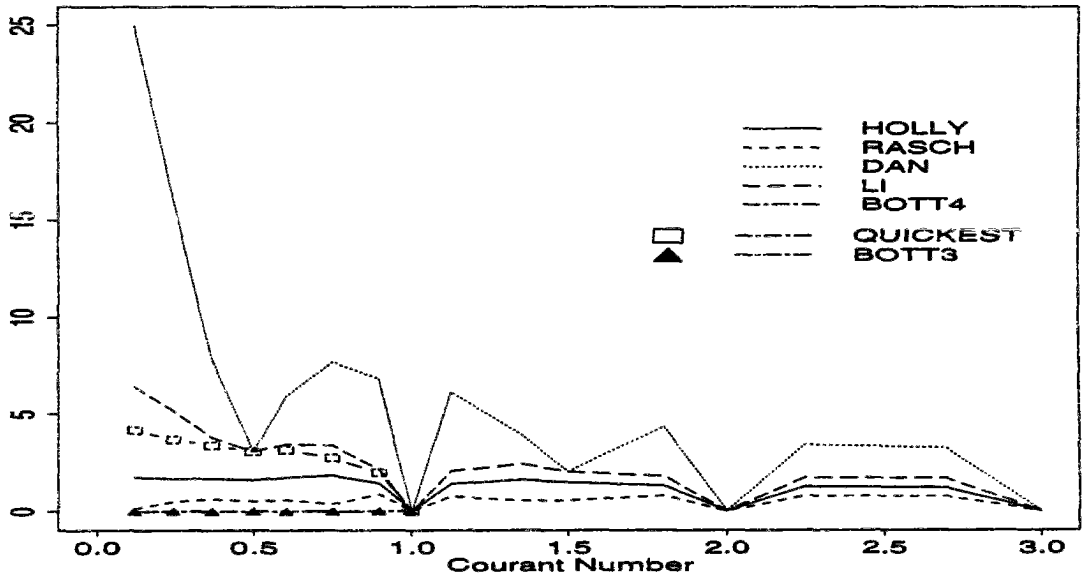


Figure 6.10: Max. Neg. concentration (in % of peak value) - Triangle-hill

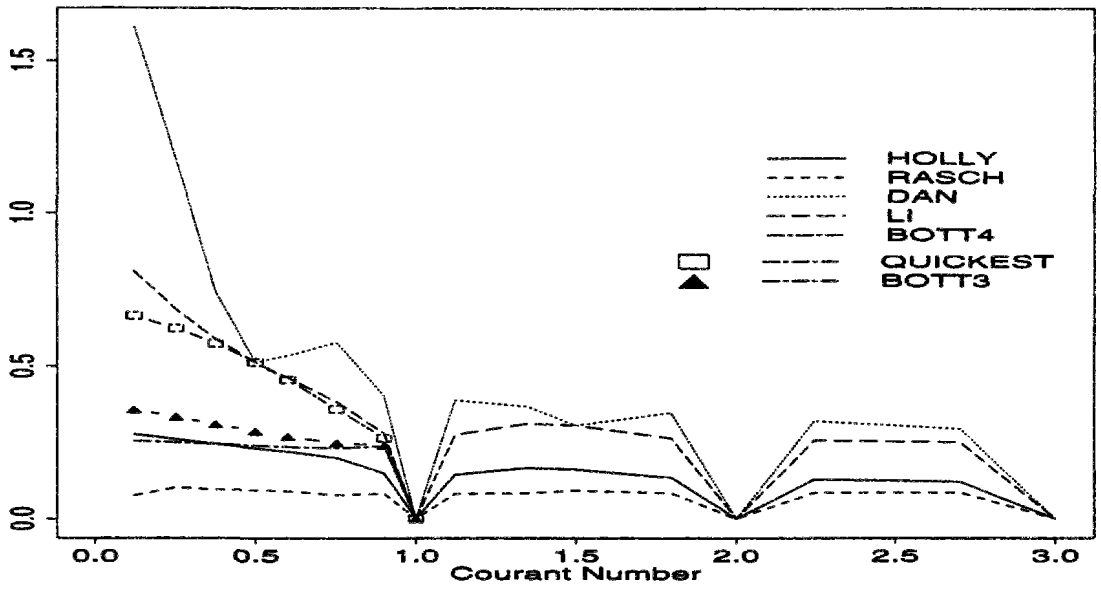


Figure 6.11: L2 error norm (in % of exact mass) - Triangle-hill

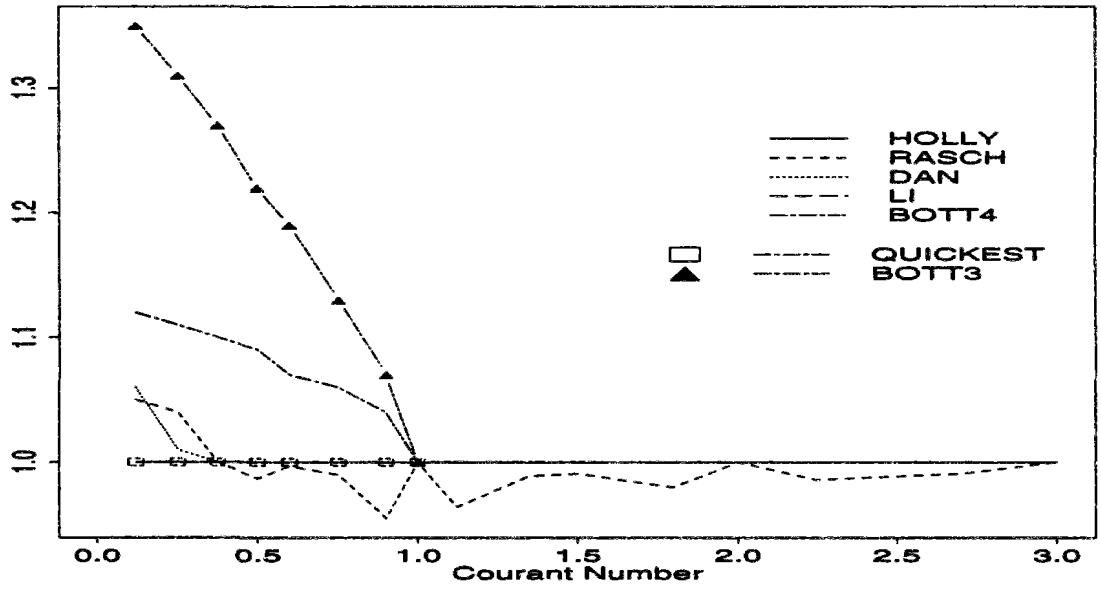


Figure 6.12: Ratio of numerical to exact variance - Triangle-hill

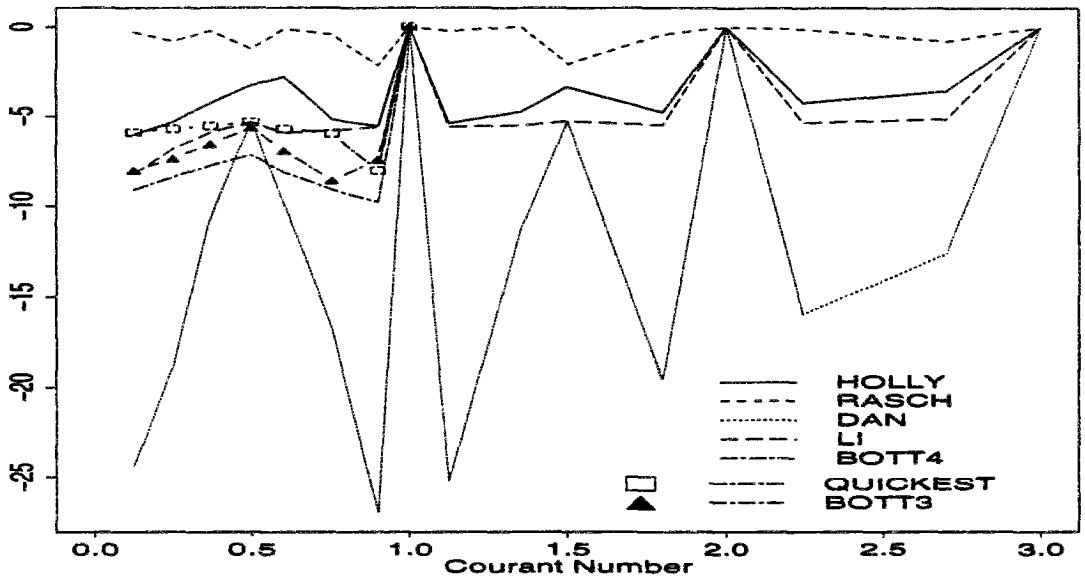


Figure 6.13: Maximum overshoot (in % of true max.) - Square Wave

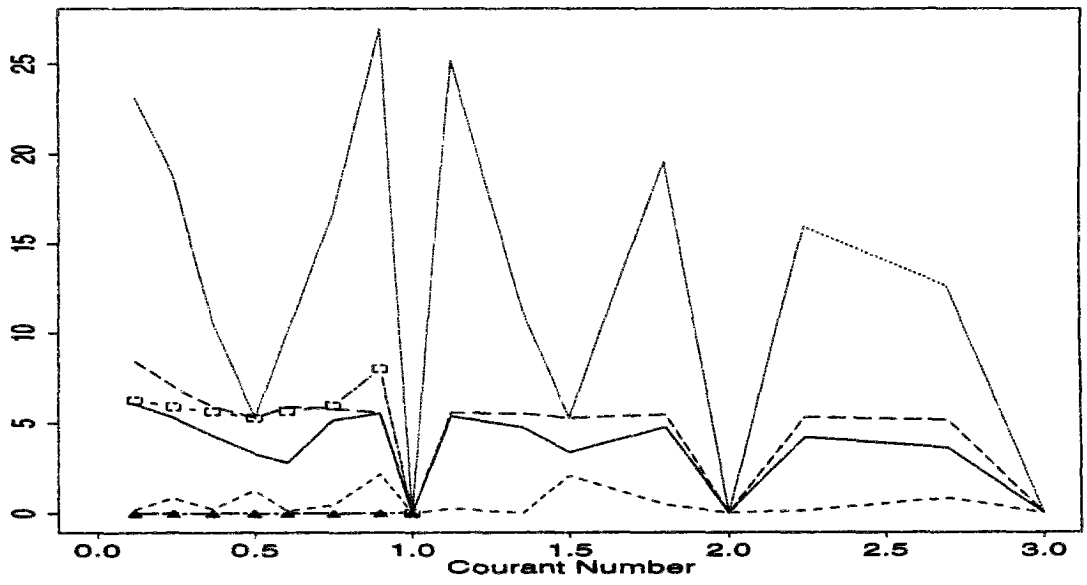


Figure 6.14: Maximum undershoot (in % of true max.) - Square Wave

3. The μ_{xx} variance ratio has not been plotted because the numerical spreading is quite moderate.

   μ_{xx} is 1 for QUICKEST and LI schemes. It is similarly 1 for DAN as soon as $c_r \geq 0.5$: for the smallest Courant numbers, DAN forecasts too narrow a shape ($\mu_{xx} = 0.857, 0.976, 0.996$ for $c_r = 0.12, 0.25, 0.375$ respectively).

   μ_{xx} for HOLLY is always equal to 1.01 except for integer Courant numbers (for which the scheme, as all backward characteristics methods, yields an exact solution).

   For BOTT3, BOTT4, RASCH, μ_{xx} stays in the range [1, 1.07].

4. With regard to the L2 error norm (fig. 6.15), the best scheme appears to be HOLLY. However, the most remarkable feature is the relatively important error induced by the RASCH scheme which previously ranked first or second according to this criterion. We shall investigate the cause of this result in the next point.

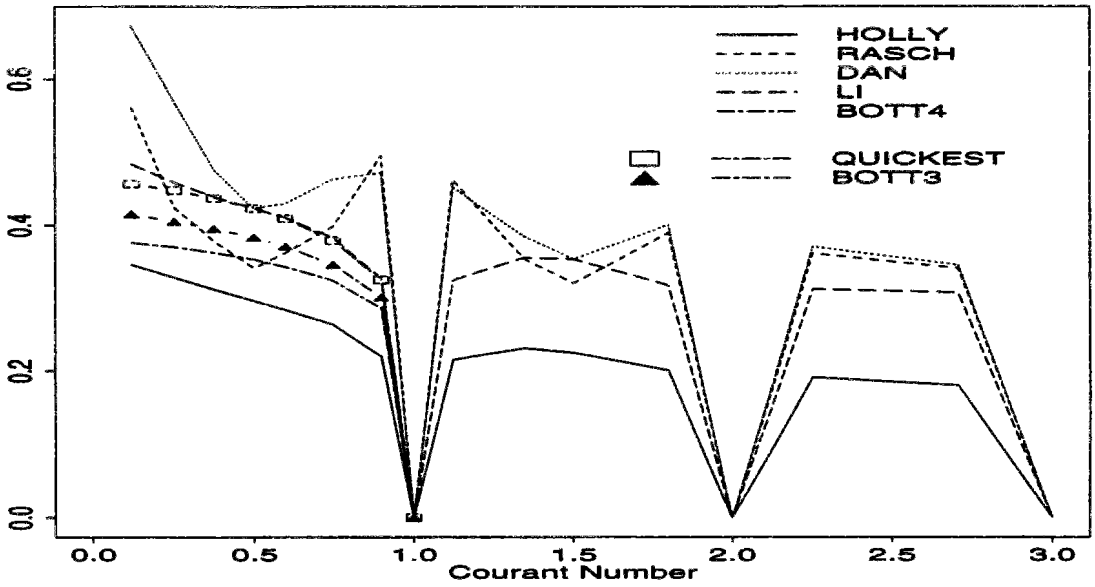


Figure 6.15: L2 error norm (in % of exact mass) - Square Wave

5. The global phase shifts display significantly higher levels than for previous tests (cf fig. 6.16) especially as regards the RASCH algorithm. This phase shift may explain the corresponding importance of the L2 error norm.

   A look at figures 6.17 to 6.20 helps to better understand what is going on. In fact it appears that, on the whole, the RASCH algorithm is the one which best preserves the transported shape. It does so at the expense of a stronger phase error, which nevertheless keeps within reasonable limit. The performance of HOLLY is also fair. BOTT3 and

BOTT4 give equivalent results : in contrast to other tests, the latter is not superior to the former.

In summary, the three tests performed which each concentration hill lead to consistent conclusions. They highlight the good performance of the two backwards characteristic methods relying on the use of the Hermite cubic interpolant : the Holly-Preissmann scheme and the Rasch-Williamson method based on the use of the Akima derivative estimate. Two of the flux form methods yield also fair results : these are the positive definite schemes based respectively on a 3^{rd} order Taylor Series expansion and on a 4^{th}-order area preserving polynomial. The three other schemes are much less satisfactory, especially that of Dan'Nguyen.

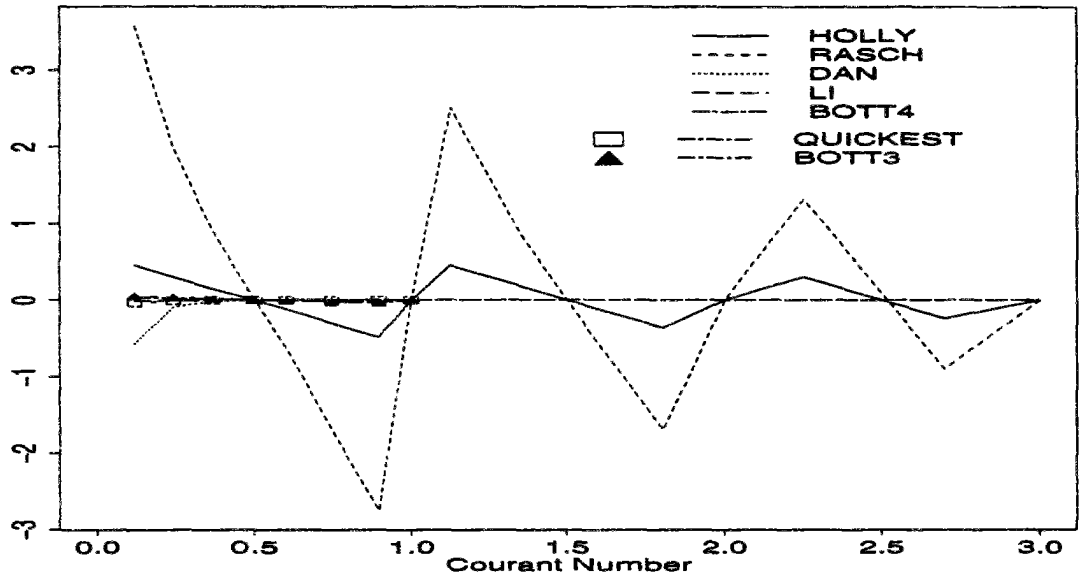


Figure 6.16: Global phase shift (in % travel distance) - Square Wave

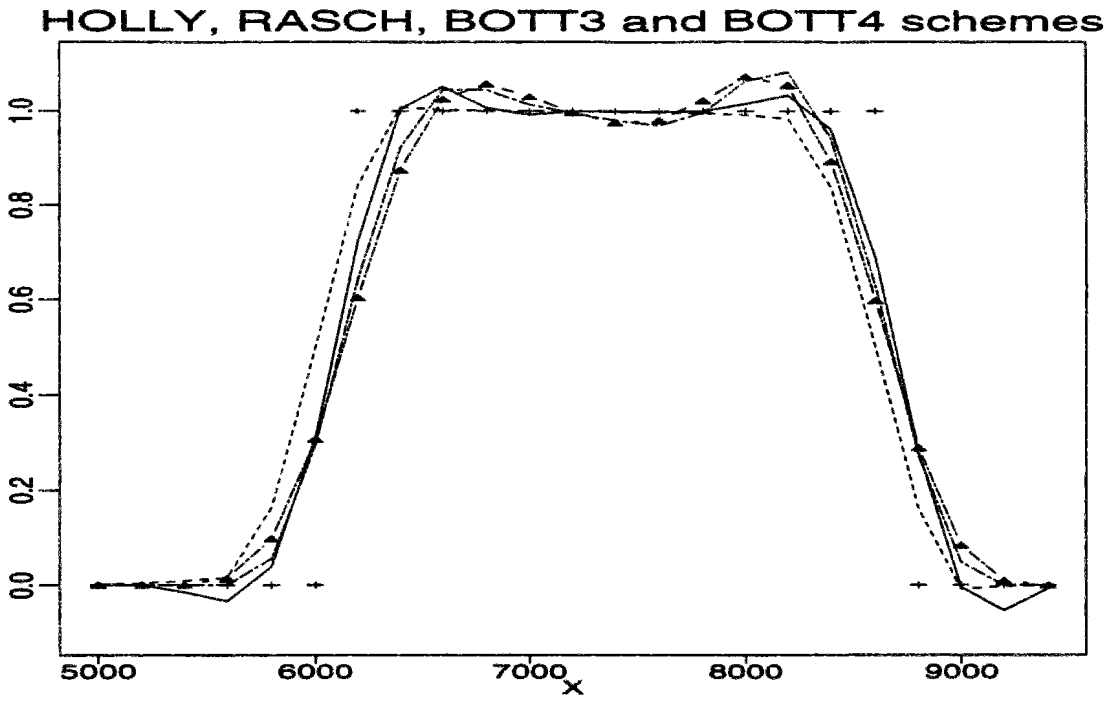


Figure 6.17: Square Wave L=12. $c_r = 0.25$

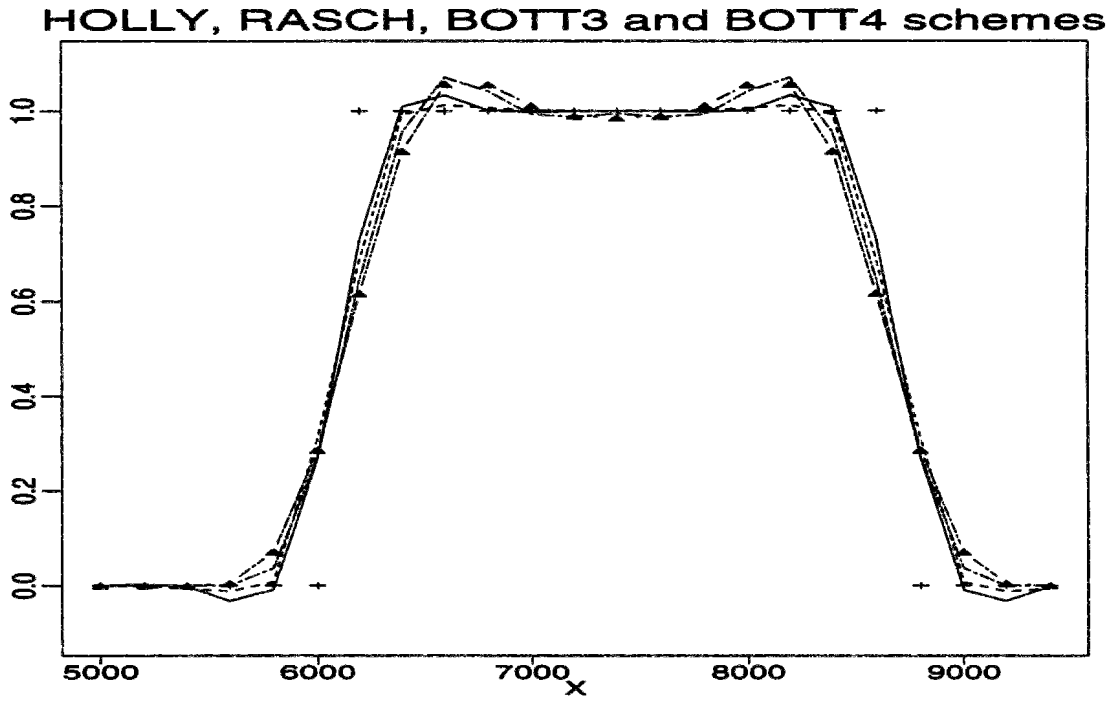


Figure 6.18: Square Wave L=12. $c_r = 0.50$

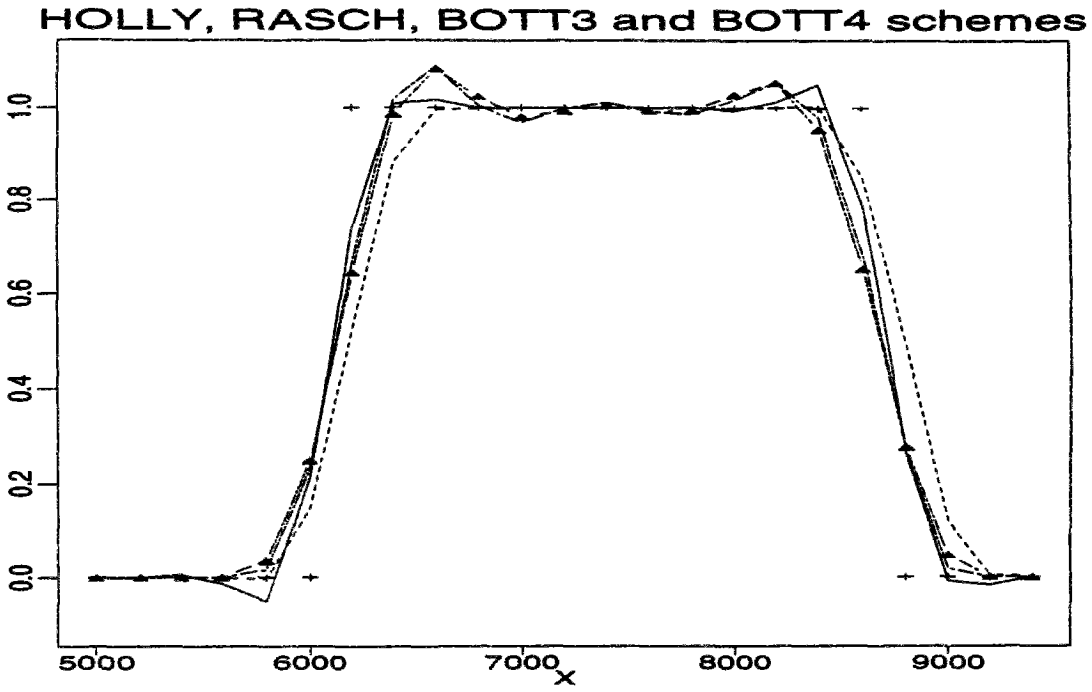


Figure 6.19: Square Wave L=12. $c_r = 0.75$

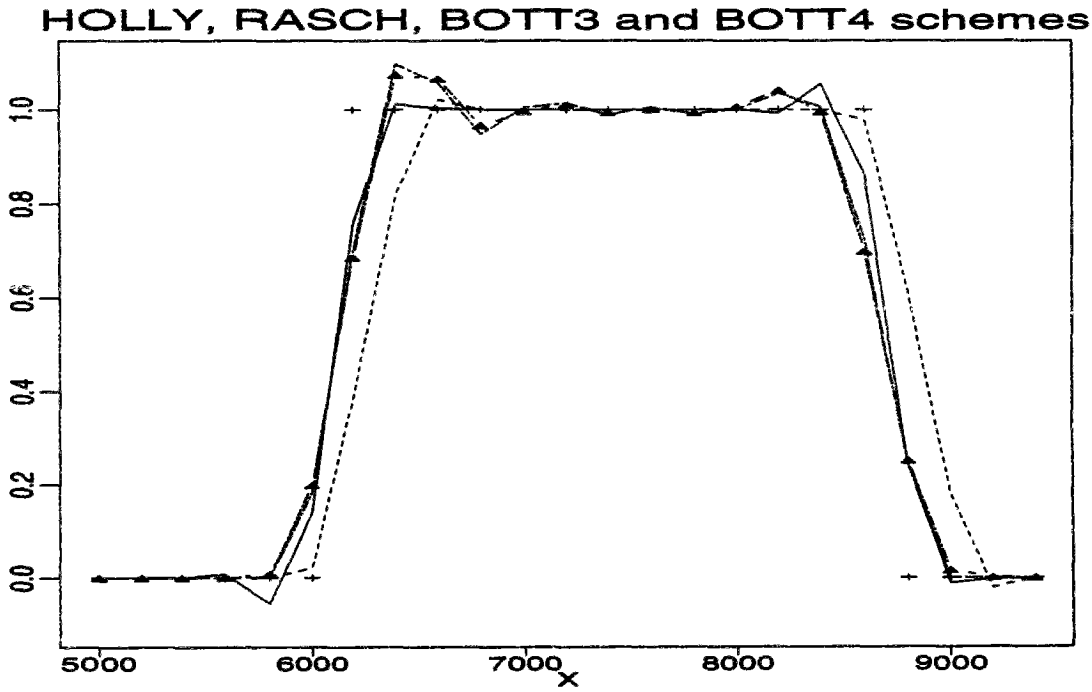


Figure 6.20: Square Wave L=12. $c_r = 0.90$

6.3.2 Influence of non-uniform grid spacing

In order to evaluate the robustness of schemes when applied to irregular grids, we have been studying the transport of the gauss-hill on grid 2. The comparison of damping, L2 error norm and numerical spreading obtained with grids 1 and 2 is illustrated in figures 6.21 to 6.23 for the four schemes RASCH, HOLLY, BOTT3 and BOTT4. On grid 2 the local Courant number is variable : we have now been plotting the error measures as a function of the number of iterations performed to obtain the numerical solution at time $t = 10800$ s. We have also indicated on these graphics the errors related to the transport on uniform grid 1 of a wider gauss-hill, whose dimensionless length is 12 ($\sigma_0 = 400$ m). The legend is as follows :

- solid lines with cross marks refer to the small gauss-hill/uniform grid test;
- dashed lines with black squares refer to the small gauss-hill/variable grid test;
- dotted lines with triangles refer to the bigger gauss-hill/uniform grid test.

1. **Damping, L2 error norm & numerical spreading**
   The utilisation of variable grid spacing seems to affect but slightly the schemes' performance. However, the RASCH algorithm appears to be more sensitive to the grid irregularity than the others.

   The use of a wider (thus, better described) initial source results in a clear improvement of the error measures, except as regards the peak value forecast by the RASCH algorithm. Once again, this scheme tends to deform slightly the transported gauss-hill shape into a more triangular distribution. The resulting overshoot is rather stable ($\simeq 5$ %), except for $c_r = 0.9$ where it suddenly reaches 8 %.

2. **Negative concentrations**
   The value of these spurious undershoots is only indicated for the Holly-Preissmann algorithm (figure 6.24). Indeed, BOTT3 and BOTT4 are positive definite schemes and the negative concentrations produced by RASCH are once again negligible (less than 0.3 % and 0.22 % of peak value with the irregular grid and the wider source respectively).

   The use of an irregular grid slightly deteriorates HOLLY's performance but, on the whole, the results remain fairly good. With the wider gauss source, negative concentrations fall below 0.15 % of the peak value (they are approximately divided by a factor ten with respect to the small gauss-hill/regular grid test).

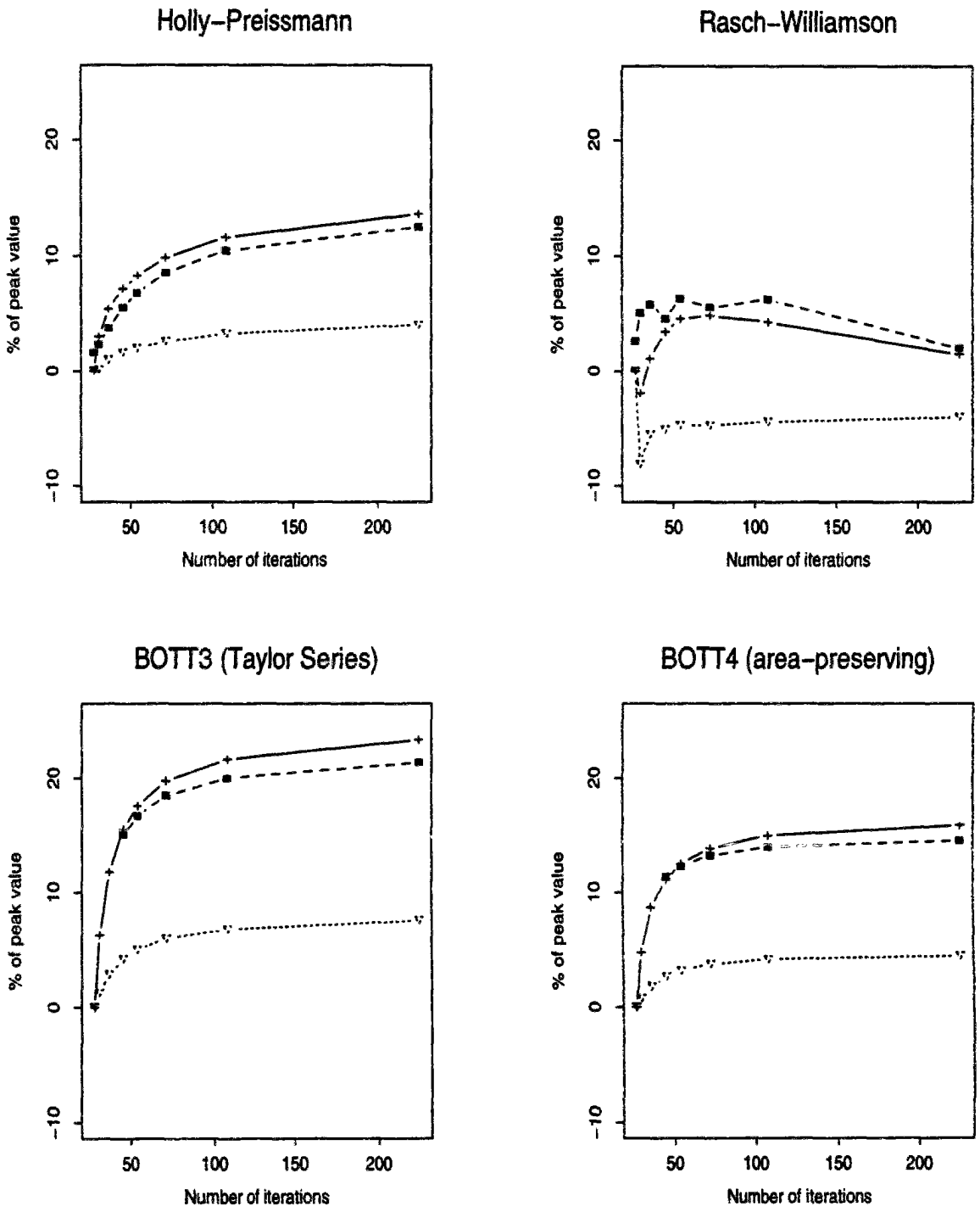


Figure 6.21: Influence of grid spacing & source length on damping

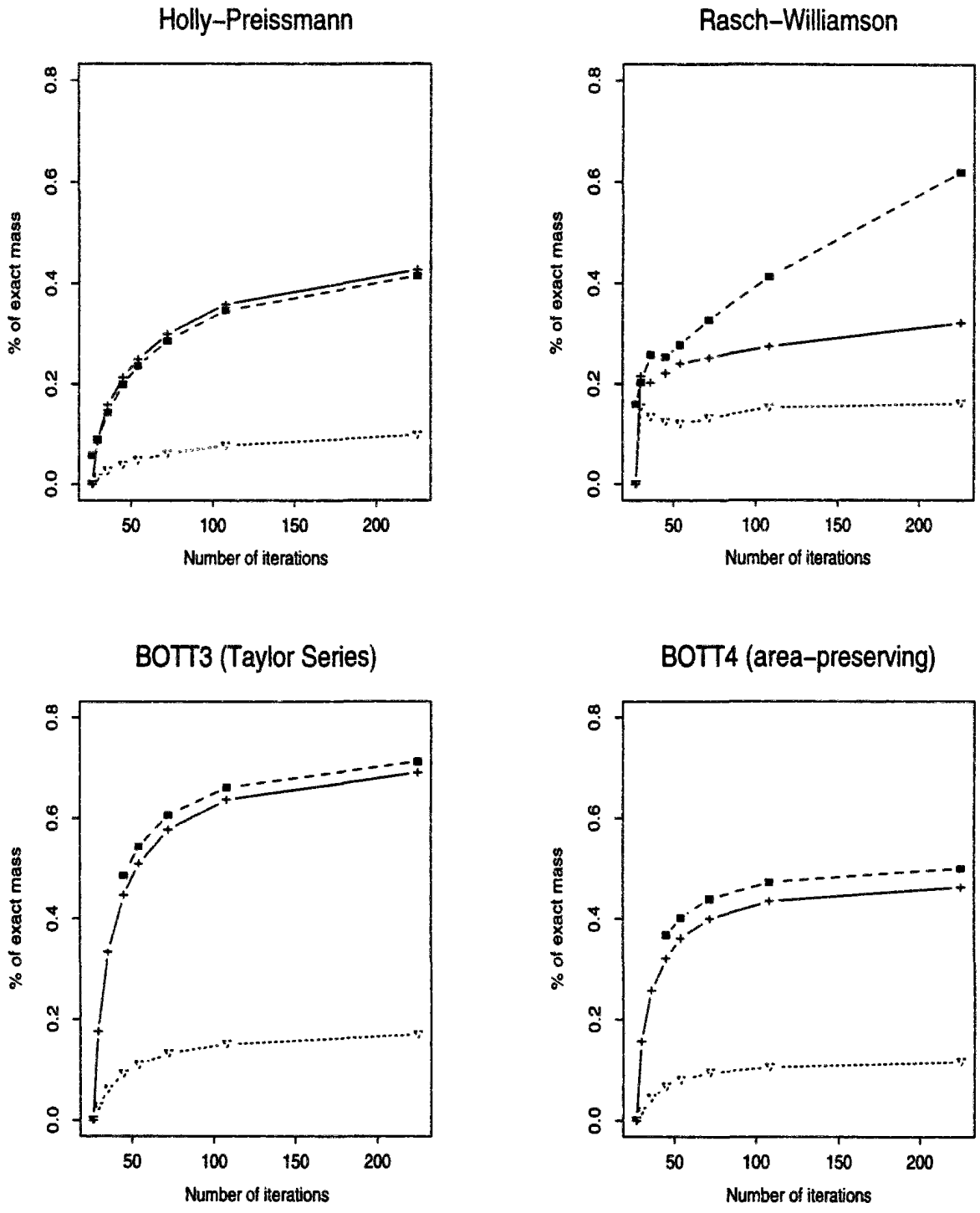


Figure 6.22: Influence of grid spacing & source length on L2 error norm

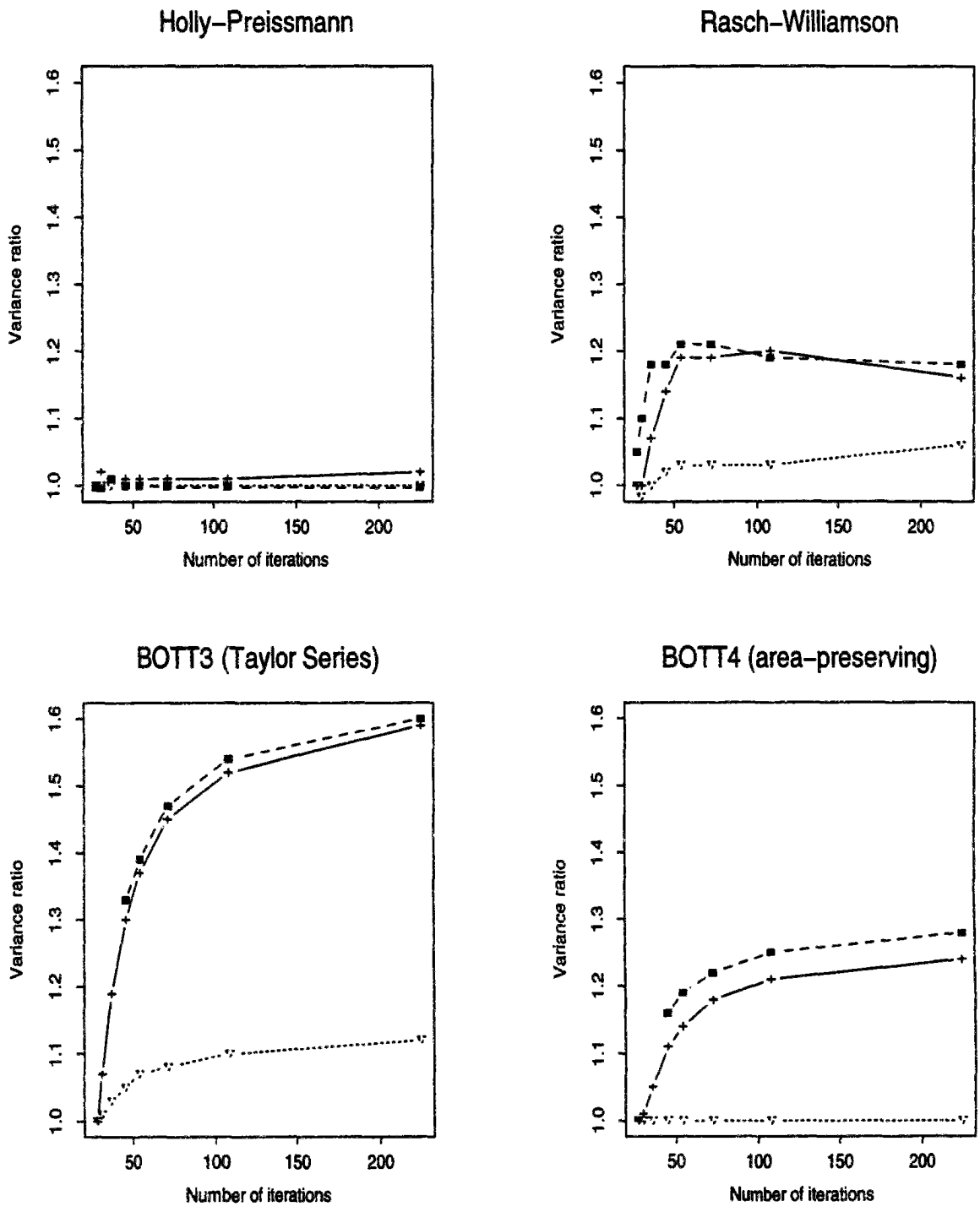


Figure 6.23: Influence of grid spacing & source length on numerical spreading

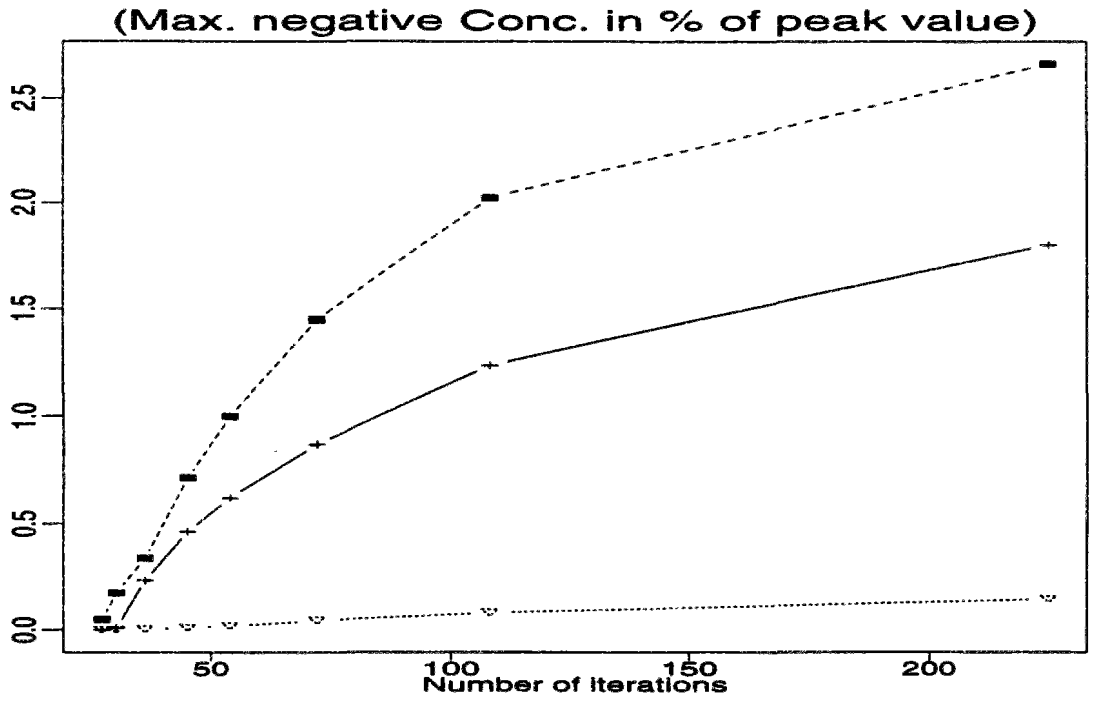


Figure 6.24: Holly algorithm undershoots : variable grid & different source lengths

3. **Mass preservation**

   This measure needs comment only for the RASCH scheme (cf fig. 6.25), the other schemes being perfectly satisfactory.

   For large numbers of iterations, the irregular spacing clearly worsens the algorithm's performance but then, results on grids 1 and 2 become quite similar : the mass loss is less than 1 %, which is quite acceptable.

   The mass preservation of the wider source is nearly perfect, except for $c_r = 0.9$, where the slight mass overestimation can be ascribed to the bigger peak overshoot observed for this particular Courant number value.

4. **Phase shift**

   Once again, there is never any error relative to the peak location.

   Global phase shifts on irregular grid keep negligible with HOLLY, BOTT3 and BOTT4 algorithms : their absolute value remains respectively inferior to 0.008, 0.12 and 0.048 % of the total travel distance. For HOLLY, there is even a neat improvement when using the variable grid instead of the regular one ! With the wider gauss source, the global phase shift vanishes (less than 0.003 %, 0.025 %, 0.002 % respectively).

   RASCH's performance is degraded by grid irregularity (cf figure 6.26) and is clearly better with a bigger initial source.

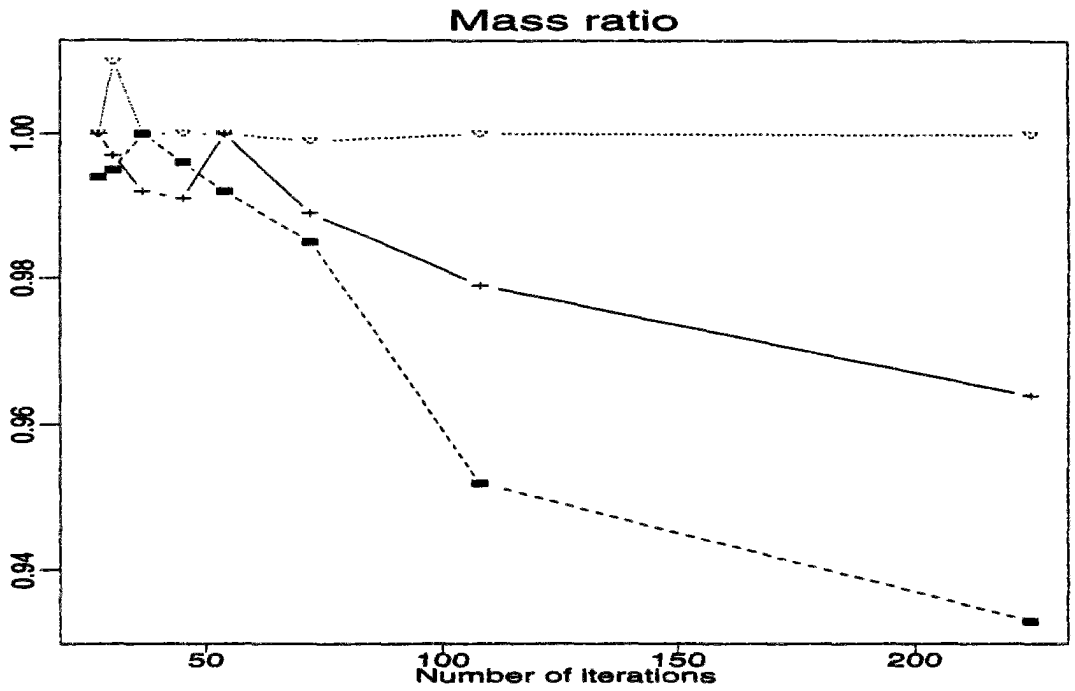


Figure 6.25: Rasch algorithm mass preservation : variable grid & different source lengths

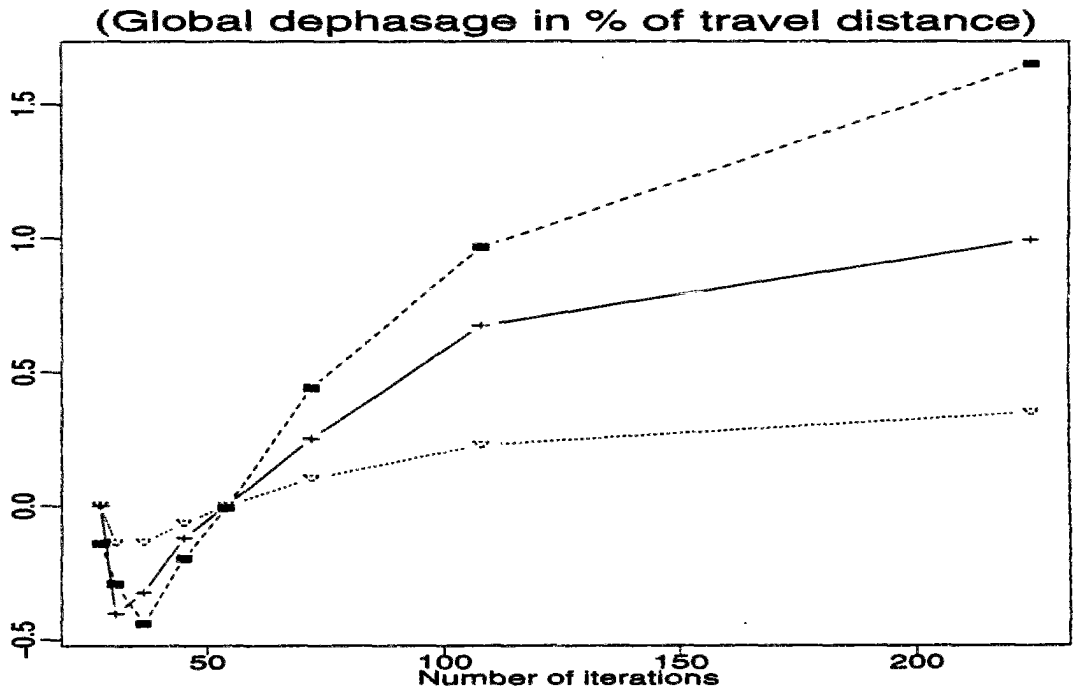


Figure 6.26: Rasch algorithm phase shift : variable grid & different source lengths

5. **Overview of computed pollutograms**
   The relatively large mass and phase shift errors displayed by the RASCH algorithm at low Courant numbers (i.e. large iteration numbers) in the irregular grid test drove us to check the computed pollutograms. These are displayed in figure 6.27 (same notations as in figure 6.8). It appears that RASCH produces too triangular pollutograms : while the upstream part of the distribution is correctly described, RASCH fails to reproduce the trapezoidal summit of the distribution and forecasts a too sharp decrease of the concentrations downstream of the peak.

In conclusion, the schemes appear to be relatively robust with respect to grid variability. Yet, the tendency of the RASCH algorithm to triangularize gauss shapes is more bothersome than for previous tests, particularly as it is accompanied by a higher mass loss.

As could be expected, the use of a wider, smoother, better defined, source results in an overall improvement of the schemes performance. Notably, the solutions computed by the BOTT3, BOTT4 and HOLLY algorithms respectively are much closer than before.

6.3.3 Unsteady sinusoidal flow

The test has been performed on the uniform grid 1. The initial source is the small gauss-hill ($L = 8$). The prescribed inflow concentrations are null. The problem has not been solved for the DAN scheme.

As regards backwards characteristic methods, the backtracking algorithm has been designed in order to deal indifferently with steady and unsteady flows (cf app. E.1.1). The fact that the flow is now time varying should probably affect more the flux-form methods, as the determination of the proper bounds of the flux integrals may become less precise.

In this test, we could have taken advantage of our perfect knowledge of the flow field to limit as much as possible the inaccuracies relative to characteristic foot or integral bounds computation. We have decided to do otherwise : when computing the solution at time level $n + 1$, bounds and characteristics are computed only with the help of the flow field *at the previous time level n*. Indeed, should the algorithms be used not only to forecast passive scalar transport but also to solve the advection stage in flow models, we would not enjoy the knowledge of flow field at both times.

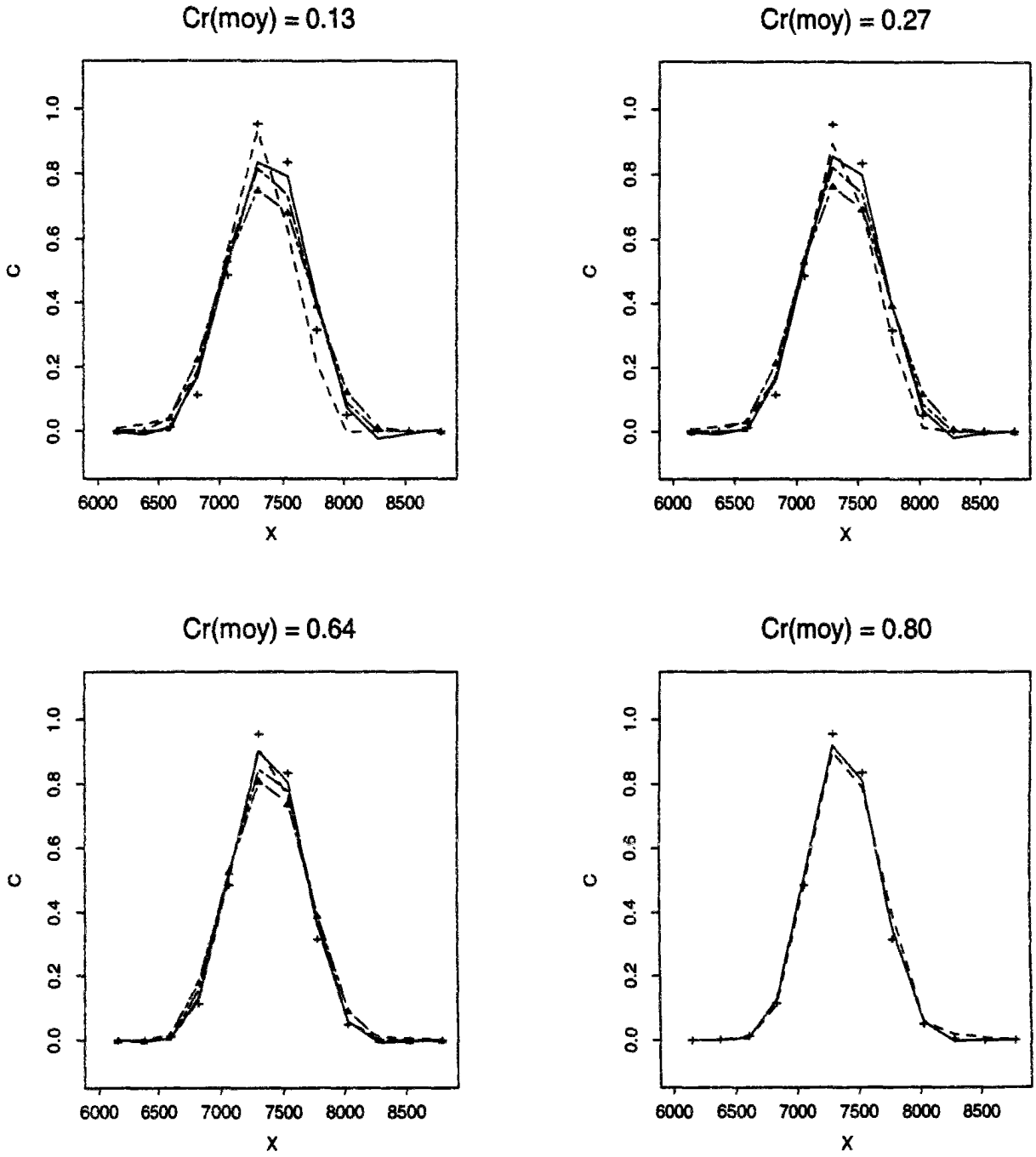


Figure 6.27: Transfer of Gauss-hill (L=8) on irregular grid : computed pollutograms

1. **Damping** (cf figure 6.28)
   The schemes rank as for the gauss-hill/steady flow test : the scheme which induces less peak attenuation is clearly RASCH, followed by HOLLY, BOTT4 and BOTT3. LI and QUICKEST/TAKACS results are nearly identical.

   According to **L2 error norm** (cf figure 6.29), the best scheme is again RASCH.

2. **Spurious undershoots** are negligible for RASCH (less than 0.0005 %). They remain bounded between 3 and 4 % for LI and QUICKEST schemes, between 0.5 and 2 % for HOLLY, and increase with the number of iterations performed.

3. **Mass preservation**
   Mass loss affects only the RASCH algorithm, as usual. The variation of the mass error according to the number of iterations is somewhat erratic. The mass loss ranges between 0.2 and 2.8 %, with an average of 1.5 %.

4. **Phase shift**
   There is no error relative to the peak position.

   We could have expected the global phase shift to be more important than usual, because of the imprecision introduced in dealing with variable velocities. However, as the velocity changes sign, errors may cancel each other. Surveying the temporal evolution of error measures (Simon, 1990a) (section 5.2) in sinusoidal flow reveals that the phase shift displays indeed local minima at full periods and local maxima at half periods when no cancellation has occurred.

   In our case (t = 10800s, full period), the resulting phase shift is negligible, even for the RASCH scheme : it is always less than 0.2 % of the travel distance. It does not exhibit any clear dependency with respect to the iteration number.

5. For BOTT3, BOTT4, HOLLY and RASCH algorithms, **numerical spreading** (cf figure 6.30) has the same order of magnitude as in steady flow conditions.

   For the LI and QUICKEST algorithms, the numerical variance is too low, probably because of the weight of negative concentrations located at the base of the computed pollutogram.

In summary, this test does not bring to light anything really new about the schemes. It confirms the relative grading of the algorithms, which all seem able to cope with this kind of smoothly varying unsteady flows.

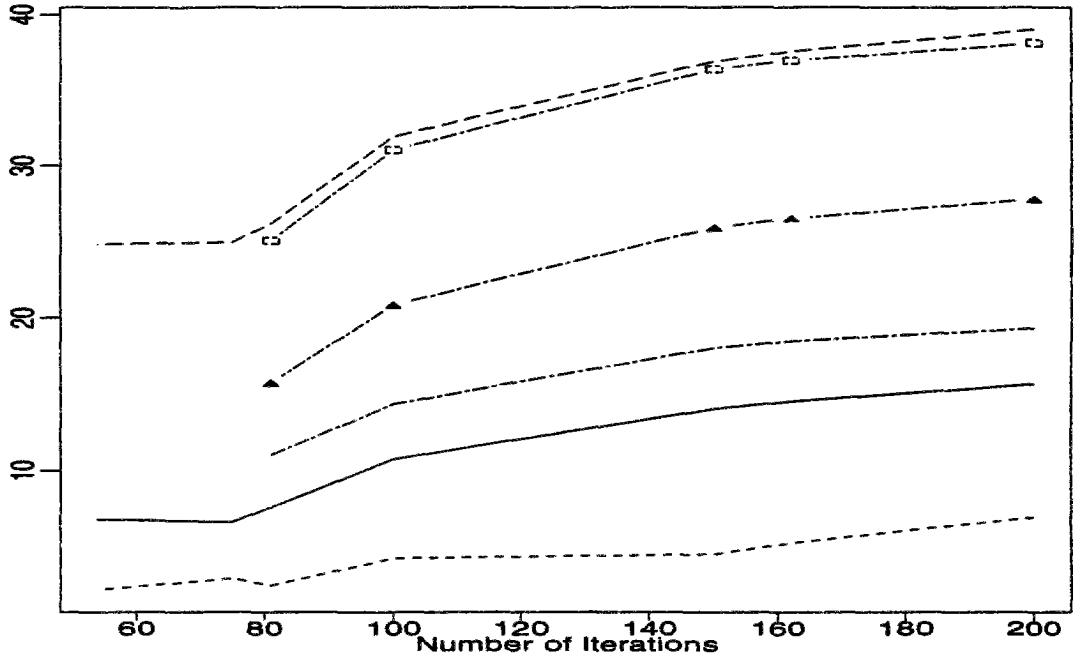


Figure 6.28: Relative error on peak value (%) - Sinusoidal flow

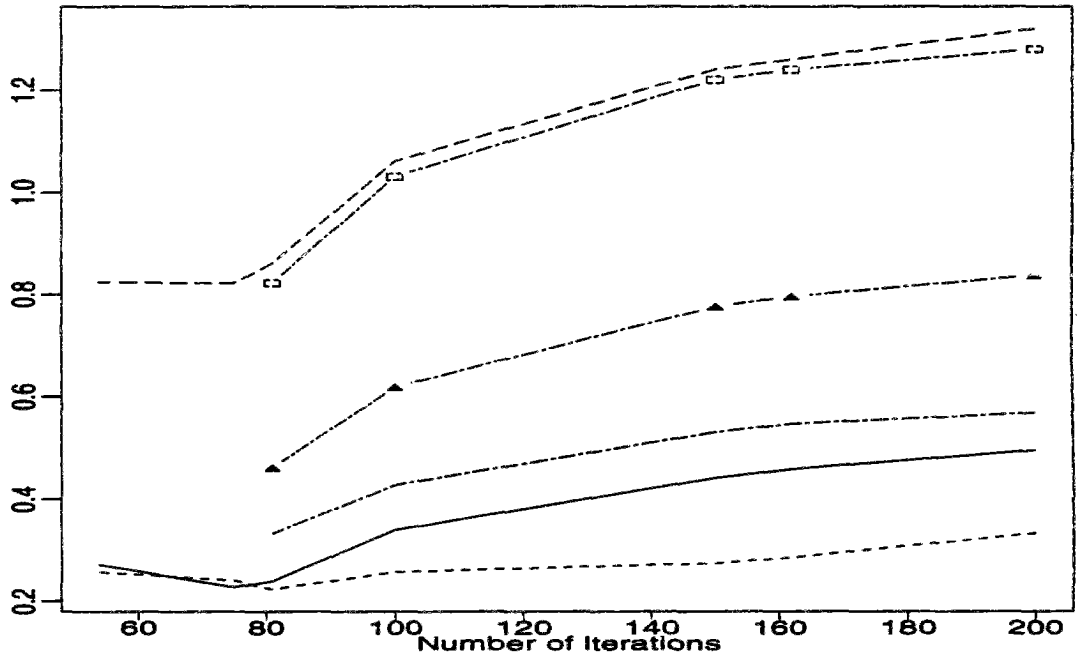


Figure 6.29: L2 error norm (in % of exact mass) - Sinusoidal flow

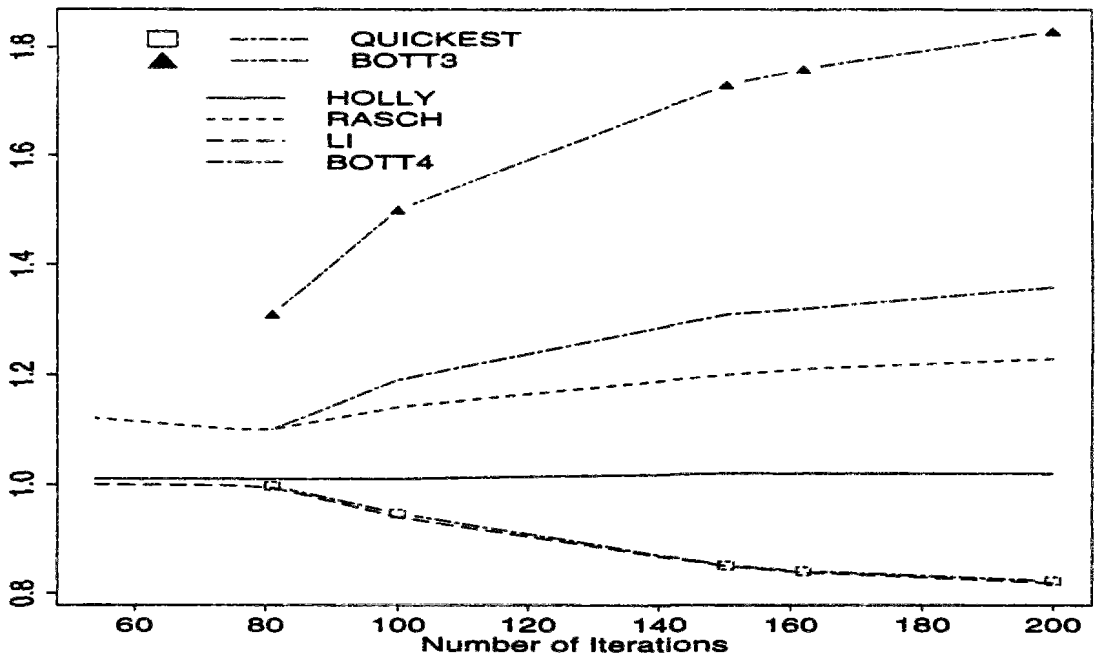


Figure 6.30: Ratio of numerical vs. exact variance - Sinusoidal flow

6.4 Advection and diffusion of gaussian profiles

Until now, we have only been dealing with pure advection. In order to obtain a more precise idea about the range of applicability of the schemes, the transport of gauss-hill sources ($L = 8$ and 12) has been studied under different dispersion conditions.

The exact solution at time t reads :

$$\begin{aligned} C(x,t) &= \frac{\sigma_0}{\sigma}.exp\left\{-\frac{(x-\bar{x})^2}{2.\sigma^2}\right\} \\ \sigma^2 &= \sigma_0^2 + 2.\Gamma.t \\ \bar{x} &= x_0 + U.t \end{aligned}$$

We once again survey the results at $t = 10800$s. Grid 1 is used, and the simulations performed with the time steps indicated in table 6.2. Table 6.5 indicates the range of diffusivities tested. This leads us to a total of 88 simulations (per each given initial L) for every flux-form method, of 165 simulations for every backward characteristics method.

Table 6.5 also mentions the corresponding Peclet number $Pe = U\Delta x/\Gamma$ which constitutes a measure of the relative strengths of advection and dispersion phenomena. Lf_n and $C\max_n$

denote respectively the final distribution dimensionless length and peak concentration for an initial source whose dimensionless length is n and peak concentration 1.

Table 6.5: Tested diffusivities and resulting Peclet numbers

Γ	100	50	20	10	5	2	1	0.5	0.2	0.1	0.
Pe	1	2	5	10	20	50	100	200	500	1000	∞
Lf_8	45	32	21	16	12.5	10	9	8.5	8	8	8
$C\max_8$	0.177	0.246	0.373	0.494	0.626	0.786	0.874	0.930	0.970	0.985	1
Lf_{12}	46	33	23	18	15.5	13.5	13	12.5	12	12	12
$C\max_{12}$	0.263	0.359	0.520	0.652	0.773	0.887	0.940	0.968	0.987	0.993	1

Some error measures are already moderate for the pure advection case (i.e. the global phase shift, the numerical spreading, the L2 error norm). They tend to improve when some physical diffusion is introduced. We shall not dwell on these but rather focus on two error measures which have hitherto proved most helpful when it comes to judge the schemes accuracy and relative rank : the peak damping and the magnitude of spurious undershoots. Their dependency with respect to the relative strengths of advection and diffusion is illustrated in figures 6.31 to 6.33. In these, we have been drawing isolines which indicate the level of errors across the $Pe - c_r$ plane. Undershoots (figure 6.33) do not occur with positive definite schemes, nor have they been plotted for RASCH algorithm, for which they are frankly negligible.

These isolines allow us to visualize quickly a **"domain of applicability"** for every scheme. In the case of the gauss-hills test case, we suggest that this domain of applicability could be defined according to the following criteria :

- *Condition 1* : Negative concentrations smaller than 5 % of the peak concentration value
- *Condition 2* : Relative damping of peak concentration less than 15 %

We may seem rather tolerant as regards the peak value preservation. Yet, in practical applications, an uncertainty of about ten percent concerning the pollutant input definition is quite common. It seems acceptable to allow the schemes to have some errors, provided they do not exceed those stemming from a possible lack of precision in the definition of the studied transport problem. On the other hand, restrictions concerning the magnitude of negative concentrations are more stringent : these purely numerical products may quickly induce worse troubles when the transport model is coupled with biogeochemical models.

We shall first compare the schemes on their common stability domain, namely $c_r \leq 1$:

- We may consider there is no restriction to the application of HOLLY, BOTT4 and RASCH algorithms, in spite of the fact that the two first schemes have a slightly too impor-

tant damping for nearly pure advective conditions and small Courant numbers and that the third generates some noticeable peak overshoot for intermediate dispersion conditions ($Pe \simeq 20$).

HOLLY, BOTT4 and RASCH all exhibit a (slight) peak overshoot in diffusion dominated problems ($Pe \leq 10$) : this is entirely due to a non-optimal resolution of the diffusion step. This phenomenon also plagues the other schemes but is less marked, as these algorithms introduce more numerical damping, which counteracts the overshoot induced by the diffusion-step resolution.

- BOTT3 is then the scheme with the wider range of applicability. It satisfies conditions 1 and 2 for all Pe inferior to 60. For $c_r = 0.4$, the range of applicability extends to $Pe \leq 100$. Then, as soon as $c_r \geq 0.6$, it includes all possible combinations of advection and dispersion processes.

- Next come the QUICKEST and LI schemes : they have similar domains of applicability, the former being nevertheless slightly better. The limit of applicability is $Pe \simeq 15 - 20$ for $c_r < 0.5$. It is approximately $Pe = 30$ for $c_r = 0.5$ precisely. There is no longer restriction to the schemes use for $c_r \geq 0.8$.

- The more limited scheme is DAN. Only the cases which correspond to $Pe \leq 10$ can be correctly treated whatever the chosen time step and corresponding Courant number. While the DAN algorithm yields slightly less damping than the QUICKEST and LI schemes, it generates more frequently negative concentrations : hence the more severe restrictions. For $c_r \geq 0.4$ the situation is somewhat better and problems with Pe up to 15 - 20 should be satisfactorily dealt with.

- A look at the isolines shape (figures 6.32 and 6.33) indicates that, as regards flux-form methods, our interest is to use systematically the maximum allowable time step. There is no restriction on the time step for backward characteristics methods, for which an optimal choice appears to be less straightforward. Figure 6.32 shows indeed that damping for instance depends both on the total number of iterations performed and on the decimal part of the Courant number : while c_r is less than 1, errors decrease as the time step increases; then, errors exhibit local maxima when the decimal part of c_r is close to 0.5 , local minima when c_r is close to be an integer.

- With a bigger initial source ($L = 12$), the schemes applicability is improved, the error isolines having the same shape as for the smaller gauss-hill.

  The maximum negative concentrations induced by HOLLY, QUICKEST and LI schemes remain always less than 0.15, 2.8 and 5 % respectively. For the DAN scheme, there are still unacceptable undershoots in the area $c_r \leq 0.4$ & $Pe \geq 10$.

The peak damping generated by BOTT3, BOTT4 and HOLLY algorithms keeps respectively inferior to 7.5, 5 and 4 %. LI and QUICKEST damping are less than 15 % except for the sub-areas defined respectively by ($c_r \leq 0.3$ & $Pe \geq 100$) and ($c_r \leq 0.2$ & $Pe \geq 300$). RASCH induces systematic peak overshoots for $Pe > 10$. In this domain, they seem rather independent of the Courant number value and stay close to 5 %, except for the particular Courant number value $c_r = 0.9$, for which they may raise up to 8 %. We have no explanation to supply for this peculiar behaviour at $c_r = 0.9$!

It should be clear that the domains of applicability we have been defining above are related only to the specific test-cases of the gauss-hills transport. They depend not only on the initial distribution, the Pe and c_r values but also on the time and place where we chose to register the errors. What would we observe if we decided, for instance, to run the schemes and compute their errors for a total computational time twice as long as the actual one? Probably two opposite effects :

- *better results for diffusion-dominated problems*, as the diffusion would have had more time to smooth the concentration profile and thus, "ease" the numerical resolution;
- *worse results for convection-dominated problems*, as, in particular, the spurious oscillations (negative concentrations) tend to grow as time goes by.

Nevertheless, this error visualization completes our assessment of the schemes performance. It confirms that HOLLY, RASCH and BOTT4 methods are the most robust ones. BOTT3 stands not far behind. The QUICKEST and LI schemes, especially the former, may still prove useful for intermediary situations when the advection is not too dominant. Use of the DAN scheme, on the other hand, does not seem advisable, except in undoubtedly diffusion-dominated problems.

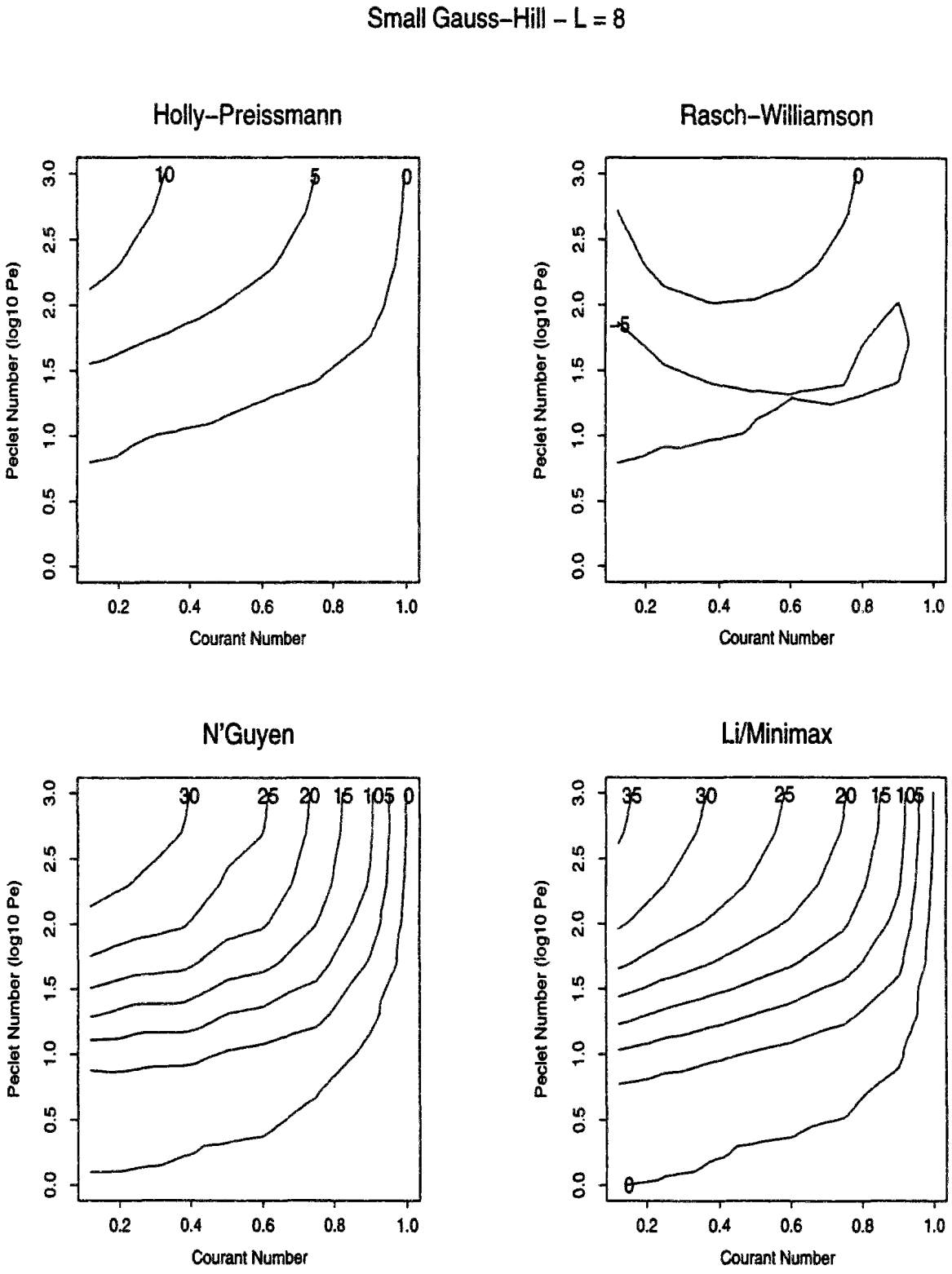


Figure 6.31: Peak damping (%) isolines v.s. Peclet and Courant numbers - Characteristics methods

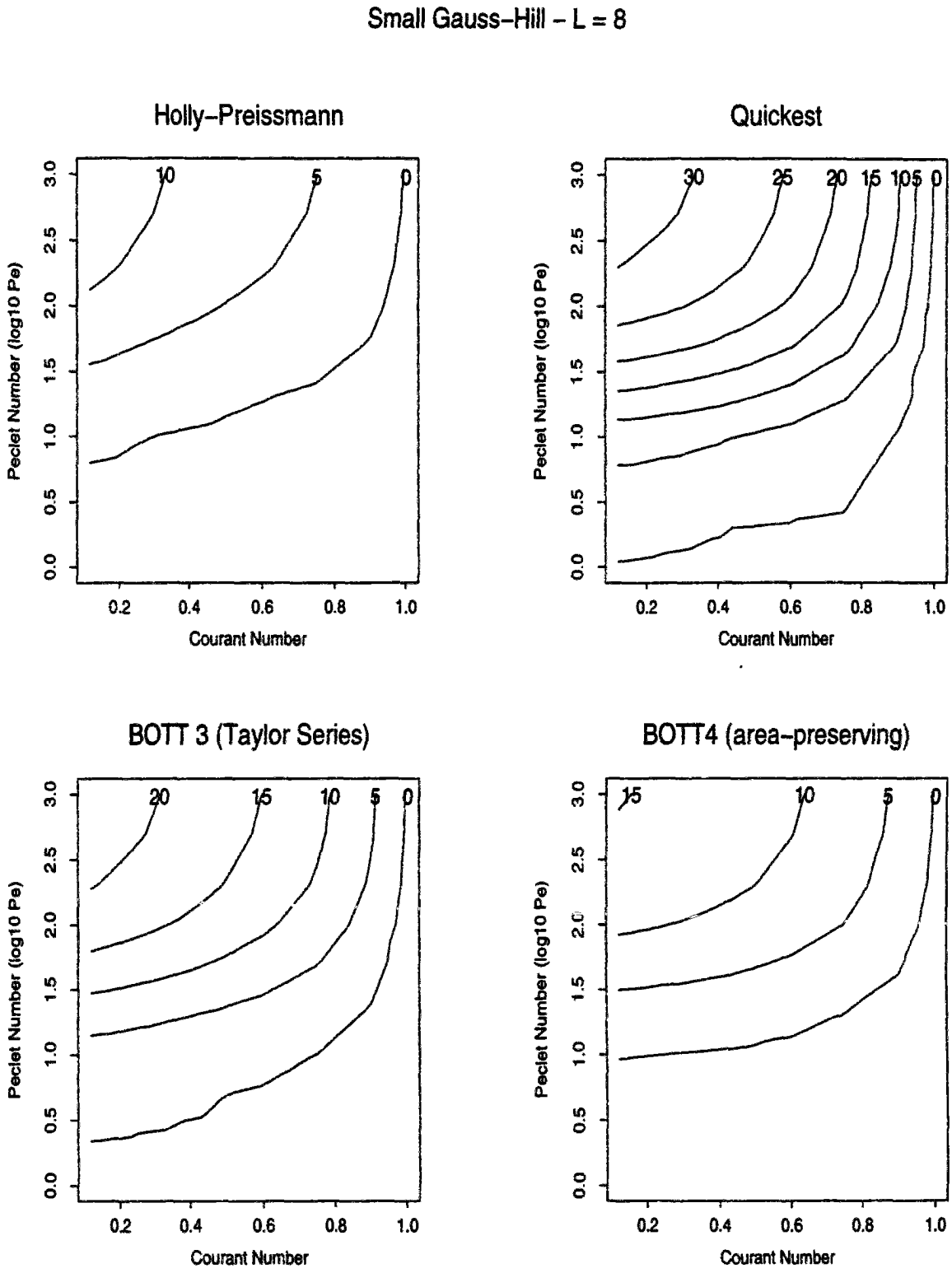


Figure 6.32: Peak damping (%) isolines v.s. Peclet and Courant numbers - Flux-form methods

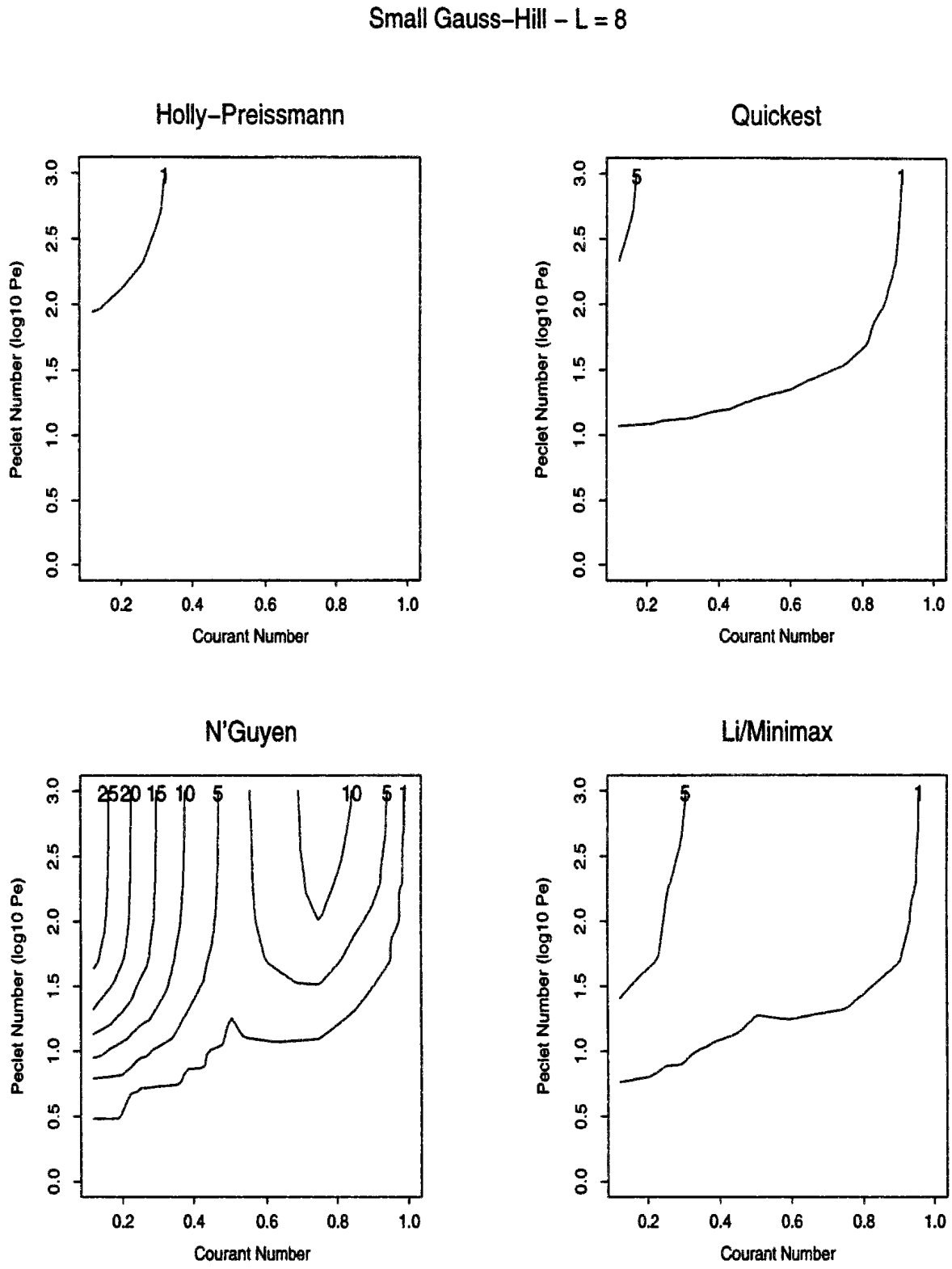


Figure 6.33: Maximum Undershoots isolines (in % of peak value) v.s. Peclet and Courant numbers

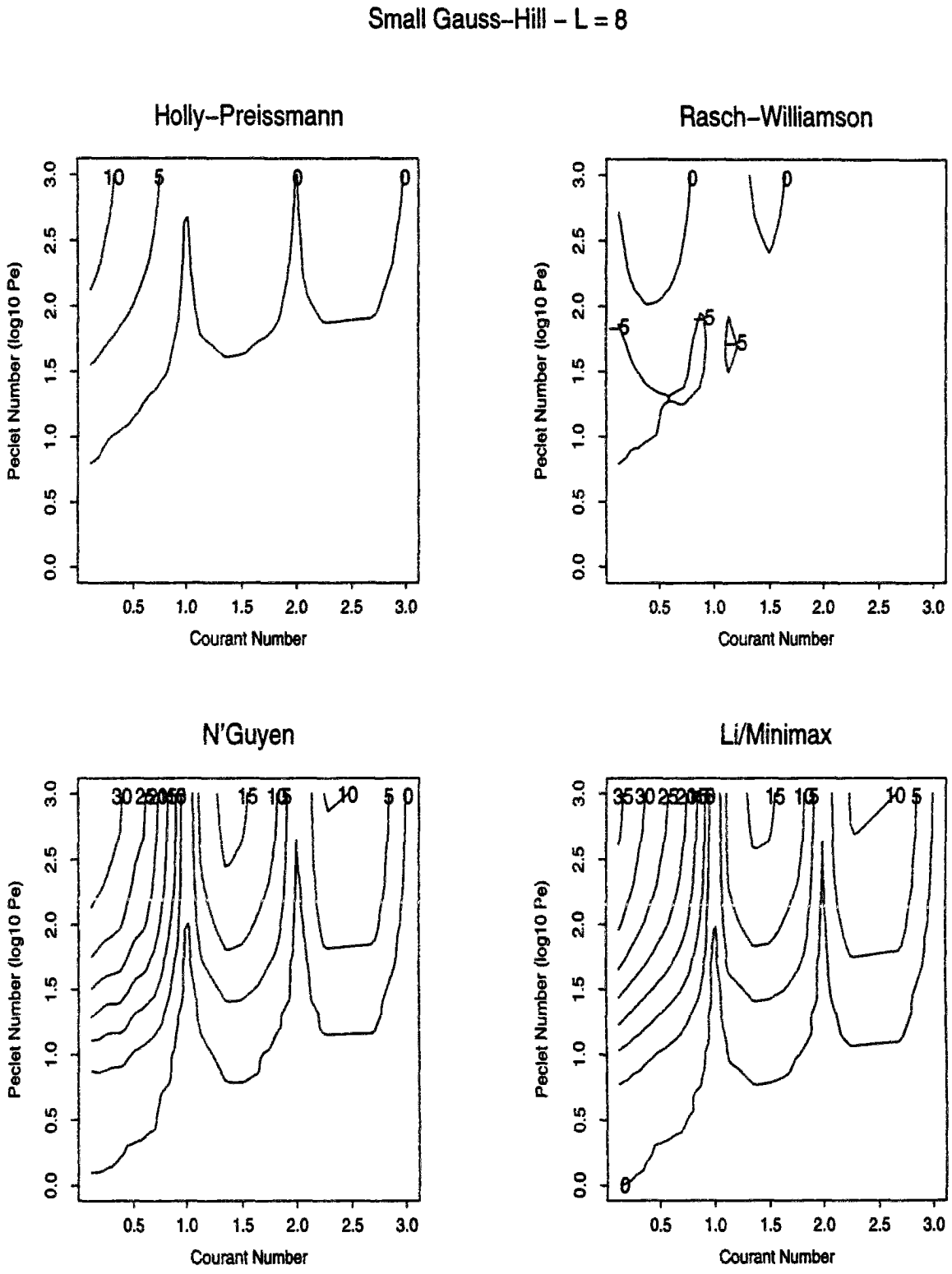


Figure 6.34: Peak damping isolines : Extended range of Courant numbers, Characteristics methods

6.5 Problems controlled by open boundaries conditions

This far we have been dealing only with tests concerned with the transport of a scalar field initially well defined within the computational domain. In the vicinity of boundaries, each of the tested schemes requires some adaptation, which generally results in a lowering of the scheme order. Thus, it appears necessary to check the schemes behaviour when the problem is controlled by its open boundary conditions. Hereafter, this analysis is presented only for the RASCH, HOLLY, BOTT3 and BOTT4 algorithms. Some complementary results relative to the MINIMAX, DAN and QUICKEST schemes are given in Appendix F.3.

6.5.1 Advancing front

In this set of problems, the concentration field is imposed by a constant mass flux, specified through constant velocity and upstream concentration :

- *Initial conditions* $C(x,0) = 0$ for $0 \leq x < \infty$;
- *Boundary conditions* $\forall t > 0$, $C(0,t) = 1$ and $\lim_{x\to\infty} C(x,t) = 0$
- *flow conditions* uniform velocity $U = 0.5$ m/s
- *Discretization* The computational domain is the same as for the concentration-hill tests. Constant and variable Grids 1 and 2 have been used. Time steps up to 400 s, as specified in table 6.2, have been applied in combination with Grid 1. Time steps used in combination with Grid 2 are indicated in table 6.3.
- *Diffusion* We performed tests with pure advection ($\Gamma = 0$ m^2.s^{-1}), intermediate ($\Gamma = 2$ m^2.s^{-1}) and important ($\Gamma = 50$ m^2.s^{-1}) diffusion.

As suggested in (Baptista *et al.*, 1988), the relevant error measures for this test case concerns the mass preservation (μ_0), the L2 error norm (Φ), the analysis of overshoot and undershoot problems (through ϵ and Ψ values). Exact solution is given by

$$C(x,t) = \frac{1}{2}\left(\operatorname{erfc}\left(\frac{x-Ut}{2\sqrt{\Gamma t}}\right) + \exp\left(\frac{Ux}{\Gamma}\right)\operatorname{erfc}\left(\frac{x+Ut}{2\sqrt{\Gamma t}}\right)\right)$$

when $\Gamma \neq 0$ and, for pure advection, is defined by

$$C(x,t) = \begin{cases} 1 & x < Ut \\ 0 & \text{otherwise} \end{cases}$$

As previously, errors are monitored after 10800 s. Their evolution with respect to the number of iterations (or time steps) required to reach this time is plotted in figures 6.35 to 6.38 for the

case of pure advection. Solid and dashed lines refer respectively to constant and variable grid results.

Considering the L2 error-norm (fig. 6.35), this test appears to be less severe than the previous tests related to the transport of concentration-hills. The hill whose shape is the most similar to the front shape is the square wave : L2 error norms are about twice as big in the square wave test than observed here.

On the other hand, the specification of the inflow boundary conditions does not allow us to ensure a perfect mass conservation, even for the schemes which performed well on this point (BOTT3, BOTT4 and HOLLY). Mass error remains less than 2% except for the RASCH algorithm with some specific combinations of spatial and temporal discretizations : errors of 3 and 2.3 % respectively on Grid 1 for the smallest Δt, respectively equal to 48 and 100s; an error of 3.5 % on Grid 2 for $\Delta t = 240$s.

All schemes induce overshoots at the upper edge of the front and undershoots at its foot, except for the positive definite schemes BOTT3 and BOTT4 (cf figure 6.39, with the same legend as fig. 6.8). However, these spurious features are small (less than 6% in any case, as illustrated in fig. 6.37 and 6.38). In figure 6.39 we may also notice that the BOTT3 and BOTT4 results are nearly indistinguishable and that the scheme most sensitive to under- and overshoots, namely HOLLY, compensates for this by the fact it displays both less dispersion and phase shift than the other schemes.

While the behaviour of BOTT3 and BOTT4 with respect to the number of iterations is nearly monotonic, HOLLY and RASCH exhibit error maxima for specific Courant number values. We could not elucidate why.

On the whole, the grid irregularity has little influence on the quality of the results.

The introduction of a small amount of dispersion improves the results. For $\Gamma = 2$ $m^2.s^{-1}$, the normalized L2 error norm is always less than 0.07 % for BOTT3, BOTT4, HOLLY and less than 0.1 % for RASCH. The relative mass error is inferior to 2 % for BOTT3, BOTT4, HOLLY, and less than 2.9 % for RASCH. The maximum overshoot over the whole range of admissible time steps for each scheme (we tested time steps up to 1200s for the backward characteristics methods) stays less than 2.7, 0.27, 2.2, 1.5 % for HOLLY, RASCH, BOTT3 and BOTT4 algorithms respectively. Similarly, the maximum undershoot remains less than 1.35 and 0.035 % for HOLLY and RASCH respectively. Errors for the strong diffusion case ($\Gamma = 50$ $m^2.s^{-1}$) are negligible.

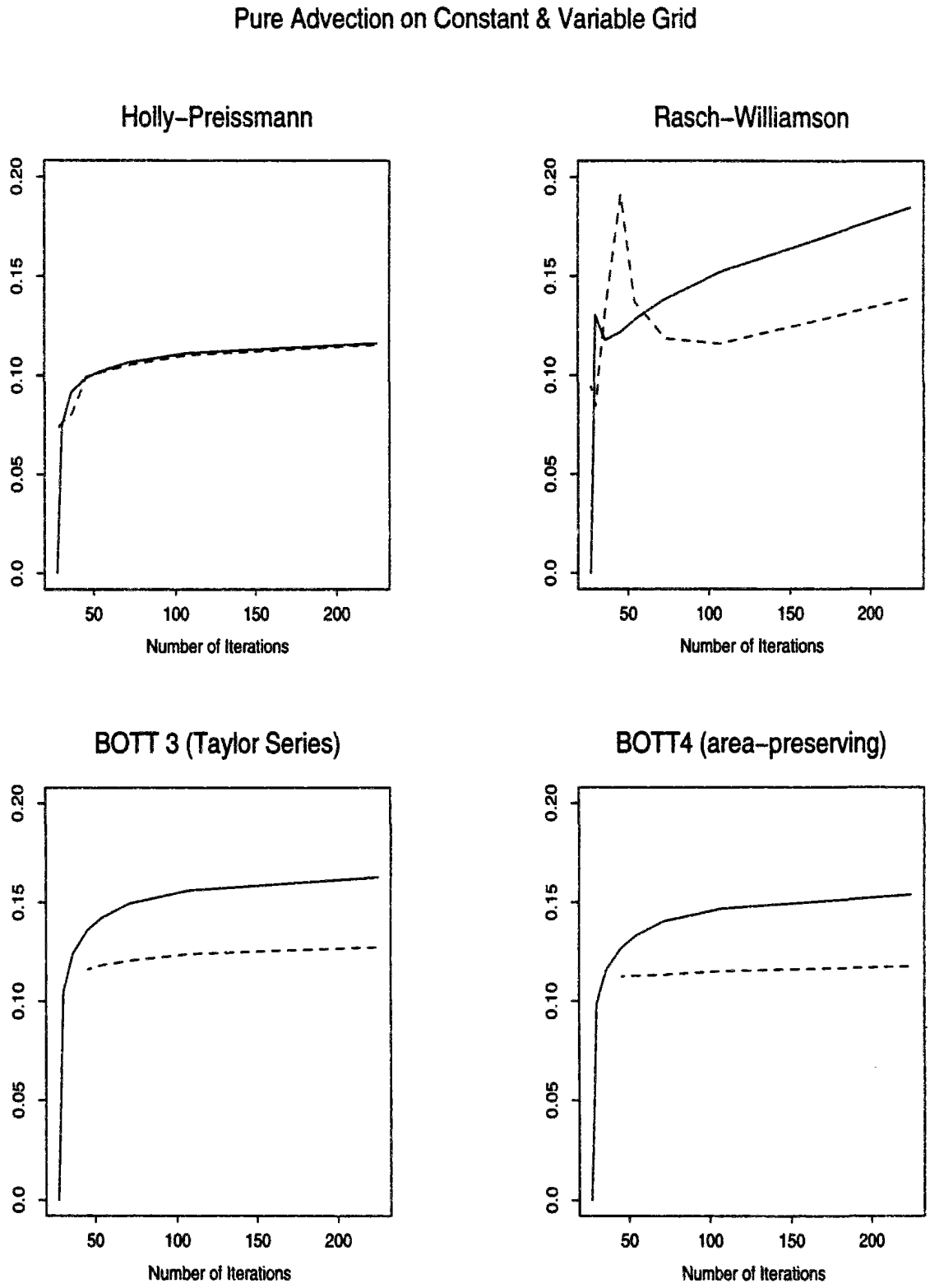


Figure 6.35: Advancing front : L2 error norm (in % of total mass)

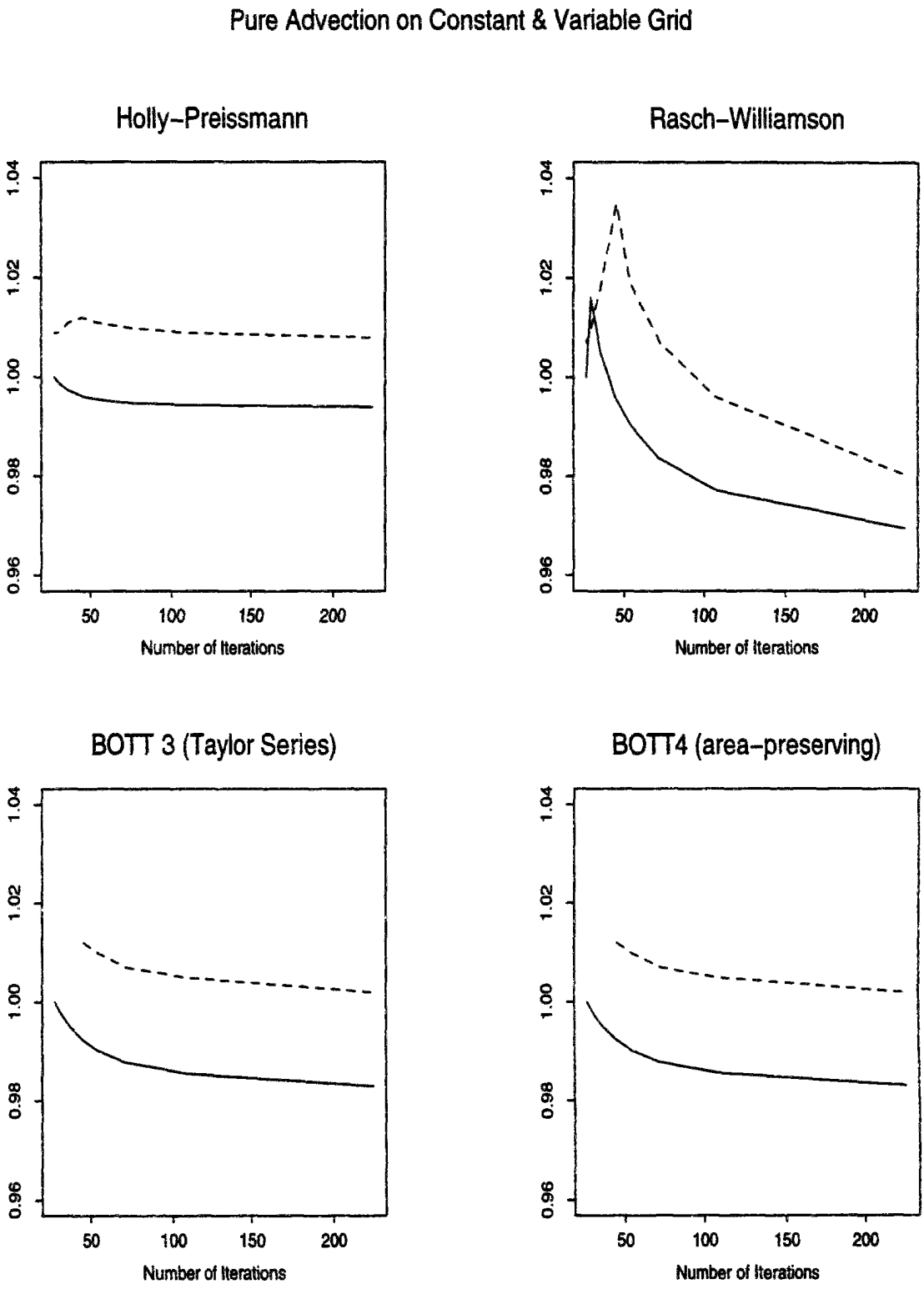


Figure 6.36: Advancing front : ratio of computed to exact mass

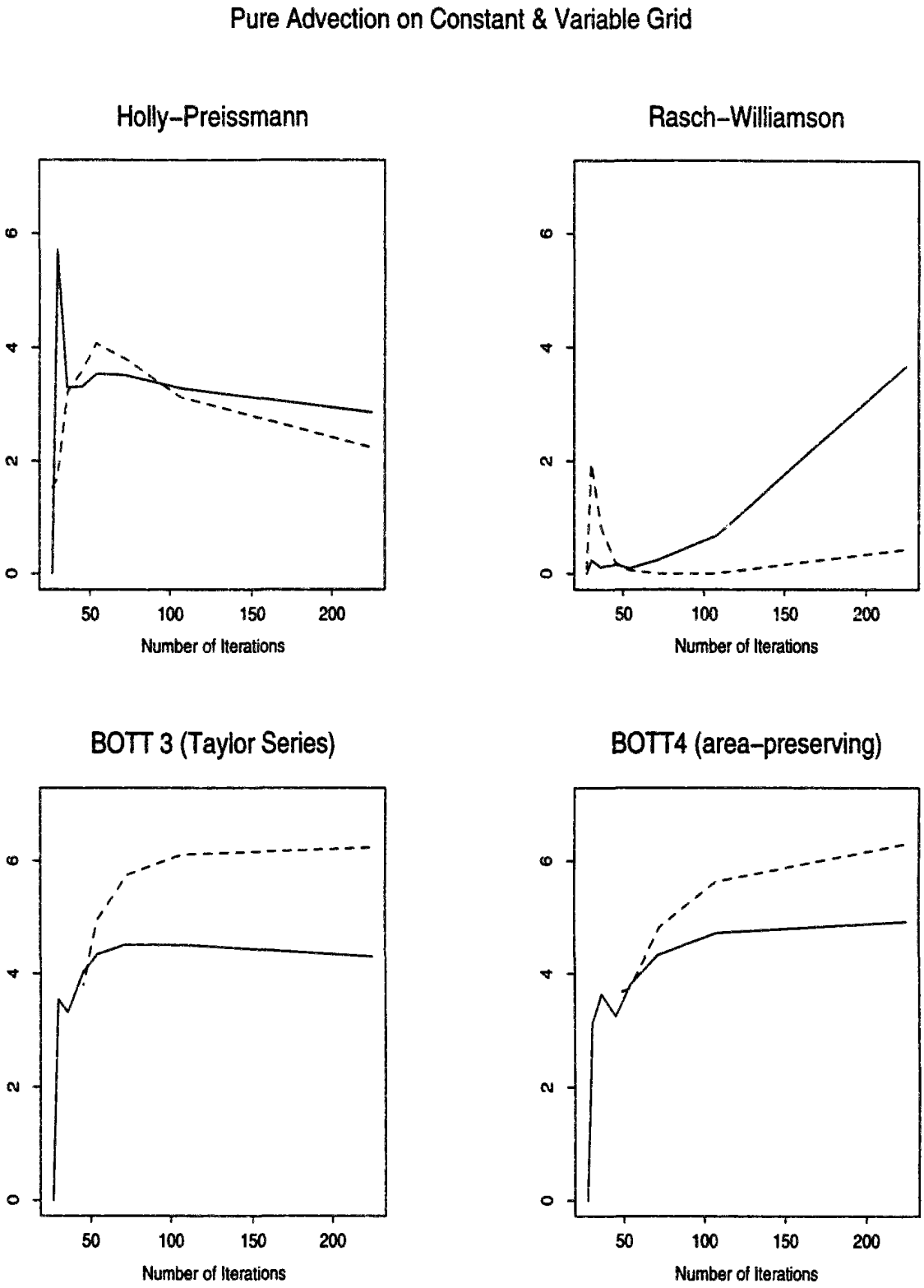


Figure 6.37: Advancing front : Maximum overshoot (in % of front concentration)

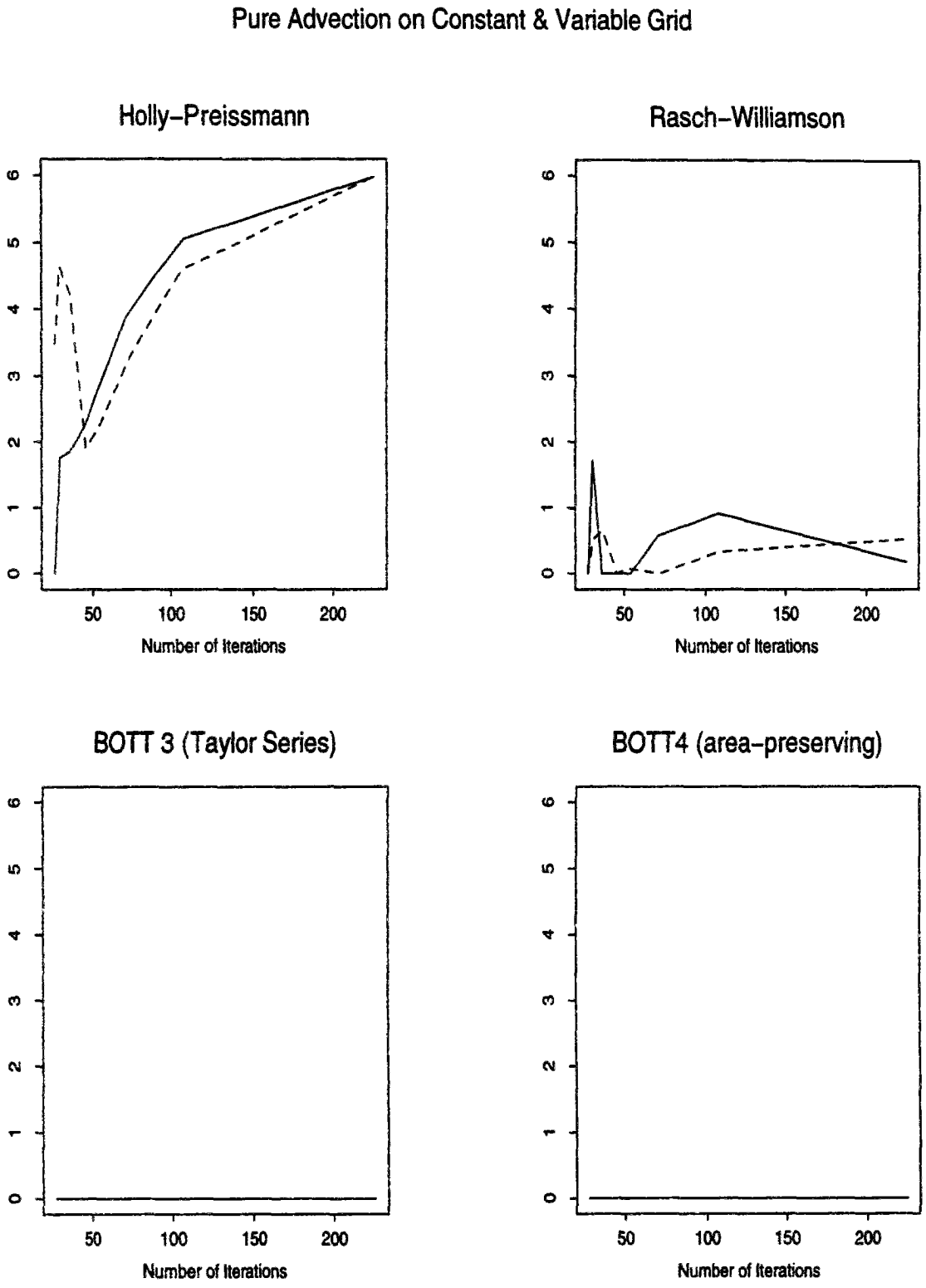


Figure 6.38: Advancing front : Maximum undershoot (in % of front concentration)

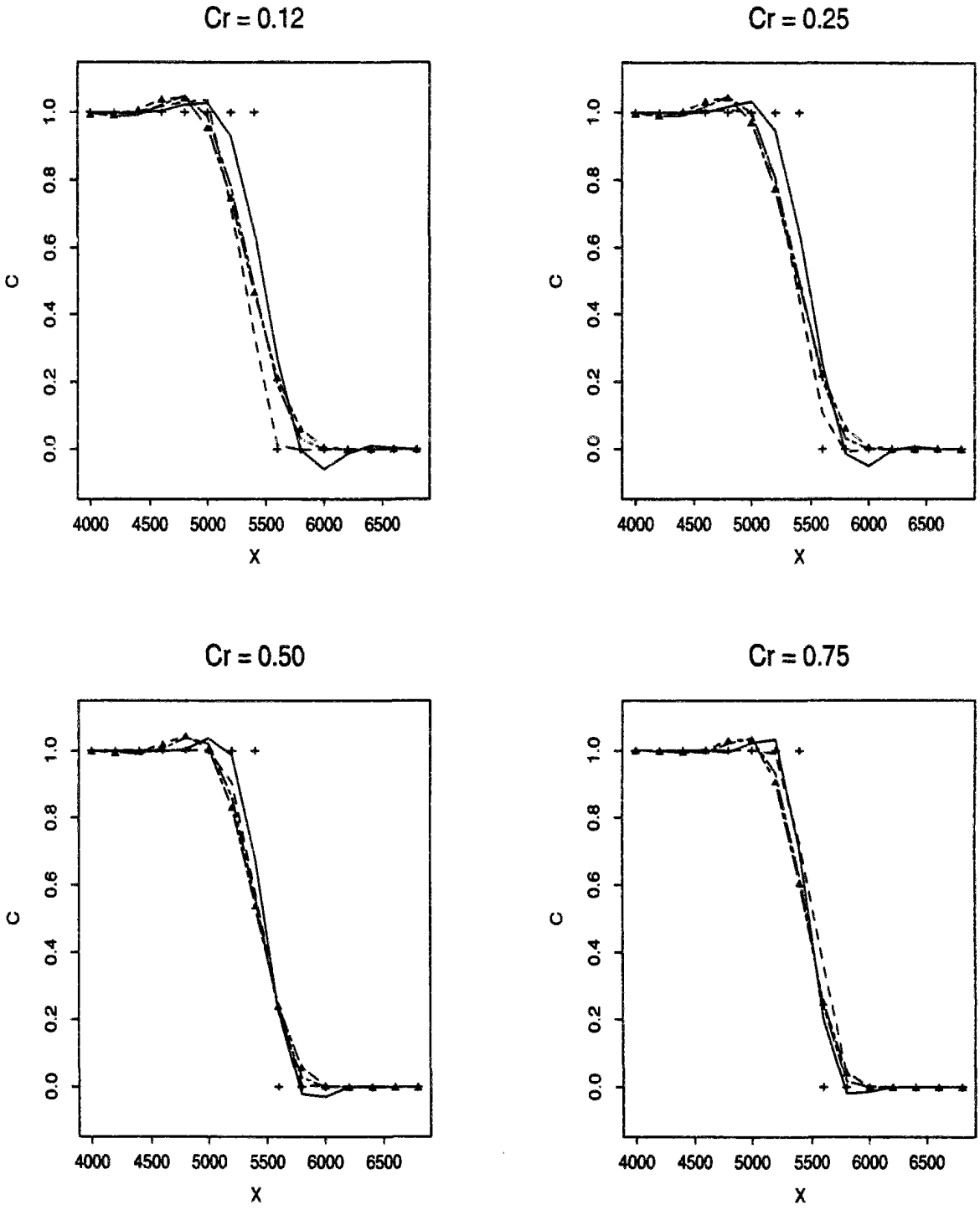


Figure 6.39: Advancing front. Overview of the performance of the best schemes

A look at appendix F.3 confirms that this test case is indeed "easier" than the previous ones : while the Dan N'Guyen scheme raises again unacceptable results, on the other hand the performances of the Quickest and Minimax algorithms are this time much closer to the best schemes' performances than for the concentration-hills tests.

6.5.2 Propagation of gauss-hill inputs

This test deals with the pure advection (in uniform flow $U = 0.5$ m/s) of more complicated inputs than the front. The initial condition is given by

$$C(x,0) = \exp -\frac{(x-x_0)^2}{2.\sigma_0^2}$$

while at the upstream boundary, the following concentration law is imposed :

$$C(0,t) = \exp -\frac{(x_0+Ut)^2}{2.\sigma_0^2}$$

σ_0 is set to 264 m. In section 6.3, x_0, the centre of mass of the initial distribution, was located 2000 m downstream the inflow boundary, so that most of the input was already spread within the computational domain. In (Baptista *et al.*, 1988), it was proposed to investigate the case $x_0 = 600$m, which means that the tail of the distribution has not yet entered the domain at the beginning of the computation. We have been also studying the more severe cases corresponding respectively to $x_0 = 0$ and $x_0 =$-600 m : in these situations, the boundary condition is essential in prescribing the input. For all computations, a uniform grid (Grid 1) has been used. The error measures are the same as for the previous concentration-hill tests and are also evaluated for $t =$10800 s. Their evolution with respect to the Courant number values is plotted in figures 6.40 to 6.46 :

- Solid lines refer to results of the first test with $x_0 = 2000$.
- Dashed lines are used for the other tests. The square, triangle and cross marks refer respectively to conditions $x_0 = 600$, 0 and -600 m.

Let us first discuss results in the area defined by $c_r \leq 1$ (figures 6.40 to 6.44 and the upper half of figure 6.45). As regards peak damping (fig. 6.40) and the L2 error norm (fig. 6.41), there exist few differences between the results obtained for $x_0 = 2000$ or 600m, namely when most of the concentration-hill lies initially within the studied domain. As could be expected, there is some degradation of the performance for the two other cases. However, this loss of accuracy is but slight, which suggests that with time steps Δt up to 400s (which corresponds to $c_r = 1$) the inflow concentration law is satisfactorily described, so that upstream imprecisions barely spoil the schemes' results. Differences are more noticeable when considering other indicators. For

instance, it appears that mass conservation is no longer strictly enforced (see fig. 6.42). Numerical spreading (fig. 6.44) and phase shift (fig. 6.43) tend also to be slightly more important. However, there is nothing dramatic in these evolutions : **the boundary condition treatment embedded in each scheme seems to be fairly robust.**

For backward characteristic methods, it is possible to investigate further the influence of the choice of bigger time steps. In the case of a fully defined initial source (cf 6.3), the choice of bigger time steps improved the performance (more precisely when c_r^a and c_r^b had the same decimal part and $c_r^a > c_r^b$, results with c_r^a were consistently better). Such is not the case here, as can be seen on figures 6.45 and 6.46, essentially for situations corresponding to $x_0 = 0$ and 600 m. There are critical time steps for which the input (especially a striking feature such as the peak value) happen to be too crudely described : this yields for instance dramatic damping. As could be logically expected, *when inflow boundary conditions play an important part, temporal (as well as spatial) discretization must be tailored so that it allows capture of all the relevant features of the input.*

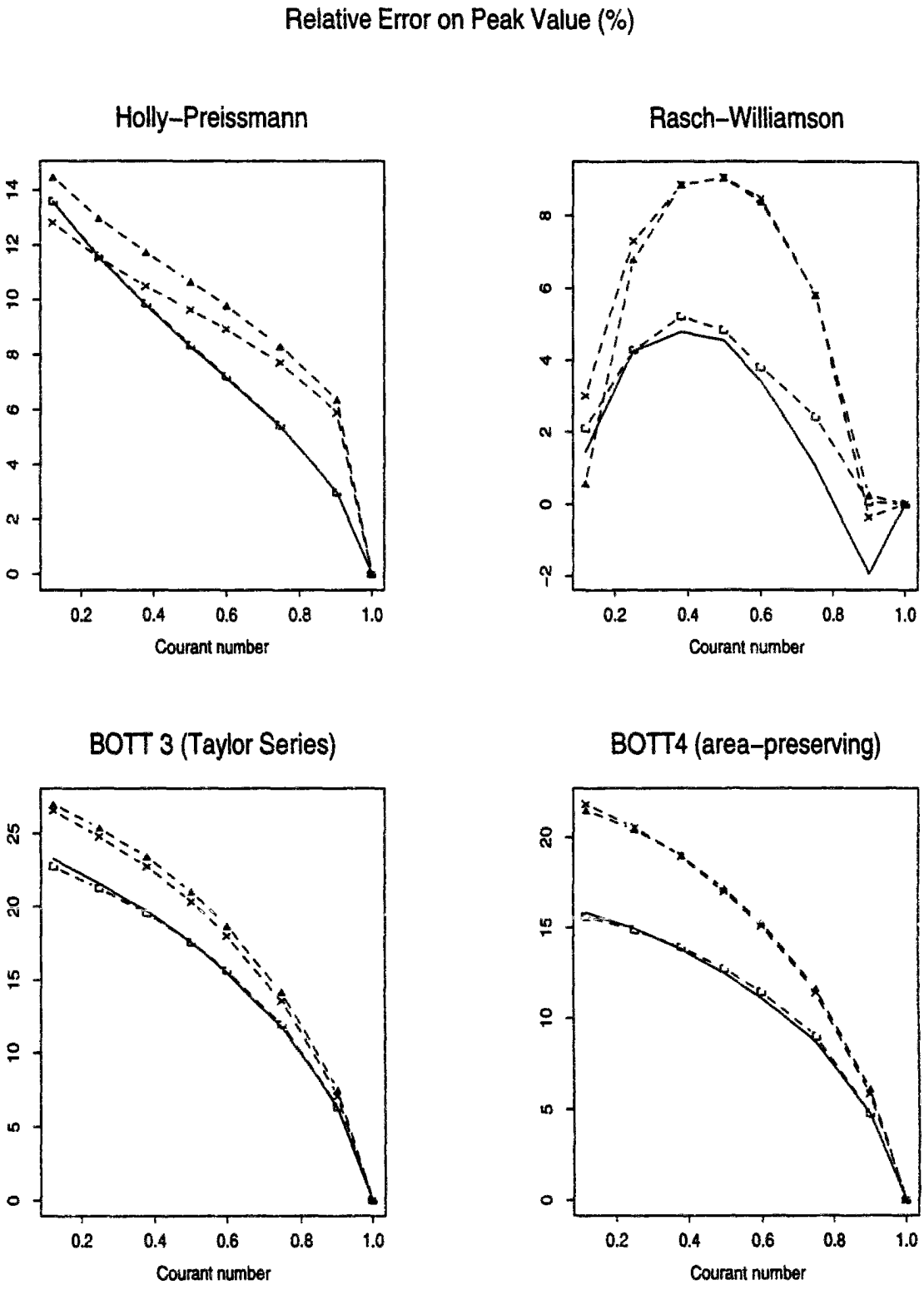


Figure 6.40: Advection of Gauss-hill inputs : Damping

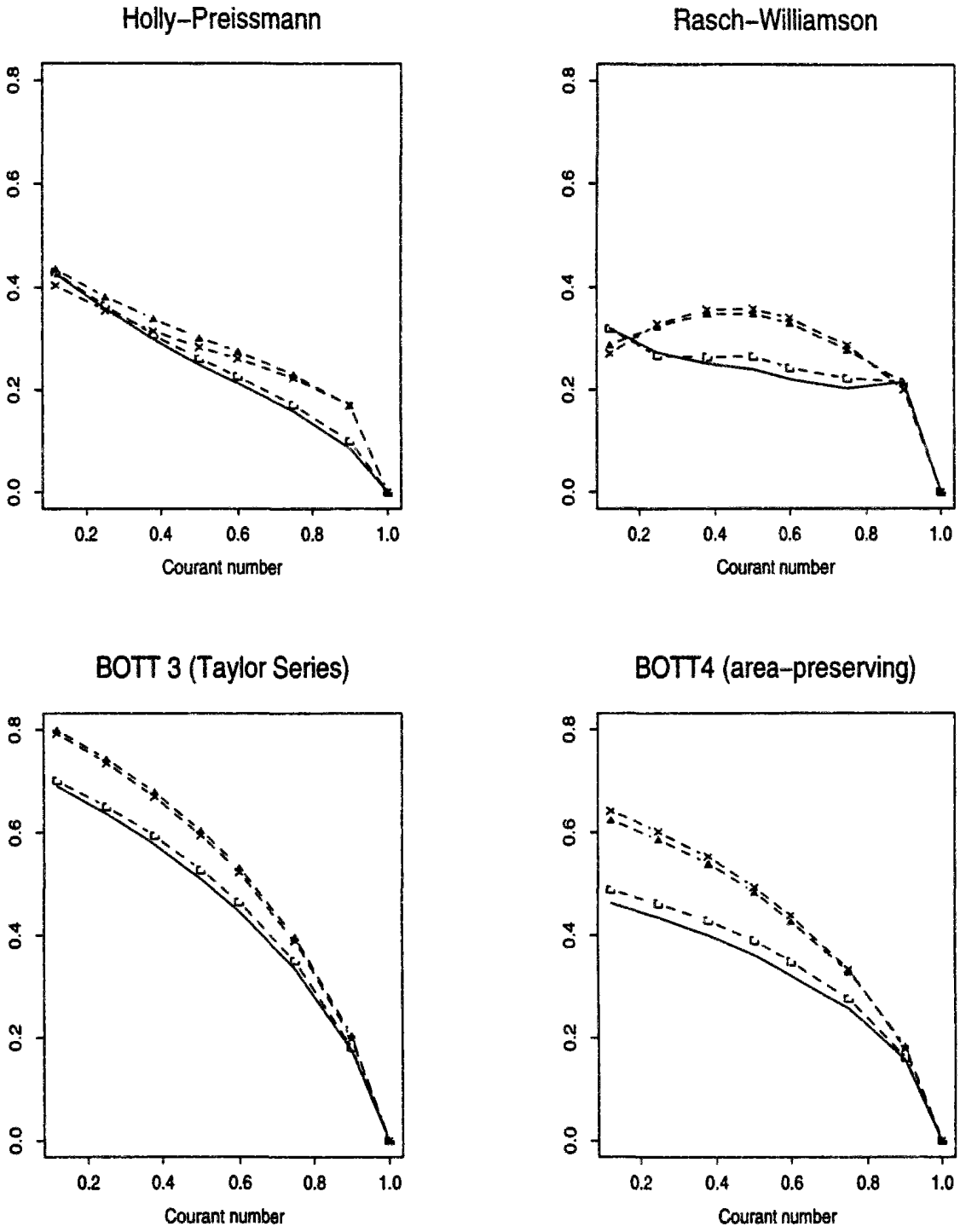


Figure 6.41: Advection of Gauss-hill inputs : L2 error norm (in % of total mass)

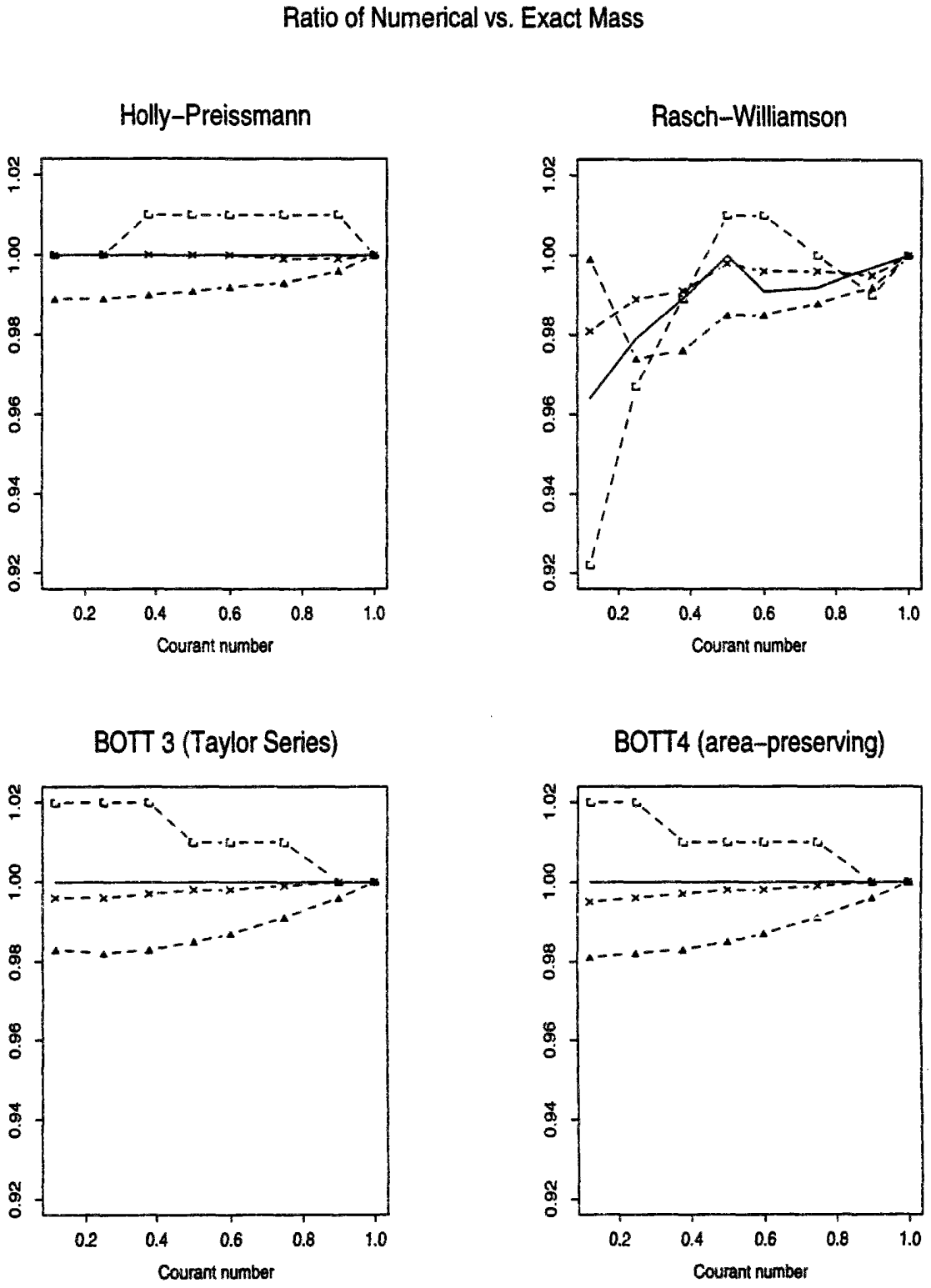


Figure 6.42: Advection of Gauss-hill inputs : Mass preservation

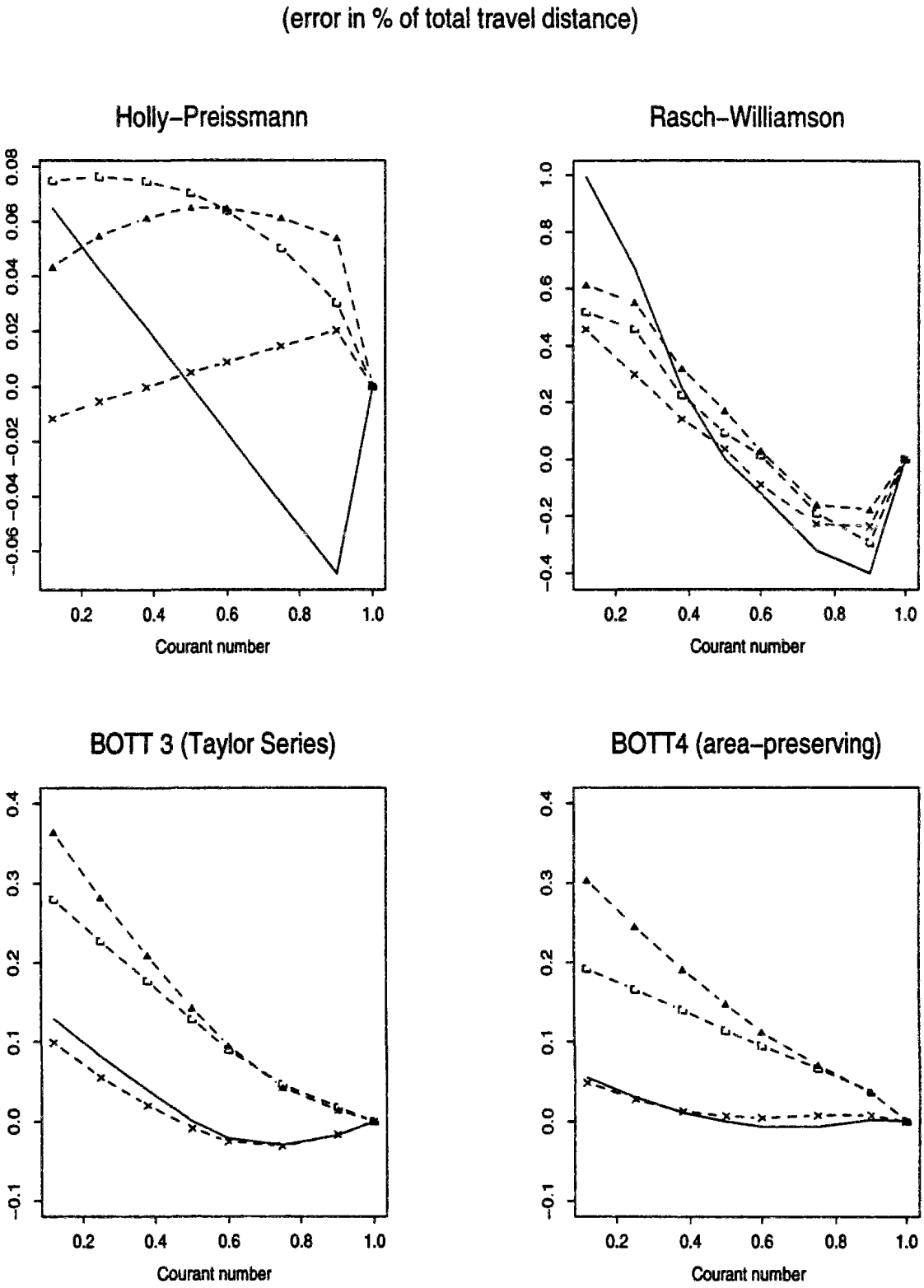


Figure 6.43: Advection of Gauss-hill inputs : Global phase shift

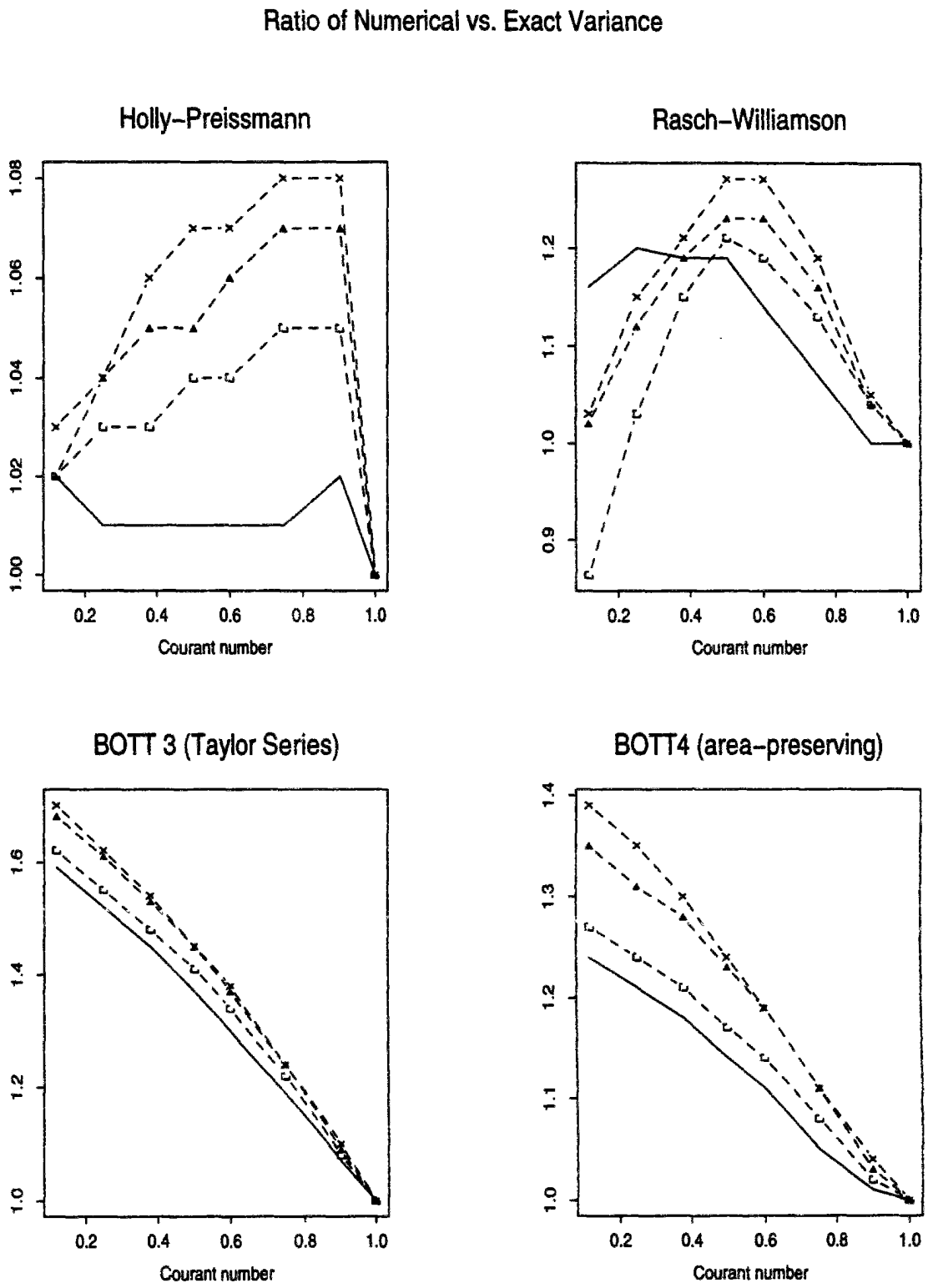


Figure 6.44: Advection of Gauss-hill inputs : Numerical spreading

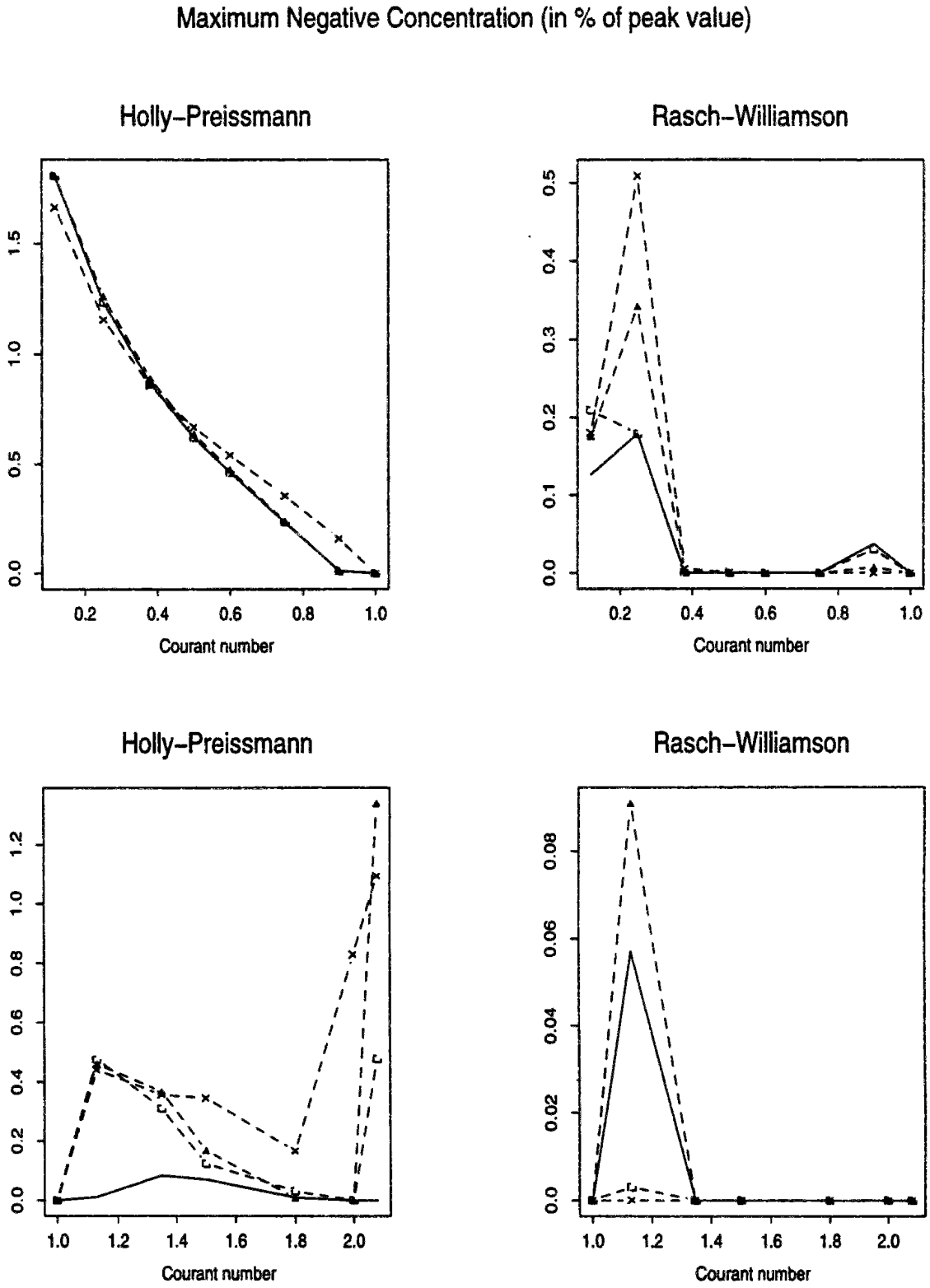


Figure 6.45: Advection of Gauss-hill inputs : Undershoots

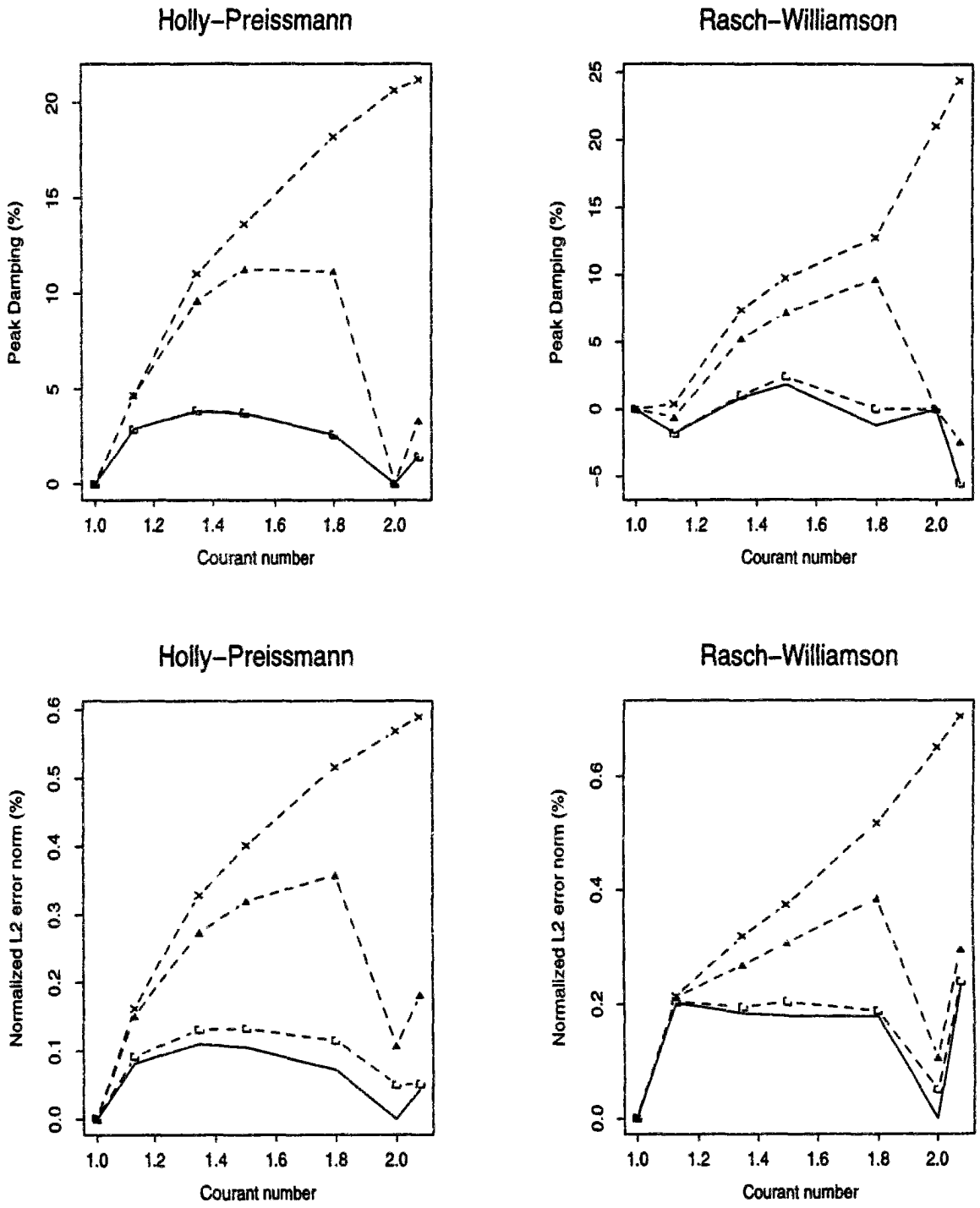


Figure 6.46: Influence of bigger time steps on Damping & L2 error norm

6.6 Computer time requirements

Hitherto we have been focussing only on the schemes' accuracy. It is time to have a look at their computational efficiency. We have thus been studying the computer time requirements of each algorithm, according to the following procedure :

i The CPU times mentioned below are an average established over about fifty simulations of gauss-hill transport. These have been run on a SUN SPARC Station 10.

ii Most tested versions of the schemes are devoid of any simplification possible in case of steady flows (e.g. computing once and then storing the characteristic foot locations) or uniform grid spacing. These standard versions could apply without any modification to any one-dimensional advection-diffusion problem. They are written in standard FORTRAN.

The only exception concerns the HOLLY algorithm : in the step dealing with the advection of the concentration derivative, we have not included the computation of corrective source terms (cf eq. 4.19 in sec. 4.1.2.1), which would be necessary in the case of non-uniform flows. Consequently, we are slightly underestimating this scheme's computational cost.

We have also been studying a simplified version of BOTT4, valid only in case of uniform grid spacing, where the approximation polynomials used in computing the flux integrals are quickly defined with the help of the formulae of table F.2 (app. F.1.1). In the "standard" version of BOTT4, a 5×5 linear system must be solved at each grid node in order to assess the polynomial coefficients (cf F.1.1). The corresponding matrices depend only on grid spacing. Consequently they are computed, inverted, and their inverse stored during the initialization stage of each run.

The generalizations to variable flows and grids of the third-order accurate flux-form method are different according to (Takacs, 1985) and (Leonard, 1979). They correspond respectively to the TAKACS and QUICKEST algorithms. Both generalizations raise the same numerical results as far as we have been testing them. We have assessed both computational requirements.

iii With an option of the Fortran compiler, it is possible to get a precise description of each run (a "profile") : the number of calls to some function or subroutine, CPU consumed by each subroutine, etc ... This device was used to distinguish between time wasted in initializing the run, managing entries and outputs of the simulation and solving the advection and diffusion steps.

The results of this study are presented in table 6.6 and 6.7 for flux-form and backward characteristics methods respectively. The indicated times correspond to the CPU time required to solve the related operation for one grid node and one time level. They are expressed in

microseconds (10^{-6} s). We have also indicated the relative cost of each scheme with respect to the quickest one, which happens precisely to be the QUICKEST algorithm.

Table 6.6: One dimensional case : CPU requirements for flux-form methods

Scheme	QUICKEST	TAKACS	BOTT3	Standard BOTT4	Simplified BOTT4
CPU advection stage (μs)	7.6	8.7	27.9	63.9	19.8
CPU diffusion stage (μs)	8.7	8.7	8.7	8.7	8.7
advection + diffusion	16.3	17.4	36.6	72.6	28.5
	Ratio of CPU by corresponding Quickest CPU				
advection	1	1.15	3.67	8.41	2.61
diffusion	1	1	1	1	1
Total	1	1.07	2.25	4.46	1.75

We shall consider the relative costs of the best performing schemes, namely BOTT3, BOTT4, HOLLY and RASCH algorithms. In particular, it is not useful to dwell on LI or DAN schemes : they have costs equivalent to those of RASCH and HOLLY but raise far worse results ! Similarly, while it is quite economical, QUICKEST/TAKACS algorithm should be avoided for advection-dominated problems (see above).

1. As regards backwards characteristics methods, the most consuming stage consists of back-tracking the fluid particles. The CPU time needed depends on the precision involved in the characteristic location, which, as explained in E.1.1, depends itself on how we break the trajectory computation. The times indicated in table 6.7 are reached when the trajectory is broken into three segments within each mesh cell. Should we decide to break it into two segments/cell only, the cost is reduced to 41.4 μs per grid node and time step. This cuts down the characteristics methods cost by 25 % approximately.

2. We may note that interpolating the derivative at the characteristic origin or estimating it by the Akima method induces similar costs.

3. As could be expected, the HOLLY diffusion step is more costly than for the other methods since two variables instead of a single one are diffused. However, the cost does not double,

Table 6.7: One dimensional case : CPU requirements for backward characteristic methods

Scheme		RASCH	HOLLY	LI/MINIMAX
ADVECTION	Backtracking	60	60	60
	Interpolating	8.6	8.5	6.5
	Total	68.6	68.5	66.5
DIFFUSION		8.7	11.7	8.7
ADVECTION + DIFFUSION		77.3	80.2	75.2
Ratio of CPU by corresponding Quickest CPU				
advection		9.02	9.02	8.75
diffusion		1	1.35	1
Total		4.74	4.92	4.61

since the most costly operations of this stage, namely computing the diffusion matrix and its inverse (cf E.2), are only performed once.

4. The need to cope with a variable grid has dramatic effect on the BOTT4 efficiency : the cost of the advection step is multiplied by a factor 3.2 . Contrary to their poor reputation, backward characteristics methods are not so dramatically the most CPU time consuming !

   Besides, when comparing relative costs, it is interesting to consider what occurs with multiple variable problems. If we work with two different transported scalars, the cost of the advection stage doubles for flux-form methods, not for backward characteristic ones, as the backtracking operation needs to be performed only once, whatever the number of transported variables !

   It appears that the criticism commonly voiced against backward characteristics methods, namely that they are too costly, is neither fair nor sound : high-order flux form methods are as costly in general situations (non-uniform grid spacing). What we have been illustrating with a scheme based on fourth-order polynomial approximations (BOTT4) would of course be accentuated with higher-order methods. A possible response would be to use different schemes according to the localisation of the grid nodes, as advocated in (Leonard & Niknafs, 1991) : in regions where gradients and curvature of the transported scalar field are smooth, a low order scheme like QUICKEST is used; in the remaining parts of the computational domain, which should probably be of limited extent, higher-order algorithms are applied. However, to implement this approach, one needs first to clarify the criteria governing the switch between the different schemes.

In summary, the schemes which achieve the best balance between accuracy and computational efficiency in one-dimensional general situations are the Holly-

Preissmann and Rasch-Williamson backwards characteristics methods. For cases with moderate or dominant diffusion, the flux-form method based on a 3^{rd}-order Taylor Series expansion (BOTT3) is an interesting alternative. Finally, when it is possible to use a uniform spatial discretization, the most interesting scheme, for 1 to 5 different transported scalars, is the BOTT4 flux-form method, based on 4^{th}-order area-preserving polynomials.

6.7 Résumé français : "Cas-tests monodimensionnels"

Comme indiqué en conclusion du chapitre 4, on ne peut se baser uniquement sur une revue bibliographique pour apprécier équitablement la performance de schémas numériques. En effet, les tests appliqués diffèrent d'un article à un autre, leur description est souvent succinte, toutes choses qui ne favorisent pas la comparaison.
Un test "idéal" doit permettre une évaluation sans ambiguité des algorithmes : pour celà, rien de tel qu'un problème où l'on dispose d'une solution analytique de référence. Compte tenu de la complexité des équations d'écoulement et de transport, ces situations sont rares. Dans le cas de l'équation de transport, la plupart ont été compilées par les organisateurs du Forum Convection-Diffusion (Baptista *et al.*, 1988), dans leur effort pour établir un cadre commun destiné à l'évaluation objective des méthodes de résolution de l'advection-diffusion. Nous nous sommes donc appuyés sur ces tests, en les élargissant quelque peu (plus de pas de temps, plus de diffusivités testés). Ces cas-tests ne concernent que des écoulements uniformes. On peut dire que chacun s'apparente à un problème pratique, typique des applications grandeur nature.

Rappelons que nous comparons 7 schémas, sélectionnés en premier lieu d'après une étude bibliographique (cf section 4.3). 3 des méthodes reposent sur le calcul de bilans de masse des flux advectifs, ces derniers étant évalués par intégration d'une distribution approchée des concentrations : il s'agit des méthodes QUICKEST, BOTT3 et BOTT4 où la distribution approchée est respectivement un polynôme d'ordre 2,3 et 4. Les 4 autres méthodes sont des schémas aux caractéristiques : HOLLY, RASCH, DAN, LI. Les 2 premiers utilisent l'interpolateur cubique hermitien, les 2 derniers des interpolateurs quadratiques.

Advection pure par un écoulement uniforme La première série de tests concerne l'advection pure de distributions de concentration caractérisées par leur "petite taille" (i.e. leur définition par un faible nombre de noeuds de calcul) et leurs forts gradients. **On se retrouve dans des conditions proches de ce qui se passe au voisinage de rejets ponctuels fortement chargés par rapport au mileu récepteur.**

On procède tout d'abord avec un écoulement permanent et une grille de calcul uniforme (section 6.3.1) ($\Delta x = 200$ m). Les distributions étudiées sont de forme gaussienne (répartie sur 8 mailles), triangulaire (base du triangle : 8 mailles) et enfin, en forme de plateau (base de 12 mailles). Les pas de temps testés sont tels que les nombres de Courant correspondant vont de 0.12 à 3 (nb : nbres de Courant supérieurs à 1 appliqués uniquement pour les schémas aux caractéristiques, la limite de stabilité des autres méthodes étant $c_r \leq 1$).

- **Transport d'une gaussienne**
  - **Amortissement du maximum**
    Les erreurs sont une fonction décroissante du nombre de Courant pour $c_r \leq 1$. Elles dépendent de sa partie décimale au delà (avec un maximum pour $c_r \equiv 0.5$).
    Le schéma qui préserve au mieux le pic est RASCH (erreur allant d'une sous-estimation de 5 % à une surestimation de 2 %), suivi de HOLLY et BOTT4 - aux résultats proches (atténuation généralement inférieure à 10 %) - puis de BOTT3. DAN, QUICKEST et LI ont des résultats quasiment identiques ...et médiocres (30 % d'erreur pour $c_r \leq 0.5$).
    L'amortissement du pic va de pair avec un certain étalement de la base de la gaussienne, deux faits qui reflètent l'existence d'une diffusion numérique parasite.
    RASCH a tendance a transporter la gaussienne en la modifiant légèrement en une forme plus triangulaire.
  - **Apparition de concentrations négatives parasites**
    BOTT3 et BOTT4 préservent la positivité par construction. RASCH et HOLLY engendrent peu de valeurs négatives (respectivement d'une amplitude égale au plus à 0.2 et 2 % de la valeur du maximum). LI et QUICKEST se comportent correctement (amplitude limitée à 5.3 et 6.5 % respectivement, le pire résultat étant obtenu pour le plus faible c_r). Par contre les résultats de DAN sont mauvais (amplitude atteignant 30 %).
  - **Conservation de la masse**
    Les schémas QUICKEST, BOTT3 et BOTT4 sont conservatifs par construction. HOLLY, DAN et LI ont une erreur négligeable. La perte de masse est un peu plus importante pour RASCH aux faibles nombres de Courant (perte de 3.6, 2.1 et 1.1 % pour $c_r = 0.12, 0.25, 0.38$ respectivement).
  - **Déphasage** Il est modéré, voire négligeable, pour tous les schémas sauf DAN.

- **Transport d'une distribution en triangle** Les résultats de ce test confirment en tous points ceux du précédent, la supériorité de RASCH étant cette fois-ci indiscutable.

- **Transport d'une distribution "en plateau"**
  En ce cas, les troubles numériques se manifestent par l'apparition de concentrations négatives à la base du plateau et par une surestimation locale des concentrations aux angles supérieurs de ce plateau.
  Bien entendu, BOTT3 et BOTT4 n'engendrent pas de valeurs négatives. Cependant, tous critères pris en compte, ce sont les schémas RASCH et HOLLY qui permettent le meilleur contrôle des erreurs parasites. DAN fournit de mauvais résultats. Ceux des 4 autres schémas sont honorables, et équivalents.

En conclusion, **cette première application met en évidence le bon comportement numérique des 2 méthodes aux caractéristiques utilisant l'interpolateur hermitien,**

en premier lieu RASCH, en second HOLLY. Les 2 algorithmes positifs par construction, à savoir BOTT4 et BOTT3, donnent également des résults corrects. Les 3 autres sont nettement moins satisfaisants, tout particulièrement DAN.

L'influence de la variabilité des pas d'espace de la grille de calcul **est testée sur le transport de gaussiennes (section 6.3.2)**, la gaussienne précédente et une distribution 50 % plus large. Les pas d'espace varient de façon sinusoidale. Leur valeur moyenne est toujours de 200 m, mais ils peuvent s'écarter de cette valeur de ±40 %. Cependant, la grille varie progressivement, d'au plus 4 % entre deux mailles adjacentes.
Les divers algorithmes apparaissent assez robustes. Le plus sensible est le schéma de Rasch, pour lequel la perte de masse aux petits pas de temps est doublée. Comme on pouvait s'y attendre, **les résultats obtenus avec une distribution initiale plus large**, donc mieux définie et aux gradients moins importants, **sont nettement meilleurs.** On note en particulier que **les solutions obtenues par HOLLY, BOTT3 et BOTT4 sont bien plus proches qu'auparavant.**

L'influence de l'instationnarité des vitesses est abordée dans l'exercice suivant, à savoir le transport d'une gaussienne par un champ de vitesses sinusoidal (section 6.3.3), ce pendant une période, et sur une grille uniforme. **Le classement relatif des algorithmes les uns par rapport aux autres est le même que dans le cas stationnaire. La dégradation des performances est minime.**

En conclusion, les tests pratiqués avec vitesse ou pas d'espace variables sont rassurants quant à leur robustesse des algorithmes. Ils n'apportent pas d'informations inédites sur le comportement de ceux-ci. Cependant, ils nous ont "forcé" à programmer les schémas sous leur forme complète, sans les simplifications possibles dans le cas de vitesse et maillage uniforme, lesquelles sont plus ou moins étendues selon les méthodes. Nous pourrons ainsi (voir plus bas) comparer plus objectivement les coûts informatiques de chacun.

Advection et diffusion de distributions gaussiennes Jusqu'ici, nous avons traité uniquement d'advection pure. En pratique, il existe toujours une certaine dispersion, induite majoritairement par la convection différentielle (cf 2.3). La dispersion physique a tendance à étaler les sources de pollution, à atténuer leurs extrema et leurs gradients : ce faisant, le calcul de leur transport devient moins délicat. **Nous allons donc essayer de jauger la sensibilité des schémas au "rapport de force" entre advection et dispersion (section 6.4).** On reprend le même canevas qu'en 6.3.1. Le champ de vitesse est stationnaire et uniforme, comme

le pas d'espace. Les distributions gaussiennes considérées sont définies respectivement par 8 et 12 mailles. **Le rapport advection/dispersion est mesuré à l'aide du nombre de Peclet** $Pe = U\Delta x/\Gamma$, Γ étant la diffusivité. Nous étudions la répartition des erreurs dans le plan de coordonnées c_r et Pe, au moyen du tracé d'isovaleurs des mesures d'erreur. Ces isocontours nous permettent de visualiser le **domaine d'applicabilité** des schémas, que nous définirons comme suit : des concentrations négatives dont l'amplitude est inférieure à 5 % du maximum de la distribution, un amortissement de celui-ci d'au plus 15 %. Nous avons retenu ces deux indicateurs d'erreur car la première série de tests nous a montré que ces mesures sont les plus "parlantes" quant au comportement d'un schéma. On le voit, nous nous autorisons une erreur assez large sur le pic et sommes plus sévères quant aux valeurs négatives. C'est compréhensible parce qu'en pratique une incertitude de 10 % sur la valeur des extrema est courante (c'est même un niveau de précision tout à fait honorable !). Par contre, si on a des valeurs négatives, les conséquences peuvent être désastreuses, avec des concentrations calculées qui deviennent totalement fantaisistes, notamment si le modèle de transport est couplé à un module biogéochimique.
Commençons par examiner ce qui se passe quand tous les schémas sont utilisables (i.e. pour $c_r \leq 1$), dans le cas de la gaussienne la plus étroite.

1. On peut considérer qu'**il n'y a pas de restriction à l'utilisation de HOLLY, BOTT4 et RASCH**, même si le second schéma amortit un peu trop les pics pour des situations trés advectives ($Pe \geq 500$) et si le troisième les surestime (de 5 à 10 %) pour les situations intermédiaires ($Pe \simeq 20$).

2. **Le domaine d'applicabilité de BOTT3 est également étendu : en gros, il correspond à** $Pe \leq 100$.

3. **QUICKEST et LI sont applicables dans un même cadre, nettement plus restreint** : $Pe \leq 15 - 20$ quand $c_r \leq 0.5$, $Pe \leq 30$ pour c_r voisin de 0.5, sans limitation pour c_r supérieur à 0.8

4. **Le schéma le plus limité est DAN** : $Pe \leq 10$, ce en raison de sa propension à engendrer des concentrations négatives trop fortes.

Le comportement des schémas vis à vis de la valeur du nombre de Courant est le suivant :

- Tant que c_r est plus faible que 1, quelque soit le schéma, ses erreurs sont une fonction décroissante du nombre de Courant, à Peclet donné.

- La situation est plus complexe pour les schémas aux caractéristiques quand $c_r \geq 1$. L'erreur est fonction à la fois de la partie entière et de la partie décimale de c_r. Pour une même partie décimale, les erreurs sont une fonction décroissante de la partie entière. Pour une même partie entière, les erreurs sont maximales pour une partie décimale de 0.5, décroissent au fur et à mesure que c_r se rapproche d'une valeur entière.

Bien entendu, les domaines d'applicabilité des schémas sont plus vastes dans le cas d'une distribution initiale plus large.

Nous ne disposons que d'un test particulier et ses résultats ne doivent pas être extrapolés sans précaution à d'autres problèmes. Cependant, **ce test nous permet d'avoir un idée de la taille relative des domaines d'applicabilité des différents schémas.** Il confirme les résultats des tests d'advection pure, à savoir que :

- **Les schémas HOLLY, RASCH et BOTT4 sont les plus performants et les plus robustes.**
- **BOTT3 n'est pas loin derrière.**
- **QUICKEST et LI peuvent se réveler utiles pour des situations "intermédiaires" : advection dominante, mais pas "écrasante".**
- **L'utilisation de DAN doit être réservée à des cas de diffusion dominante.**

Interférence des conditions limite Jusqu'ici, nous avons travaillé dans des cas où la source de pollution à transporter était parfaitement définie, sous forme d'une condition initiale. **En pratique, on a souvent à traiter de sources entrant par une frontière du domaine de calcul.** Chacun des algorithmes que nous étudions requiert une certaine adaptation au voisinage des frontières, qui s'accompagne d'une réduction de l'ordre du schéma. Nous examinons en section 6.5 (pour les schémas RASCH, HOLLY, BOTT3 et BOTT4) et en annexe F.3 (pour LI, DAN, QUICKEST) si ceci a des conséquences néfastes.

Le premier jeu de tests concerne **le transport d'un front de concentration, entrant par l'amont (section 6.5.1).** La vitesse est uniforme et stationnaire. On applique 3 valeurs de diffusivité, correspondant respectivement à un cas d'advection pure, d'advection dominante et de diffusion dominante. On travaille successivement avec une grille uniforme et une grille variable. Cette série de tests apparaît globalement moins sévère que les précédents. L'écart de performance entre les différents schémas se réduit (notamment BOTT3 et BOTT4 sont équivalents), hormis en ce qui concerne DAN.

Le second jeu de tests consiste en **l'advection pure d'inputs gaussiens (section 6.5.2).** Ce test permet de conclure que dans ce type de situations, il ne suffit pas d'avoir un bon schéma. Il faut aussi choisir des pas de temps qui permettent de ne pas perdre trop d'information en décrivant l'input. Une fois qu'on a respecté cette condition, **le traitement des conditions limite apparaît satisfaisant pour tous les schémas.**

Temps de calcul Finalement, on s'intéresse au coût informatique, mesuré en terme de temps CPU (section 6.6). Nous avons calculé, sur la base d'une cinquantaine de simulations, le coût relatif de chaque méthode par rapport à celle qui s'est révélée la plus économique, à savoir QUICKEST. **Dans le cas d'un problème univariable, on obtient que :**

1. **Les coûts des méthodes aux caractéristiques (LI, RASCH, HOLLY) sont très proches. Ils sont en gros 5 fois plus lents que QUICKEST.**
2. **BOTT3 est 2.25 fois plus lent que QUICKEST.**
3. Il y a une grande différence entre **la méthode BOTT4 standard et sa version simplifiée (pas d'espace uniforme)**, lesquelles **consomment respectivement 4.5 et 1.75 fois plus de CPU que QUICKEST.**

Pour les méthodes aux caractéristiques, l'étape la plus coûteuse est celle qui consiste à **remonter les caractéristiques** : elle **consomme 75 % du temps CPU.** Cette étape est commune à toutes les variables susceptibles d'être véhiculées par l'écoulement. Ainsi, si l'on travaille avec deux composés dissous au lieu d'un, le coût d'une méthode aux caractéristiques ne doublera pas, comme c'est le cas pour les autres types d'algorithme. Par exemple, dans le cas de RASCH, il augmentera de 20 % uniquement.

On a souvent accusé les méthodes aux caractéristiques d'être très coûteuses. Cette critique semble peu fondée.

- Le coût des méthodes aux caractéristiques performantes (HOLLY, RASCH) et des méthodes conservatives permettant d'atteindre la même qualité de résultats (BOTT4) est finalement assez proche dans un cas d'application général (grille variable).
- Les méthodes aux caractéristiques autorisent plus d'"économies d'échelle" dans les problèmes multivariables (existence d'une "grosse" étape de calcul commune)

Conclusion générale des tests monodimensionnels En résumé, pour les tests pratiqués, qui sont pour la plupart représentatifs de situations où l'advection est dominante, **dans un cadre général, les schémas qui présentent le meilleur rapport "qualité/prix" sont les méthodes aux caractéristiques d'Holly-Preissmann et Rasch-Williamson.** Dans le cas particulier où l'on peut travailler avec un pas d'espace constant, le meilleur rapport, jusqu'à un maximum de 4 variables, est offert par BOTT4. Dans des cas de diffusion modérée ou dominante, quelque soit la discrétisation, une alternative intéressante est d'utiliser le schéma BOTT3.

Chapter 7

Two-dimensional test cases

7.1 Choice of test cases

The application of one-dimensional test cases has already allowed us to outline some classification of the selected schemes. However, it is essential to check how these algorithms extend to multi-dimensional situations. We shall restrict our analysis to the two-dimensional case, as we are only interested in the implementation of cross or depth-averaged surface flow and transport models.

The two-dimensional test cases most frequently applied appear to be the following :

- The transport of concentration-hills in a rotational flow field has been frequently investigated (see for instance (Baptista *et al.*, 1988; Holly & Usseglio-Polatera, 1984; Smolarkiewicz, 1983; Takacs, 1985; Tremback *et al.*, 1987; Bott, 1989a; Dan N'Guyen, 1988; Williamson & Rasch, 1989; Chen & Falconer, 1992)). Various shapes (gauss, cosine and cone-hills) and sizes (base radius from 5 to 15 Δx) of concentration-hills have been used as initial conditions.
- The transport of gauss-hills in shear flow is one of the benchmark tests suggested for the Convection-Diffusion forum (Baptista *et al.*, 1988).
- The convection of a step profile in an oblique uniform flow field has also been studied (eg. (Holly & Usseglio-Polatera, 1984)), notably in order to assess the efficiency of limiters in the reduction of over- and undershoot problems (see (Leonard, 1988; Gaskell & Lau, 1988)).
- Finally, the distortion of concentration-hills in a deformational flow field has also been tackled (eg. (Smolarkiewicz, 1983; Tremback *et al.*, 1987; Bott, 1989a)). However, for this test case, there is no analytical reference solution available so that comments on the schemes performance can be only qualitative.

We discarded the deformational flow field as it involves flows mostly observed in meteorological situations and does not possess any reference solution.

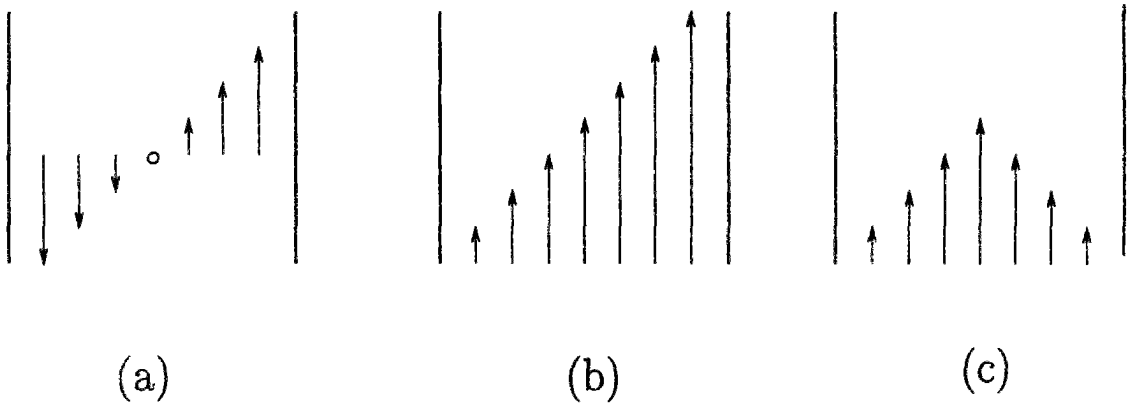


Figure 7.1: Examples of shear flow for 2D advection-dispersion tests

The shear flow test is apparently closer to the kind of flows we can find in riverine applications than the rotational flow. (Figure 7.1 indicates the kind of shear flows for which an analytical solution is available). However, we develop hereafter only the tests relative to a rotational flow field, for the following reasons :

- We judged the rotational flow more interesting because it concerns a truly two-dimensional flow. In the case of the shear flow, the flow is unidirectional. Besides, if one wants to have analytical solutions available, one must assume the flow to be uniform in the longitudinal direction. The rotational flow test is more demanding. Indeed, it involves strong, variable, advection in both directions so that one can fully assess the validity of the two-dimensional extensions of the algorithms.

- We worked some time ago on the shear flow test (cf detailed results in (Simon, 1990a)). It leads to the same relative ranking of the schemes than does the rotational flow test. This latter test allows furthermore to observ things which do not occur with the shear flow (notably the degradation of Holly-Preissmann and Rasch-Williamson algorithms for respectively too large and too small a time step, cf section 7.3.1) For the sake of brevity, we avoided to repeat twice the same conclusions !

- Last, but not least, visualizations of the results were nicer !

The applied sets of parameters, boundary and initial conditions are derived from those suggested in (Baptista *et al.*, 1988) and (Holly & Usseglio-Polatera, 1984) (the latter includes notably the case of anisotropic diffusion). Considering the outcome of the one-dimensional tests, we

have been studying only five schemes : the most successful schemes (namely HOLLY, RASCH, BOTT3 and BOTT4) and the simplest and "cheapest" one (QUICKEST).

7.2 Test presentation

The mathematical problem is governed, in cartesian coordinates (x, y), by equation :

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} + V\frac{\partial C}{\partial y} = \Gamma_{xx}\frac{\partial^2 C}{\partial x^2} + \Gamma_{xy}\frac{\partial^2 C}{\partial x \partial y} + \Gamma_{yx}\frac{\partial^2 C}{\partial y \partial x} + \Gamma_{yy}\frac{\partial^2 C}{\partial y^2} \tag{7.1}$$

with boundary conditions

$$C(x, y, t) \rightarrow 0 \quad \text{as} \quad x^2 + y^2 \rightarrow \infty \tag{7.2}$$

where Γ_{xx}, Γ_{yy}, Γ_{xy}, Γ_{yx} denote the components of the diffusion tensor and U, V are the components of velocity vector $\vec{U}$ along the $x-$ and $y-$ directions respectively.

1. **Flow conditions**
   The flow describes a counterclockwise rotation :

   $$U = -\omega y \quad \text{and} \quad V = \omega x$$

   ω is the angular frequency of rotation. Let $\mathcal{P}$ be the period of rotation : $\omega = 2\pi/\mathcal{P}$. Two sets of tests have been performed, the first (tests A) with $\mathcal{P} = 3000$s, the second (tests B) with $\mathcal{P} = 12000$s.

   Let $\Gamma_{\parallel}$ and $\Gamma_{\perp}$ be respectively the diffusivities in the streamline direction and perpendicular to it. At point M, Γ_{xx}, Γ_{yy}, Γ_{xy} and Γ_{yx} are linked to $\Gamma_{\parallel}$ and $\Gamma_{\perp}$ by the following relations :

   $$\begin{aligned} \Gamma_{xx} &= \Gamma_{\parallel}\cos^2\varphi + \Gamma_{\perp}\sin^2\varphi \\ \Gamma_{yy} &= \Gamma_{\parallel}\sin^2\varphi + \Gamma_{\perp}\cos^2\varphi \\ \Gamma_{xy} &= \Gamma_{yx} = \left(\Gamma_{\parallel} - \Gamma_{\perp}\right)\cos\varphi\sin\varphi \end{aligned}$$

   φ being the angle between the flow velocity at M and the $x-$axis. For the rotational flow field, we have at point M(x, y) with corresponding polar coordinates (r, θ), $\varphi = \theta + \pi/2$.

   The first set of problems (tests A) include only pure advection tests. In the second set (tests B) we apply anistropic diffusion : $\Gamma_{\parallel} = 0.5363$ m^2.s^{-1} and $\Gamma_{\perp} = 0.2681$ m^2.s^{-1}.

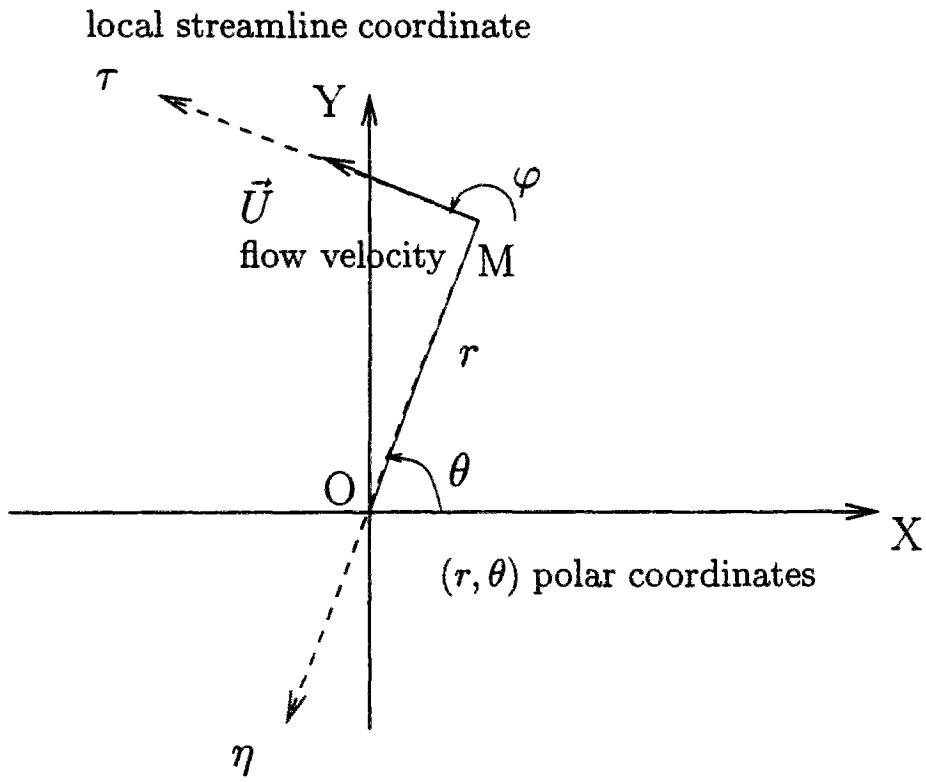


Figure 7.2: Rotational flow field : Notations

2. **Computational domains and their discretization**

For tests A, computations were performed on 35 × 35 points grids defined as follows :

- Grid 3 ($x, y \in [-3400, +3400]$)

$$\begin{aligned} x(i,j) &= 200(i-1) - 3400 \quad i = 1, 35 \\ y(i,j) &= 200(j-1) - 3400 \quad j = 1, 35 \end{aligned}$$

- Grid 4 (with $x(21,j) = 0$ and $y(i,21) = 0$)

$$\begin{aligned} x(i+1,j) - x(i,j) &= 200 - 50 \cos\frac{\pi(i-1)}{35} \quad i = 1, 35 \\ y(i,j+1) - y(i,j) &= 200 - 50 \cos\frac{\pi(j-1)}{35} \quad j = 1, 35 \end{aligned}$$

For tests B, computations were performed on a 27 × 27 points grid :

- Grid 5 ($x, y \in [-1300, +1300]$)

$$\begin{aligned} x(i,j) &= 100(i-1) - 1300 \quad i = 1, 27 \\ y(i,j) &= 100(j-1) - 1300 \quad j = 1, 27 \end{aligned}$$

3. **Initial conditions**

Two different kinds of initial concentration field C_0 are considered. Let G_0 with cartesian coordinates (x_0, y_0) and polar coordinates (r_0, θ_0) be the centre of mass of the initial distribution. Let $\mathcal{R}_0$ be the moving coordinate system, whose origin is G_0 and whose coordinate axes are respectively parallel and perpendicular to the local streamline direction at G_0. Let (τ, η) be the new coordinates associated to $\mathcal{R}_0$.

The initial distributions are given by :

$$\textit{gauss-hill} \quad C_0(\tau,\eta) = \exp\left(-\frac{\tau^2}{2\sigma_{0_\parallel}^2} - \frac{\eta^2}{2\sigma_{0_\perp}^2}\right) \tag{7.3}$$

$$\textit{cone-hill} \quad C_0(\tau,\eta) = \begin{cases} 1 - \sqrt{\dfrac{\tau^2+\eta^2}{l_0^2}} & \text{if } \tau^2+\eta^2 \leq l_0^2 \\ 0 & \text{otherwise} \end{cases} \tag{7.4}$$

where,

- $\sigma_{0_\parallel}$ and $\sigma_{0_\perp}$ are the initial standard deviations, respectively in the streamline direction and perpendicular to it (gauss-hill);
- l_0 is the radius of the initial scalar field (cone-hill).

For tests A, initial conditions were either a gauss-hill with $\sigma_{0_\perp} = \sigma_{0_\parallel} = 264$m either a cone-hill with radius $l_0 = 800$m. When constant grid 3 was used, (x_0, y_0) was set to $(0, -1800)$. When variable grid 4 was used, G_0 was coinciding with node ($x(21)$, $y(11)$). Let $L_{0_\parallel}$ and $L_{0_\perp}$ denote the source dimensionless lengths respectively in the streamline direction and perpendicular to it. On grid 3, both the gauss and cone sources satisfy $L_{0_\parallel} = L_{0_\perp} = 8$.

For tests B, the initial condition was a gauss-hill with $\sigma_{0_\parallel} = 141.43$m and $\sigma_{0_\perp} = 100$m. Thus, on grid 5 $L_{0_\parallel} = 8.5$ and $L_{0_\perp} = 6$. G_0 was located at $x_0 = 0$ and $y_0 = 600$m.

4. **Reference solutions**

At time t, the exact centre of mass G_t is defined by $r^{ex}(t) = r_0$ and $\theta^{ex}(t) = \theta_0 + \omega t$. Let $\mathcal{R}_t$ be the moving coordinate system associated with G_t (see definition in previous paragraph).

For the case of pure advection (tests A), the exact scalar field at time t is described in $\mathcal{R}_t$ by relations 7.3 and 7.4 for the gauss-hill and cone-hill cases respectively.

Let us consider a case with diffusion (tests B). There is no exact solution readily available for the cone-hill test. For the gauss-hill test, should the diffusivities $\Gamma_\parallel$ and $\Gamma_\perp$ be constant over the flow field, the exact solution at time t is given in coordinate system $\mathcal{R}_t$ by :

$$
\begin{aligned}
C(\tau,\eta,t) &= C_{\max}\exp\left(-\frac{\tau^2}{2\sigma_{\parallel}^2}-\frac{\eta^2}{2\sigma_{\perp}^2}\right) \qquad (7.5)\\
\text{where } \sigma_{\parallel}^2 &= \sigma_{0_{\parallel}}^2+2\Gamma_{\parallel}t\\
\sigma_{\perp}^2 &= \sigma_{0_{\perp}}^2+2\Gamma_{\perp}t\\
C_{\max} &= \left(\sigma_{0_{\parallel}}.\sigma_{0_{\perp}}\right)/\left(\sigma_{\parallel}.\sigma_{\perp}\right)
\end{aligned}
$$

In both sets of tests, error measures (specified in tables 7.1 & 7.2, extracted from (Baptista *et al.*, 1988)) are computed after one revolution. In tests B, the exact distribution after a full period rotation has now dimensionless lengths $L_{0_{\parallel}} = 11$ and $L_{0_{\perp}} = 7.7$ and a peak concentration equal to 0.6084 .

Table 7.1: Error measures for 2D convection-diffusion tests (part 1)

Sym-bol	Description	Definition	Comments & exact value	
Φ	L2 error norm normalized by total mass	$\Phi(t) = \dfrac{\left(\int_\Omega \left[C^{\mathrm{nu}}(r,\theta,t) - C^{ex}(r,\theta,t)\right]^2 r\,dr\,d\theta\right)^{1/2}}{m(t)}$	Integral meas. overall error of numerical solution	0
ϵ	Relative error on peak concentration	$\epsilon(t) = \dfrac{C^{ex}_{\max}(t) - C^{\mathrm{nu}}_{\max}(t)}{C^{ex}_{\max}(t)}$	Point measure of numerical damping	0
Ψ	Absolute value maximal neg. concentration, normalized by ex. peak val.	$\Psi(t) = \left\lvert \dfrac{C^{\mathrm{nu}}_{\max,\mathrm{neg}}(t)}{C^{ex}_{\max}(t)} \right\rvert$	Point measure wiggles in numerical solution	0
μ_0	0^{th}-moment of conc. field, normalized by exact value	$\mu_0(t) = \dfrac{1}{m(t)}\int_\Omega C^{\mathrm{nu}}(r,\theta,t)\,r\,dr\,d\theta$	Integral meas. of mass preservation	1

(nb: superscripts "nu" and "ex" refer respectively to the numerical and reference solution)

Table 7.2: Error measures for 2D convection-diffusion tests (part 2)

Sym-bol	Description	Definition	Comments & exact value	
ξ_r	Normalized errors in peak position	$\xi_r(t) = \dfrac{r_0 - r^{\mathrm{nu}}_{\max}(t)}{r_0}$	phase shift introduced by numerical solution	0
ξ_θ		$\xi_\theta(t) = \dfrac{\theta^{ex}_{\max}(t) - \theta^{\mathrm{nu}}_{\max}(t)}{2\pi t/\mathcal{P}}$		0
μ_r	Normalized errors in centre of mass position	$\mu_r(t) = \dfrac{E^{ex}_r(t) - E^{\mathrm{nu}}_r(t)}{r_0}$	Integral meas. phase shift introduced by numerical solution	0
μ_θ		$\mu_\theta(t) = \dfrac{E^{ex}_\theta(t) - E^{\mathrm{nu}}_\theta(t)}{2\pi t/\mathcal{P}}$ with $E_r(t) = \dfrac{1}{m(t)} \int_\Omega r C^{\mathrm{nu}}(r,\theta,t)\, r\, dr\, d\theta$ $E_\theta(t) = \dfrac{1}{m(t)} \int_\Omega \theta C^{\mathrm{nu}}(r,\theta,t)\, r\, dr\, d\theta$		0
μ_{rr}	Centred 2^{th}-moments of conc. field, normalized by exact value	$\mu_{rr}(t) = \dfrac{\int_\Omega [\, r - E^{\mathrm{nu}}_r(t)\,]^2\, C^{\mathrm{nu}}(r,\theta,t)\, r\, dr\, d\theta}{\int_\Omega [\, r - E^{ex}_r(t)\,]^2\, C^{ex}(r,\theta,t)\, r\, dr\, d\theta}$	Integral meas. of numerical spreading	1
$\mu_{\theta\theta}$		$\mu_{\theta\theta}(t) = \dfrac{\int_\Omega [\, \theta - E^{\mathrm{nu}}_\theta(t)\,]^2\, C^{\mathrm{nu}}(r,\theta,t)\, r\, dr\, d\theta}{\int_\Omega [\, \theta - E^{ex}_\theta(t)\,]^2\, C^{ex}(r,\theta,t)\, r\, dr\, d\theta}$		1

5. **Temporal discretization**
Finite difference methods such as QUICKEST, BOTT3 and BOTT4 have a stability condition, namely

$$\max(c_{r,x}, c_{r,y}) \leq 1 \tag{7.6}$$

$c_{r,x}$ and $c_{r,y}$ being the local Courant numbers in the $x-$ and $y-$ directions respectively. For tests A, condition 7.6 is enforced all over the computational domain if and only if the time step Δt is less than 29s (grid 3) or 21s (grid 4) respectively. If we impose 7.6 to be satisfied only on the subdomain where concentrations are not negligible (which corresponds approximately to $\sqrt{x^2+y^2} \leq 2600$m), the stability condition is somewhat relaxed and becomes "Δt less than 36 and 28s respectively". For tests B, 7.6 induces $\Delta t \leq 150$s.

Backwards characteristics methods, on the other hand, have no stability restriction.

Tables 7.3, 7.4 and 7.5 indicate the time steps applied for both families of algorithms in tests A and B respectively. In these tables, we mentioned also the number of iterations required to reach the final computational time and the Courant number at the centre of mass of the transported distribution.

Table 7.3: Rotational flow : Time steps for tests A & FDM methods

	Both grids				Grid 3
Δt (s)	10	20	25	30	37.5
$\mathcal{N}$	300	150	120	100	80
c_r	0.189	0.377	0.471	0.566	0.707

Table 7.4: Rotational flow : Time steps for tests A & Characteristics methods

Δt (s)	10	20	30	50	100	200	500
$\mathcal{N}$	300	150	100	60	30	15	6
c_r	0.189	0.377	0.566	0.943	1.885	3.770	9.425

Table 7.5: Rotational flow : Time steps for tests B

	All methods			Characteristic methods			
Δt (s)	100	120	150	200	300	600	1000
$\mathcal{N}$	120	100	80	60	40	20	12
c_r	0.314	0.377	0.471	0.628	0.943	1.885	3.142

7.3 Discussion of test results

7.3.1 Pure advection of concentration-hills : Tests A

The error measures introduced in tables 7.1 and 7.2 are displayed in figures 7.3 to 7.6 as regards the gauss-hill and figures 7.7 to 7.10 for the cone-hill, apart from the errors relative to the peak location. Indeed, all schemes forecast correctly the peak position except RASCH in two cases (gauss-hill transport with the smallest time step, i.e. 10s, both on regular and variable grid) : in such cases, the error amounts to the width of one mesh cell.

Errors generated by Holly-Preissmann scheme are plotted with solid lines, errors of Rasch-Williamson scheme appear in dashed lines. Errors introduced by the flux-form methods are indicated in dotted lines : square and triangular marks denote respectively the QUICKEST and BOTT3 methods errors, while the BOTT4 errors are plotted without marks.

Some computed distributions are also displayed :

- Figures 7.11 to 7.13 allow comparison of the outcome of the different schemes as regards the gauss-hill transport on uniform grid for time step $\Delta t = 30$s, which is the largest allowable time step for the flux-form methods.
- Figures 7.14 to 7.16 illustrate the evolution of Holly-Preissmann and Rasch-Williamson solution for different time steps, namely $\Delta t = 10$ and 100s (as requested in the convection-diffusion forum (Baptista *et al.*, 1988)) and $\Delta t = 500$s.
- Figures 7.17 to 7.19 present the solution of the cone-hill transport on uniform grid for $\Delta t = 30$s.

1. **L2 error-norm (Φ)**
   As can be checked on figures 7.3 and 7.7, the levels of Φ are rather insensitive to the choice of the spatial discretization.

The cone-hill source is somewhat wider and thus better defined than the gauss-hill distribution. This results in a lower level of errors.

The behaviour of the L2 error norm with respect to the total number of iterations varies according to the related scheme :

- Φ is fairly constant for the flux-form methods. It decreases but slightly as the number of iterations is reduced. QUICKEST displays the highest level of error, approximately twice the error raised by BOTT3. As previously, BOTT4 turns out to be the best of these three schemes.
- The RASCH error decreases steadily with the number of iterations. An important degradation of the scheme performance for the smallest time step (corresponding to 300 iterations) is noticeable. Indeed, in that case, RASCH generates spurious undershoots in the wake of the gauss-hill, as can be observed on figure 7.14.
- The HOLLY error evolution is not monotonic. First the error diminishes as the number of iterations decreases, then it grows again. As we shall see and discuss later, other error measures behave similarly.

According to this criterion, and considering all four tests (gauss and cone-hills on uniform and variable grids), the schemes rank generally as follows (from the best to the worst) : HOLLY, RASCH, BOTT4, BOTT3, QUICKEST.

2. **Mass preservation** (μ_0)
All schemes appear to be fairly conservative, except that of RASCH (cf fig. 7.3 and 7.7).

The flux form methods exhibit slight errors :

- QUICKEST error is less than 0.3 % for the gauss-hill test and the cone-hill test on uniform grid. It is more important for the cone-hill transport on the variable grid but stays nevertheless inferior to 0.9 % in this case.
- BOTT3 mass loss is always less than 0.04 % and the BOTT4 error is smaller than 0.01 %.

As these algorithms are written in a conservative form, these mass errors are to be ascribed to the treatment of the open boundary conditions.

Although it is not inherently a conservative method, HOLLY scheme behaves near exactly. Its error remains inferior to 0.4 % except on the variable grid for the time step corresponding to the smallest number of iterations : then, the mass overestimation is about 1.5 % .

As for the L2 error measure, μ_0 worsens drastically when the RASCH scheme is applied with a small time step : with 300 iterations, the mass loss reaches 30 % for the cone-hill on the variable grid, 25 % for the gauss-hill on the regular grid. Yet, provided that the iterations are fewer than 100, the mass error keeps within acceptable bounds (less than 2 %).

3. **Spurious undershoots** (Ψ)

The important mass loss of the RASCH method for small time steps can be ascribed to the negative concentrations generated by the scheme (cf solution on figure 7.14). Indeed, Ψ remains less than 2 % of the peak value except for the two cases when RASCH displays a dramatic mass loss : there it reaches 10 and 8 % for the gauss and cone-hill respectively (cf figures 7.4 and 7.8).

The behaviour of the HOLLY scheme is exactly the opposite of RASCH : undershoots are moderate and fairly constant (approximately 1 % and 1.5 % for the gauss- and cone-hill respectively) while the number of iterations is superior to 50 (gauss-hill) or 100 (cone). Then they increase sharply as the iteration number keeps decreasing and reach respectively 4 % (uniform grid) and 8 % (cone-hill on variable grid). As the previous error indicators (and the next related to peak damping and numerical spreading), Ψ highlights a degradation of HOLLY's performance for large time steps, which is obvious in figure 7.16. It is time to ask what is the cause of such trend. Errors cannot be ascribed to inaccurate estimation of the trajectories : in such a case, the same troubles would plague the RASCH scheme, whose solution is steadily improving as the time step increases. Similarly, increasing the time step reduces bicubic interpolation errors, as the interpolations performed are fewer. Consequently, the only plausible explanation is that, at large time steps, the explicit computation (cf E.1.3) of the corrective terms which affect the scalar derivatives transport equations becomes too crude ! This conclusion is further supported in Appendix F.4 where we investigated the sensitivity of Holly-Preismann results with respect to the method applied to estimate the derivative correction terms. While we reckon that improvement of HOLLY for large time steps could probably be obtained when using more clever algorithms than what we have been testing so far, we have not been dwelling any longer on that problem.

Lastly, let us consider the flux-form methods : the positive schemes BOTT3 and BOTT4 obviously do not induce any negative concentrations while the third scheme, namely QUICKEST, raises some moderate undershoots (less than 2 %).

4. **Damping** (ϵ)

This kind of error is rather insensitive with respect to the chosen discretization.

As can be observed on figures 7.4 and 7.8, the scheme which best preserves the peak concentration is undoubtedly RASCH. Notably, results relative to the cone-hill are excellent (cf fig. 7.8), with errors less than 4 % in any case : as in one-dimensional tests, RASCH appears to deal better with distributions whose shape is triangular (compare also figures 7.17 to 7.19).

Next comes Holly-Preismann scheme, for which errors are in the range 5 to 25 % for the gauss-hill case (fig. 7.4), between 10 and 20 % for the cone-hill test (fig. 7.8).

Damping raised by the flux-form methods is fairly independent of the number of iterations. With QUICKEST, it is about 60 % for the gauss-hill test, 50 % for the cone-hill. For BOTT3, it is respectively 40 and 30 %; for BOTT4, 30 and 20 %. These underestimations of the peak concentration are rather unacceptable. Only the characteristics methods (for RASCH in most cases, for HOLLY with intermediate time steps) allow accurate reproduction of the peak value.

5. **Phase shift : centre of mass location** (μ_r & μ_θ)
HOLLY, BOTT3 and BOTT4 forecast accurately the location of the centre of mass of the transported scalar field (cf figures 7.5 and 7.9). QUICKEST performs also well. Yet, it appears to be more sensitive to the non-uniformity of the grid (see the cone-hill test results in figure 7.9).

   As in the one-dimensional cases, RASCH displays more phase shift, notably for small time steps (i.e. a large number of iterations). Yet, as soon as the total number of iterations is inferior to 150, the error levels are acceptable (errors on the radius inferior to 2 and 1 % for the gauss- and cone-hill respectively, errors on the angle less than 0.5 % of the angle described). RASCH performs undoubtedly better with the cone-hill distribution, for which the errors are approximately half the errors observed in the gauss-hill test.

6. **Numerical spreading** (μ_{rr} & $\mu_{\theta\theta}$)
Numerical spreading is evaluated in polar coordinates with the help of two second-order moments ratios, one relative to the radius (μ_{rr}), the other one relative to the angle ($\mu_{\theta\theta}$) (cf table 7.2).

   The less dispersive scheme is quite obviously HOLLY, except for the biggest time step : on uniform grid, $\mu_{\theta\theta}$ worsens clearly for the gauss-hill (fig. 7.6) as for the cone-hill (fig. 7.10).

   RASCH behaves fairly well for the cone-hill test (fig. 7.10) except once again for the smallest time step, on the variable grid : there, both μ_{rr} and $\mu_{\theta\theta}$ values are unacceptable. For this specific combination of spatial and temporal discretization, large negative concentrations occur. RASCH dispersion is similarly too important for a large number of iterations in the gauss-hill case (cf fig. 7.6).

   Let us now consider the flux-form methods. In the gauss-hill case, the best performance is apparently QUICKEST. However, a look at the solution (cf fig. 7.13) indicates that QUICKEST does display numerical spreading. Yet, in the computation of second-order moments, this is somewhat compensated by negative concentrations which, contrary to what occurs for the RASCH scheme, are rather symmetrically located. In fact, the flux-form method which exhibits truly less dispersion is BOTT4, followed by BOTT3 (cf figures 7.12 and 7.13). The schemes rank similarly as regards the cone-hill test (cf fig. 7.10, and more generally distributions displayed on fig. 7.17 to 7.19).

In fact the only schemes which allow us to solve correctly this test case (provided some relevant choice of the time steps) **are the backwards characteristic methods.** The Rasch-Williamson scheme best reproduces the peak value but slightly distorts the gauss-hill distribution. The Holly-Preismann scheme induces more damping but best preserves mass and gauss-hill shape. In the range of allowable time steps which ensure their stability, all three flux-form methods exhibit too much damping.

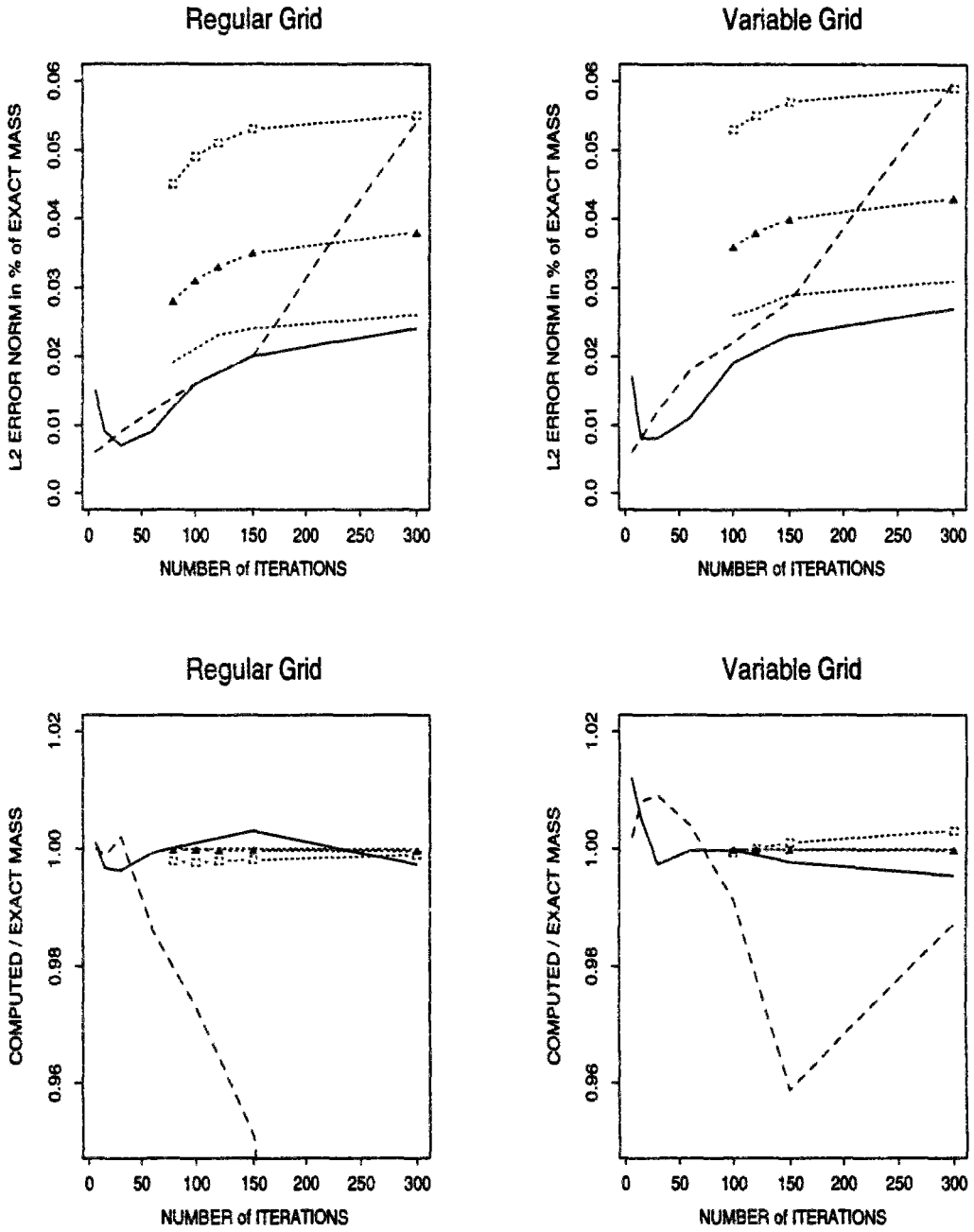


Figure 7.3: Test A / Gauss-hill : L2 error norm & mass preservation after a full revolution

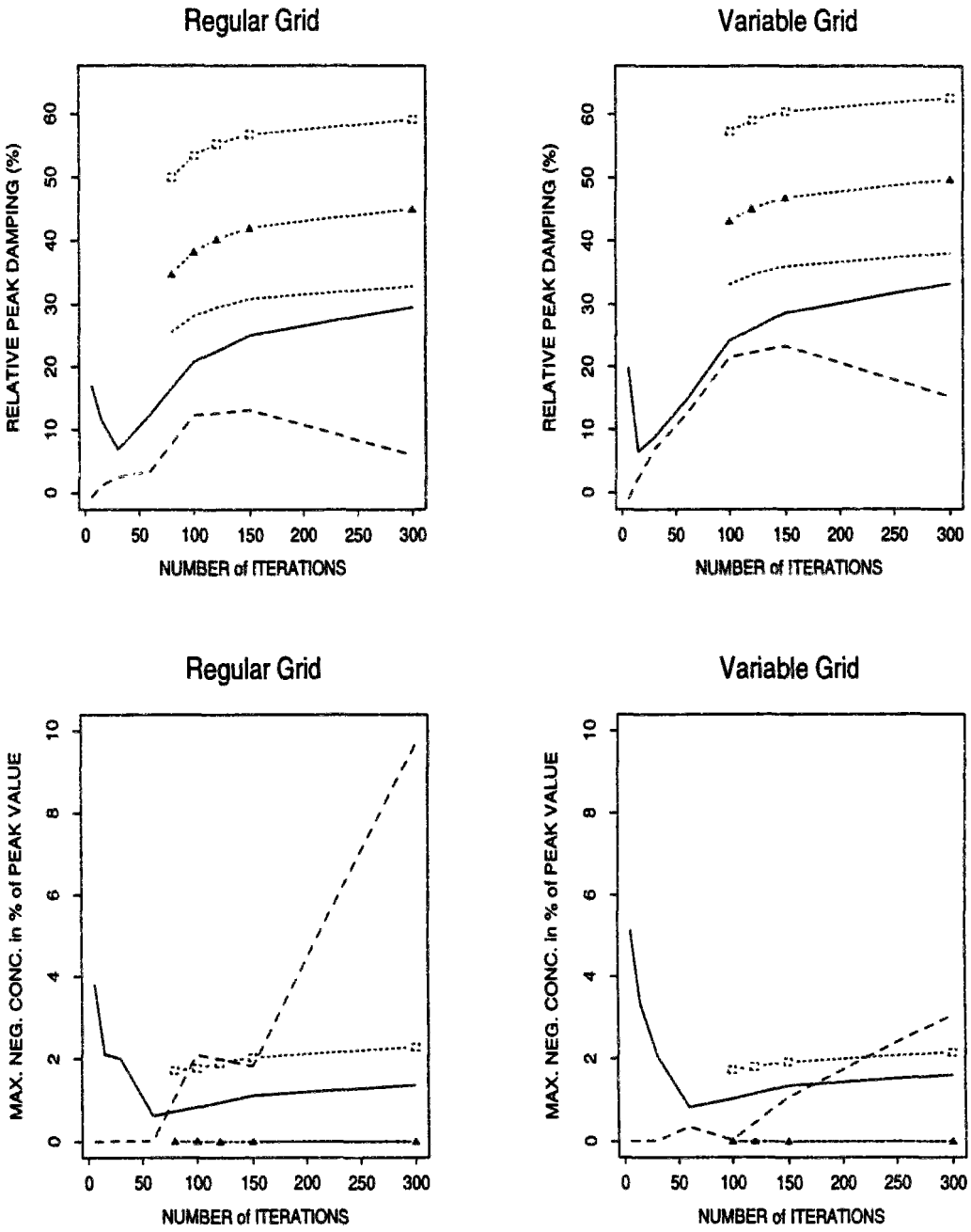


Figure 7.4: Test A / Gauss-hill : Damping & spurious undershoots after a full revolution

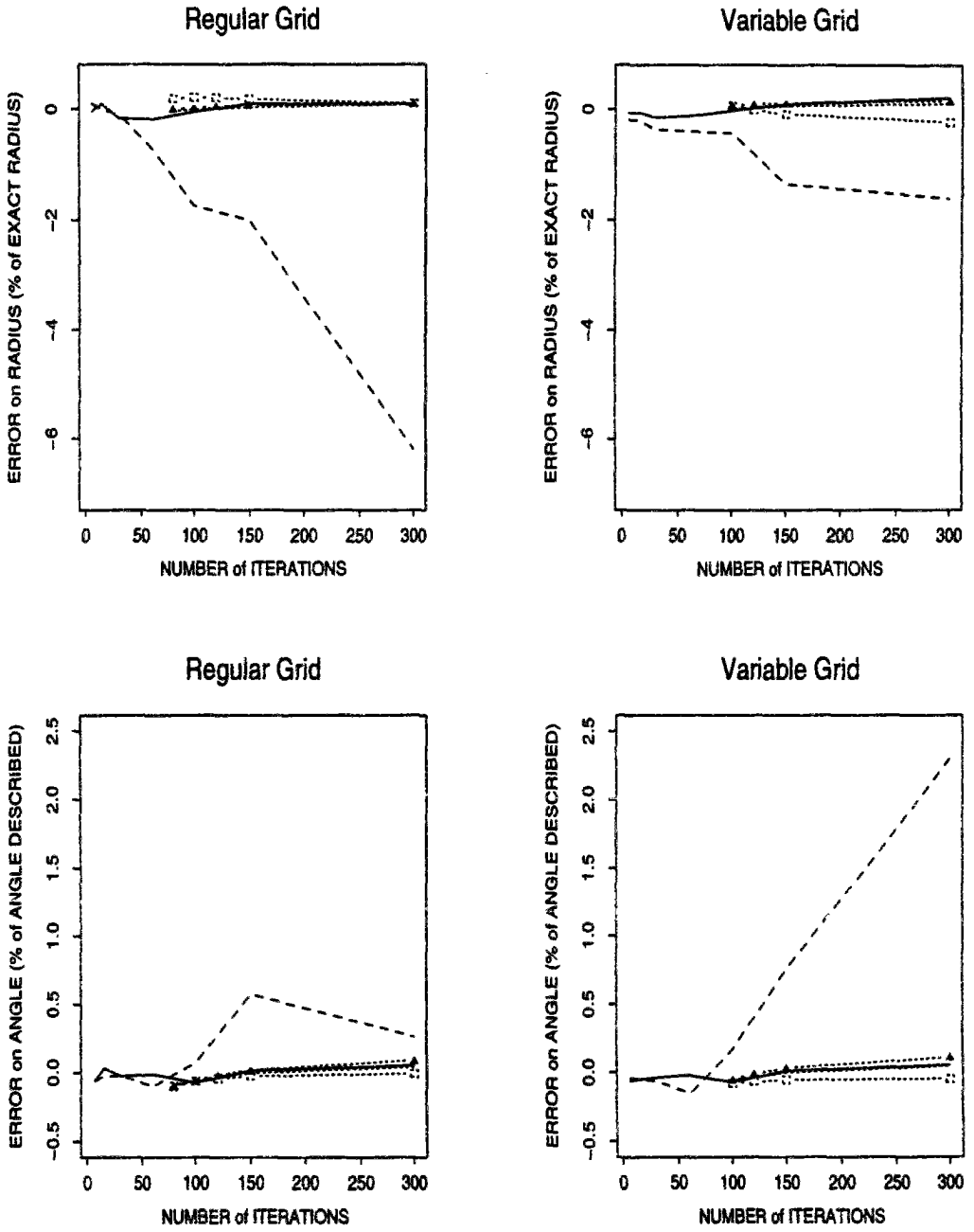


Figure 7.5: Test A / Gauss-hill : Errors on centre of mass location after a full revolution

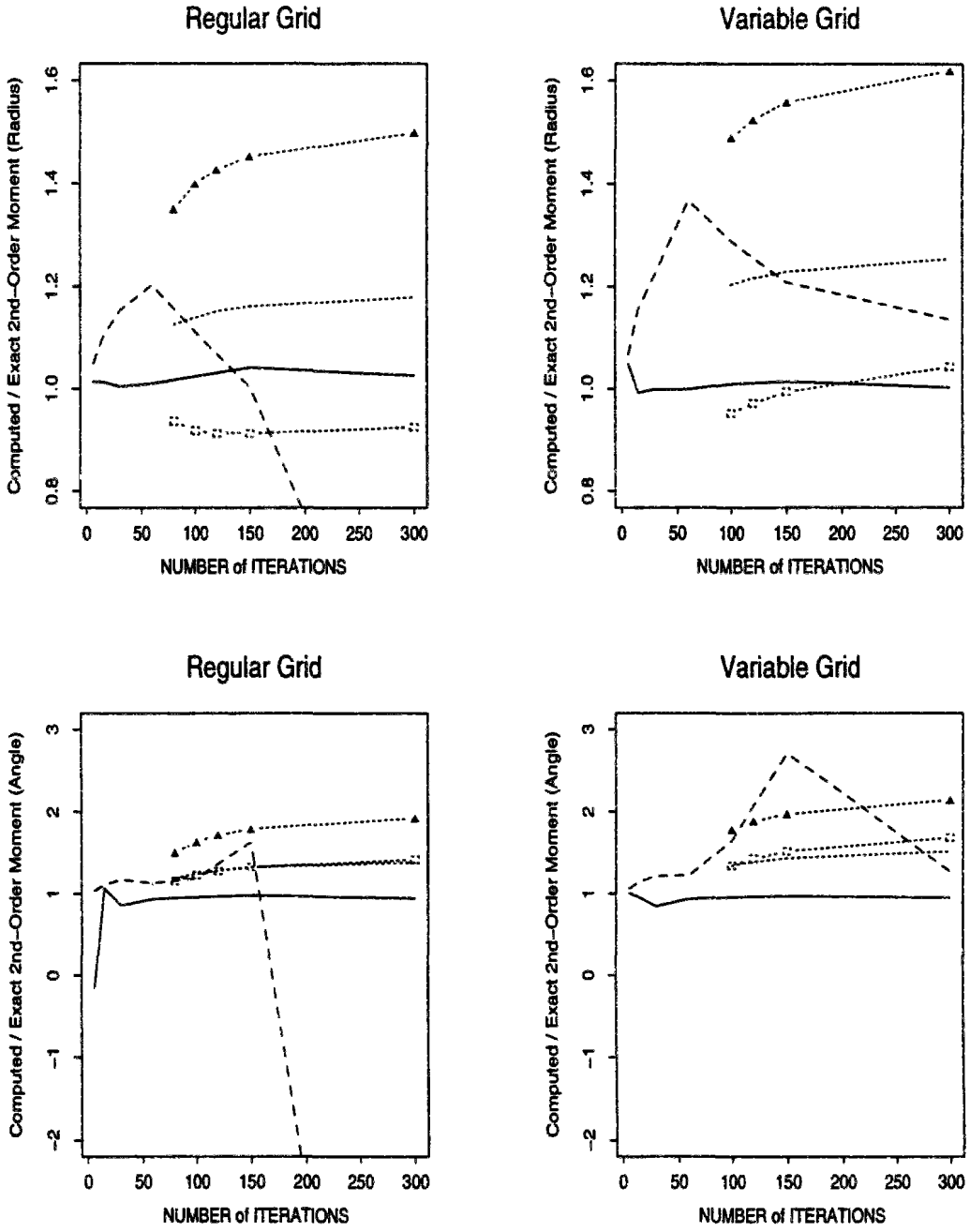


Figure 7.6: Test A / Gauss-hill : Numerical spreading after a full revolution

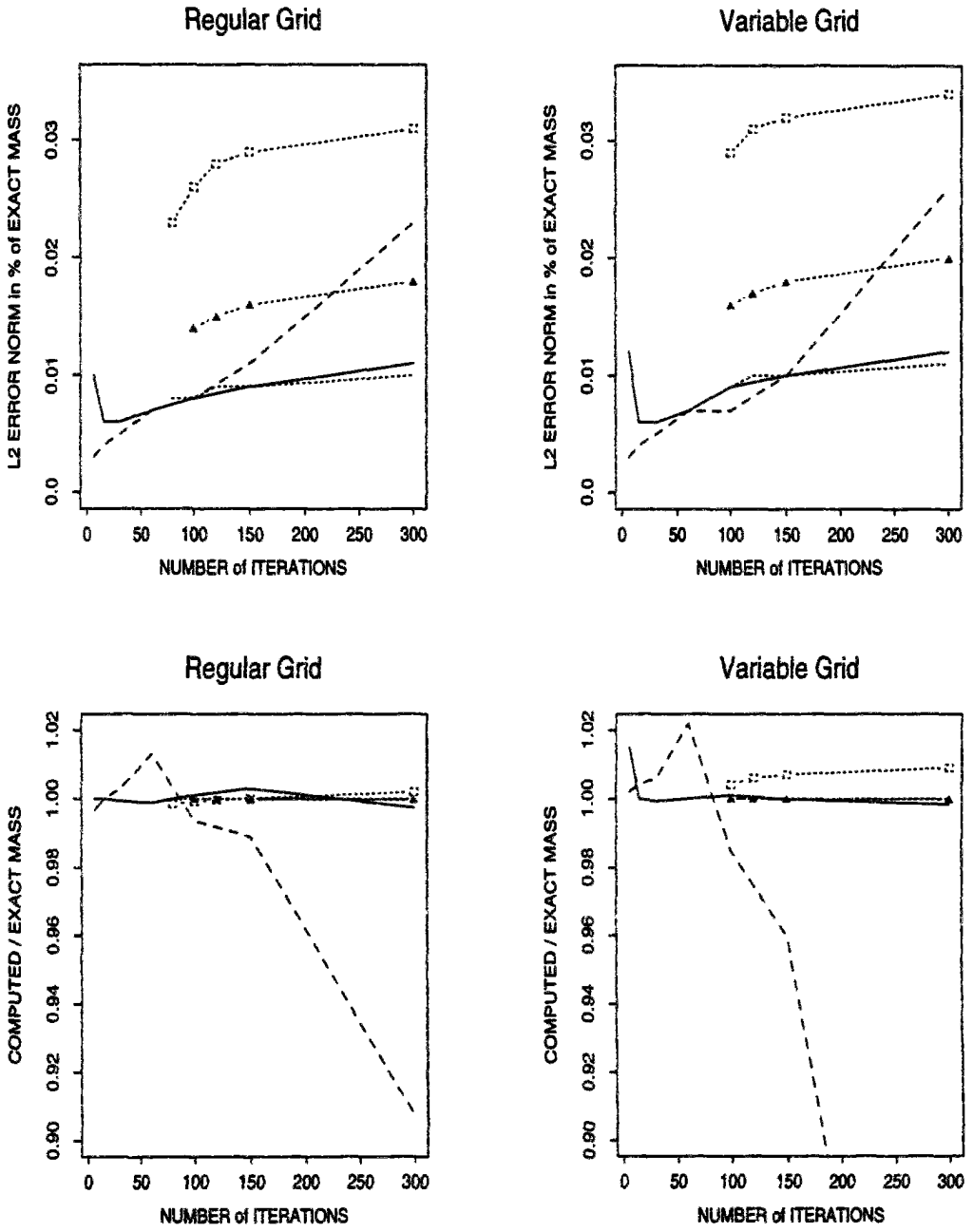


Figure 7.7: Test A / Cone-hill : L2 error norm & mass preservation after a full revolution

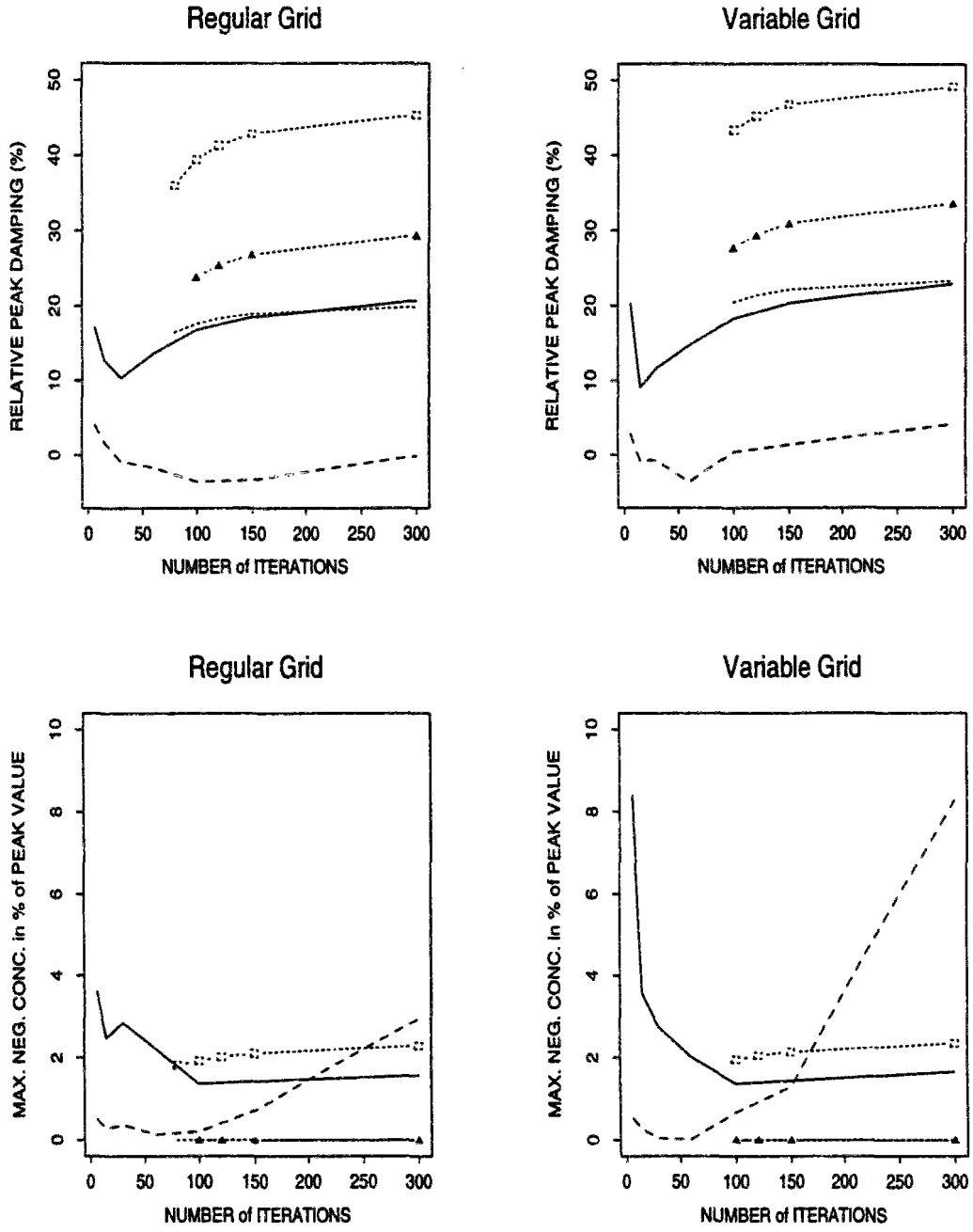


Figure 7.8: Test A / Cone-hill : Damping & spurious undershoots after a full revolution

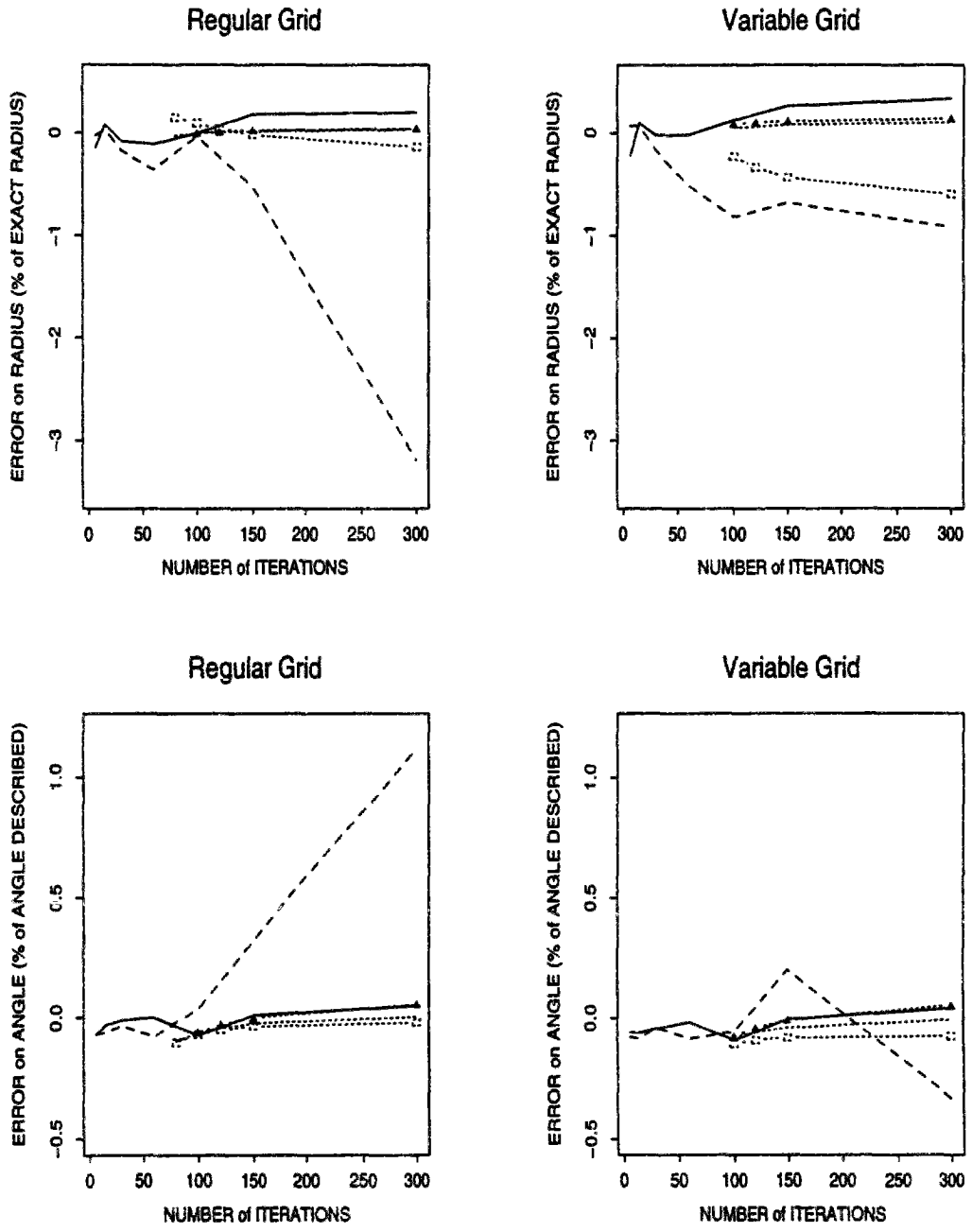


Figure 7.9: Test A / Cone-hill : Errors on centre of mass location after a full revolution

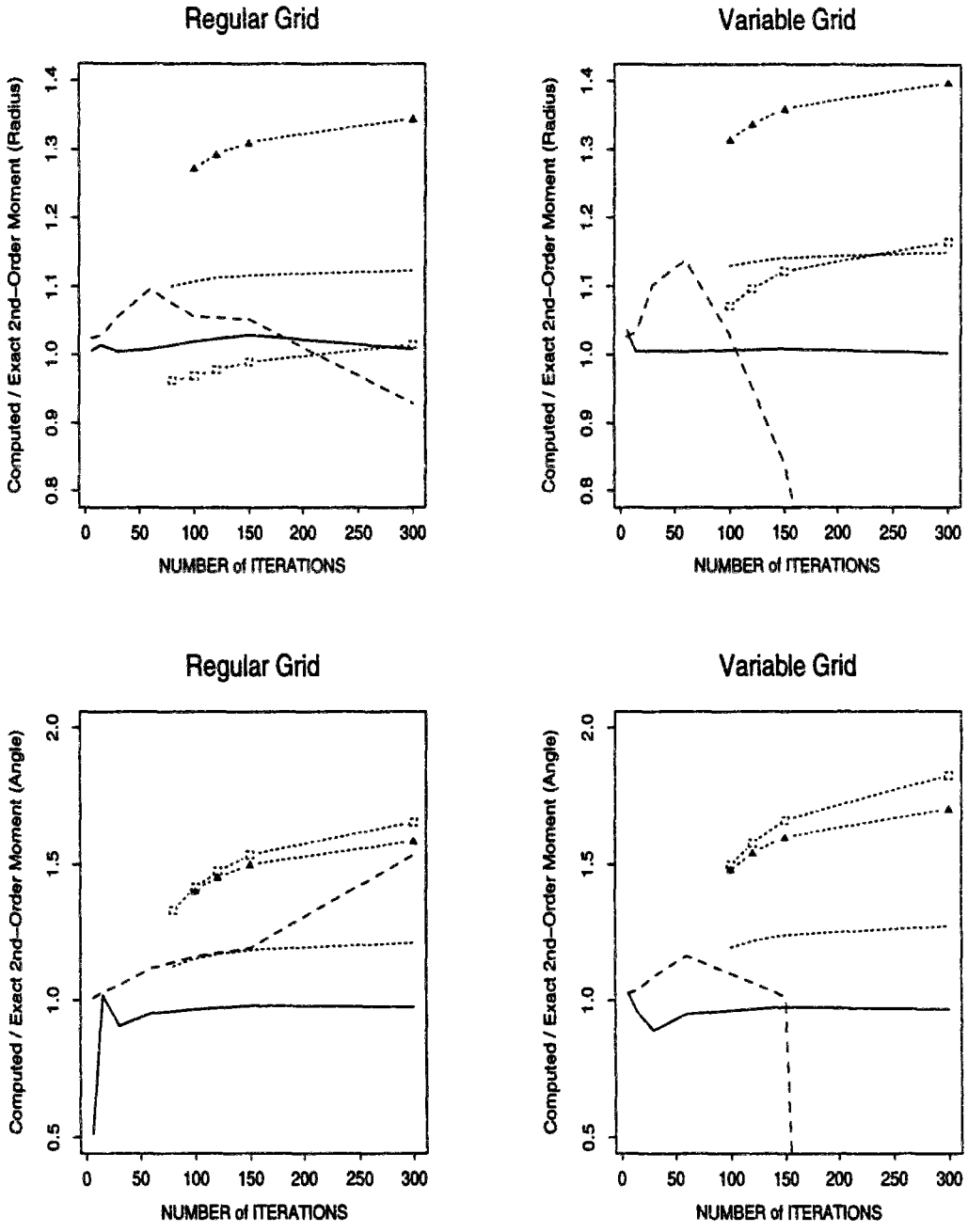


Figure 7.10: Test A / Cone-hill : Numerical spreading after a full revolution

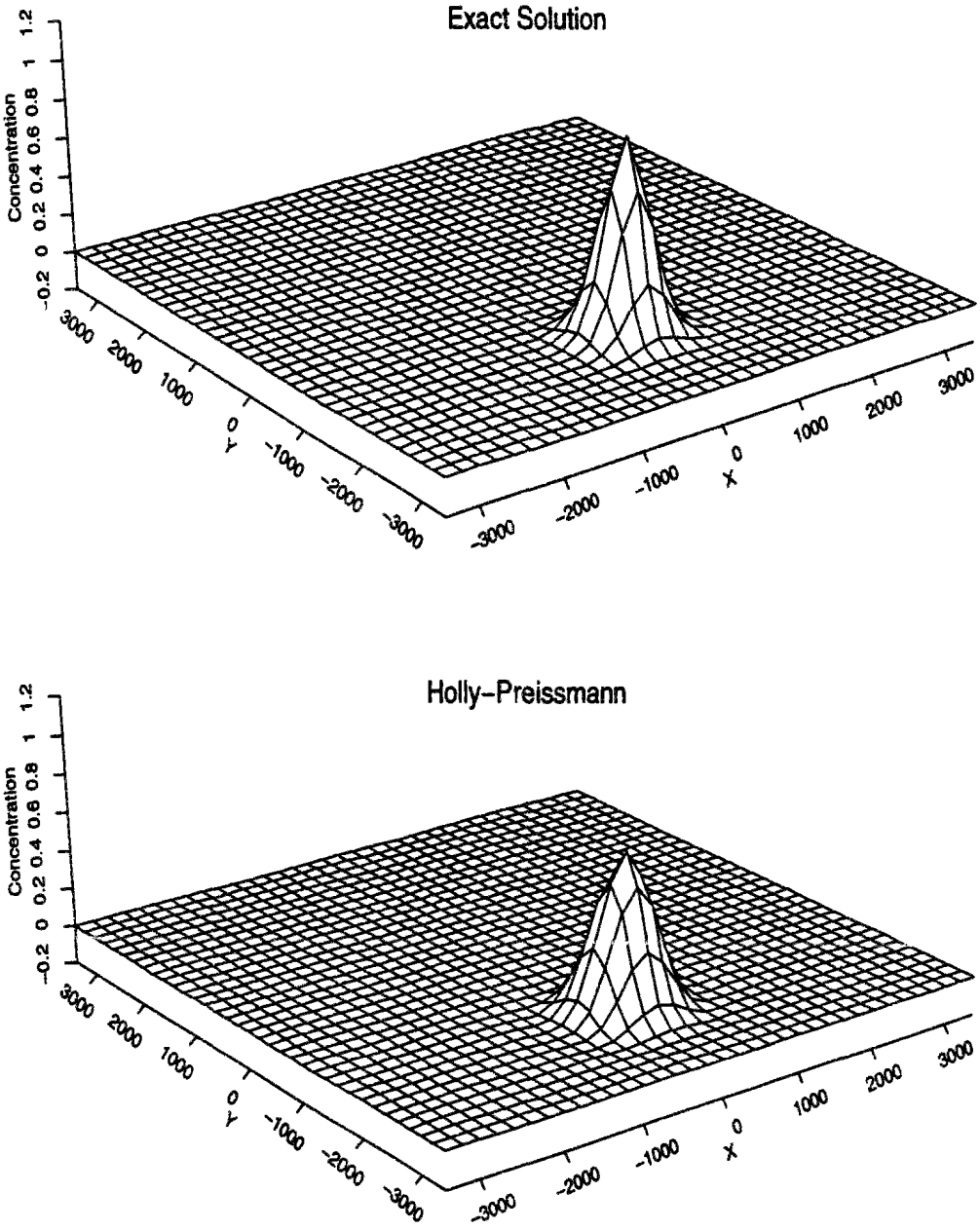


Figure 7.11: Gauss-hill : Exact & HOLLY solutions after a full revolution (Δt = 30s)

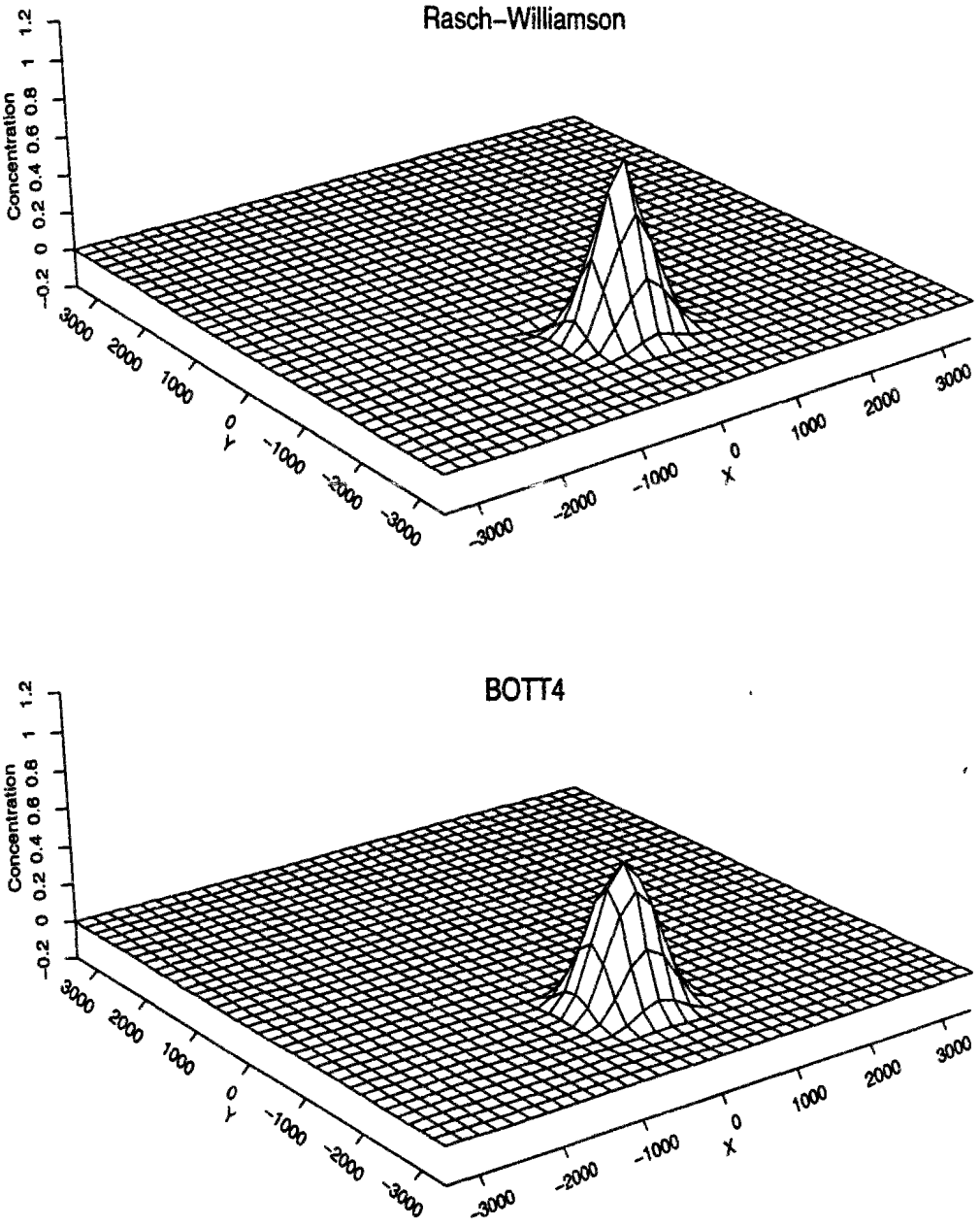


Figure 7.12: Gauss-hill : RASCH & BOTT4 solutions after a full revolution (Δt = 30s)

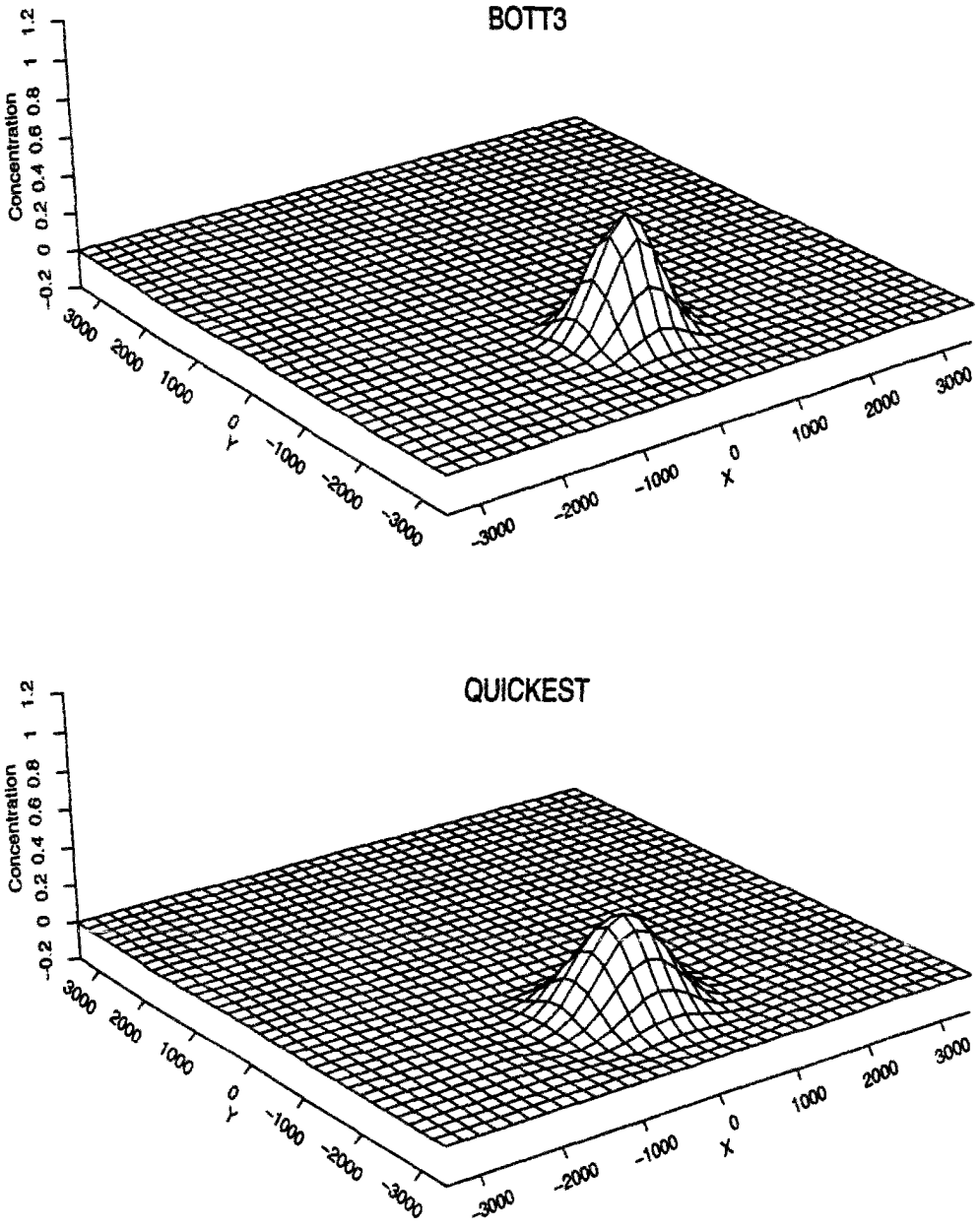


Figure 7.13: Gauss-hill : BOTT3 & QUICKEST solutions after a full revolution (Δt = 30s)

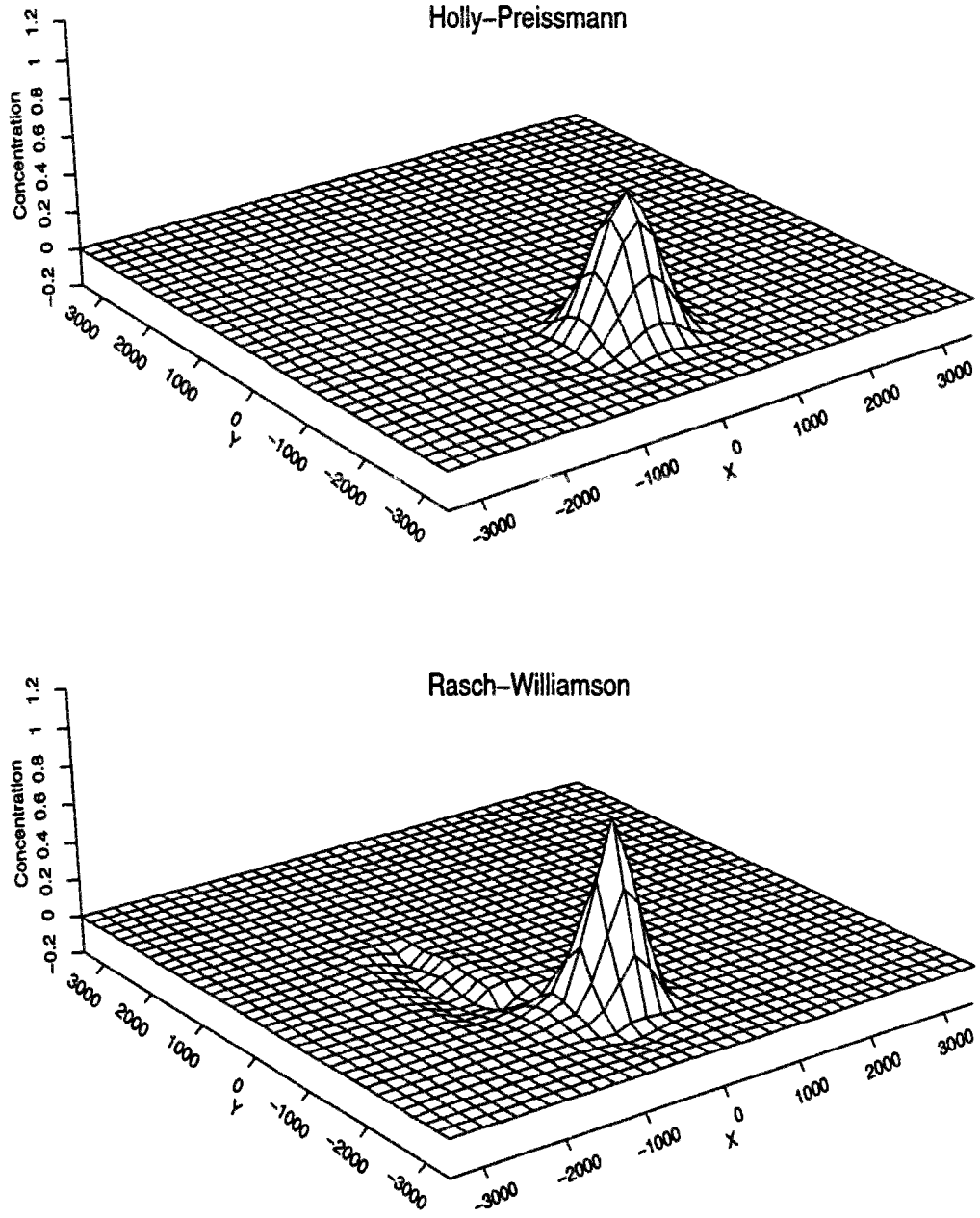


Figure 7.14: Gauss-hill : HOLLY & RASCH solutions for $\Delta t = 10$s

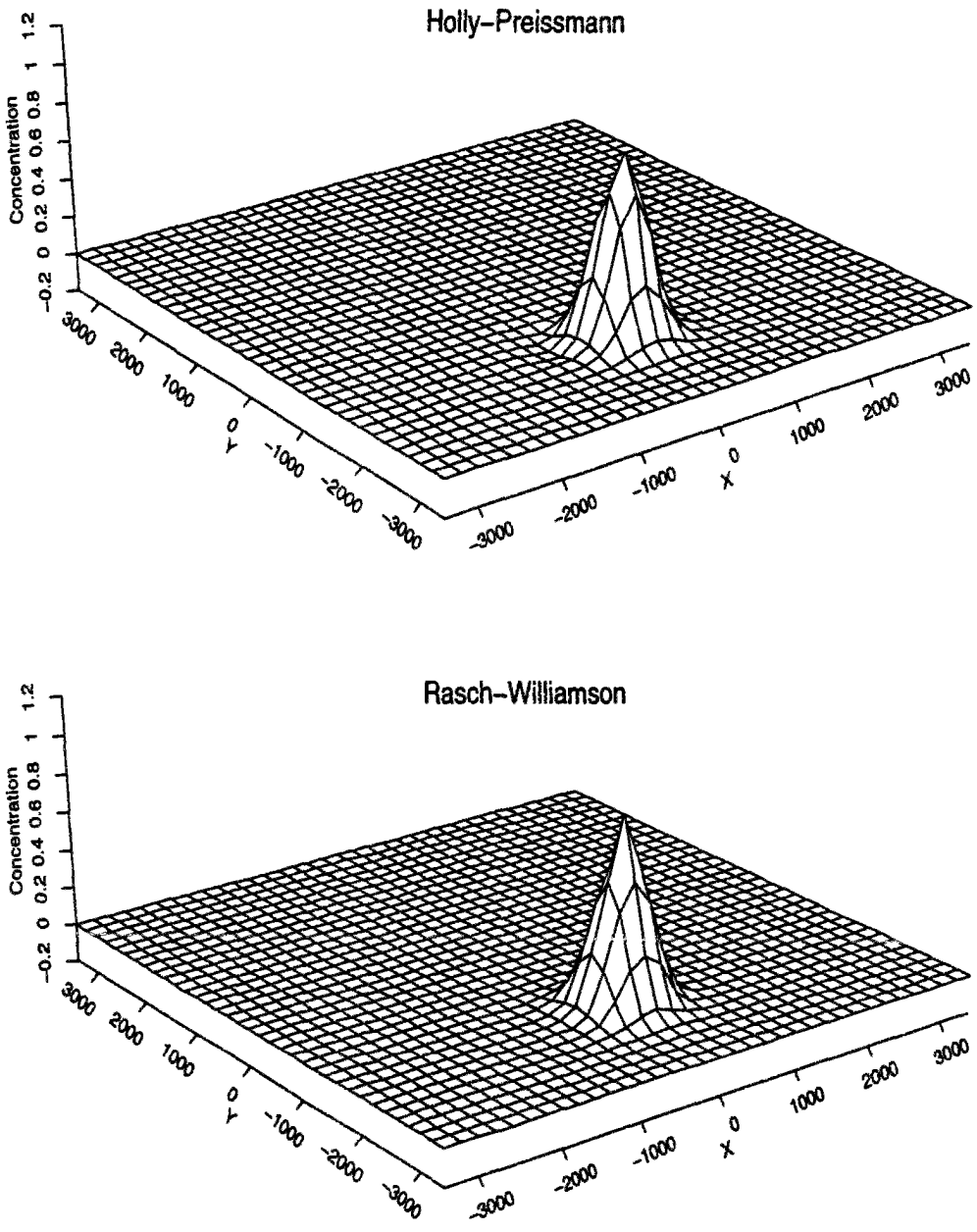


Figure 7.15: Gauss-hill : HOLLY & RASCH solutions for $\Delta t = 100$s

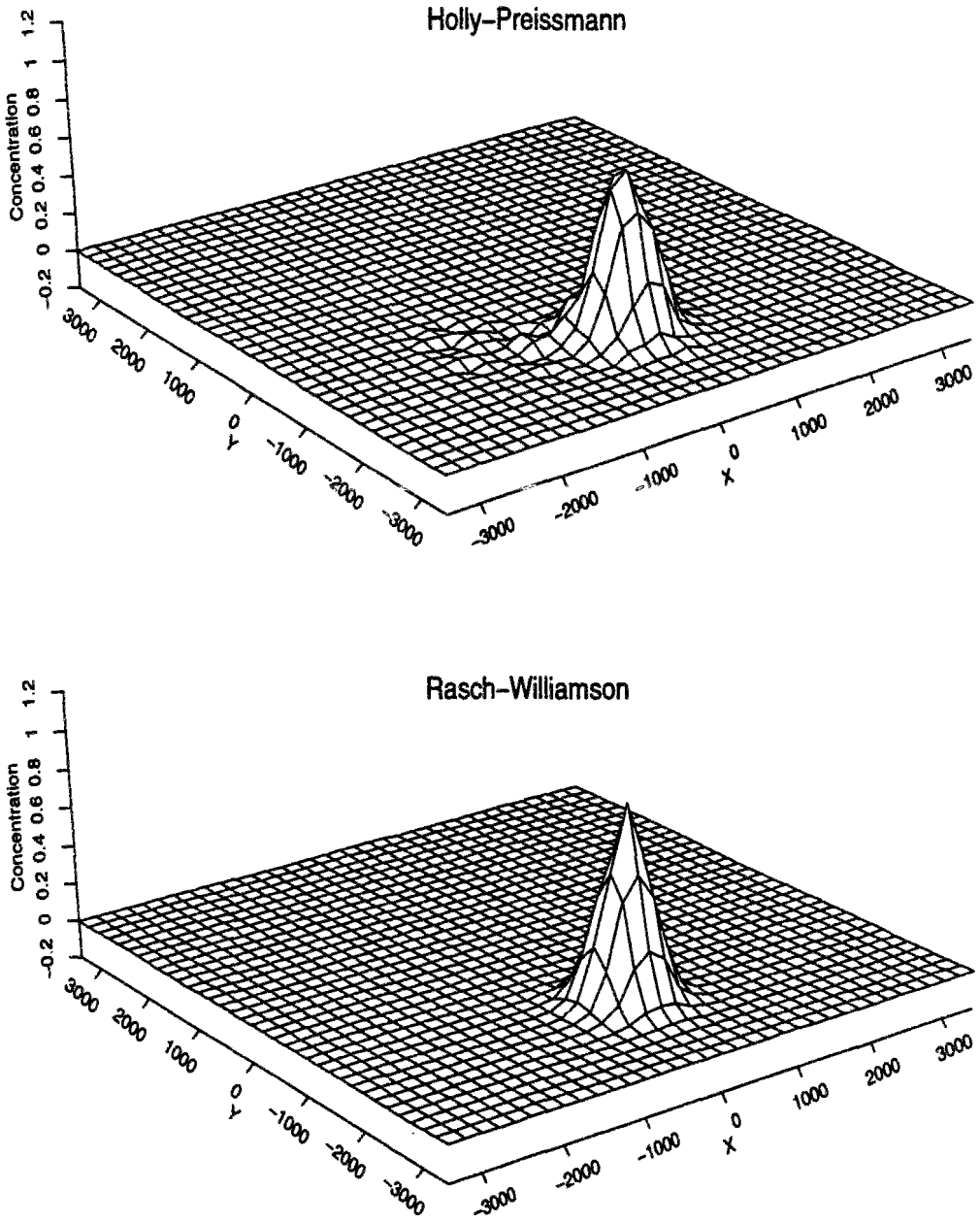


Figure 7.16: Gauss-hill : HOLLY & RASCH solutions for $\Delta t = 500$s

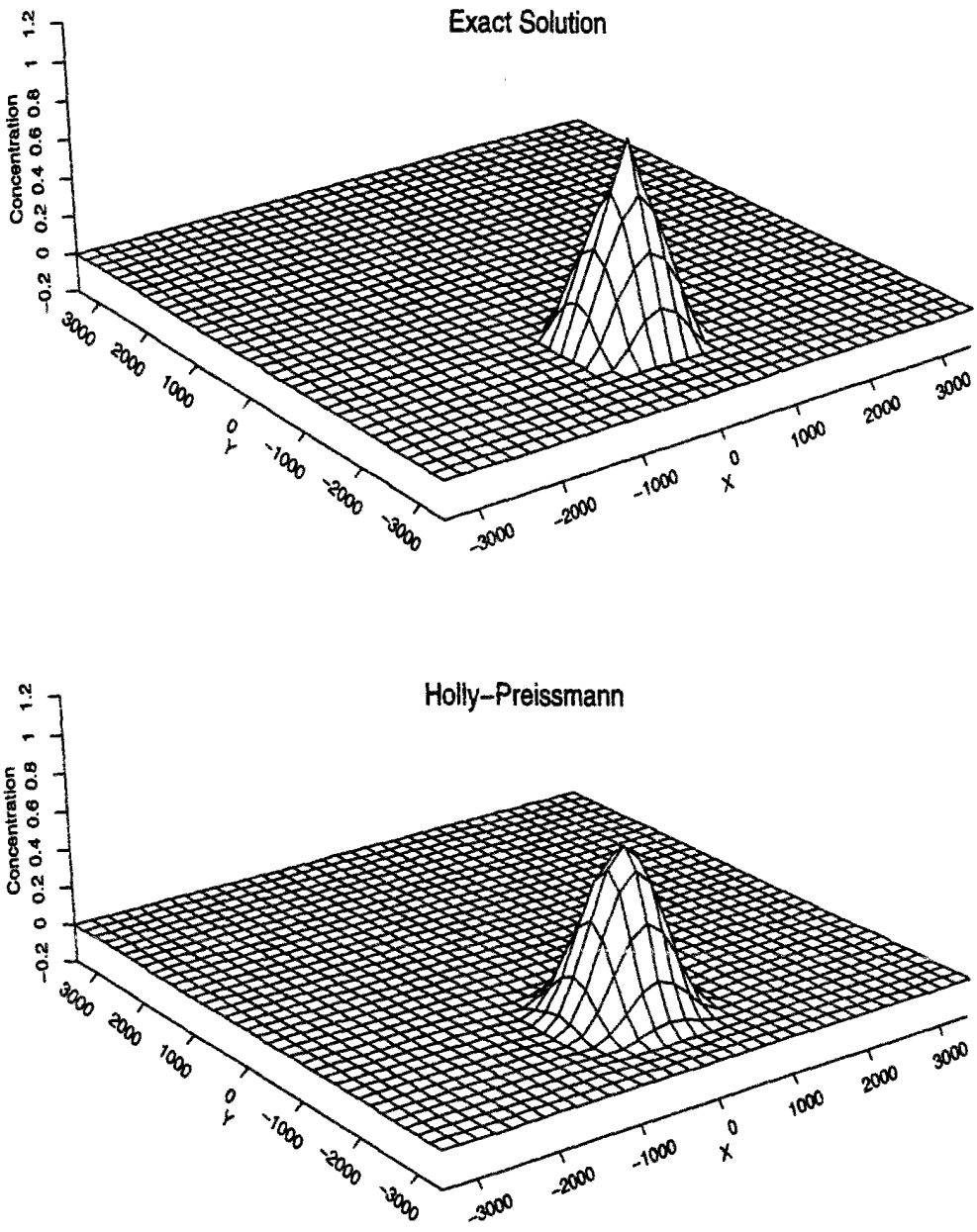


Figure 7.17: Cone-hill : Exact & HOLLY solutions after a full revolution (Δt = 30s)

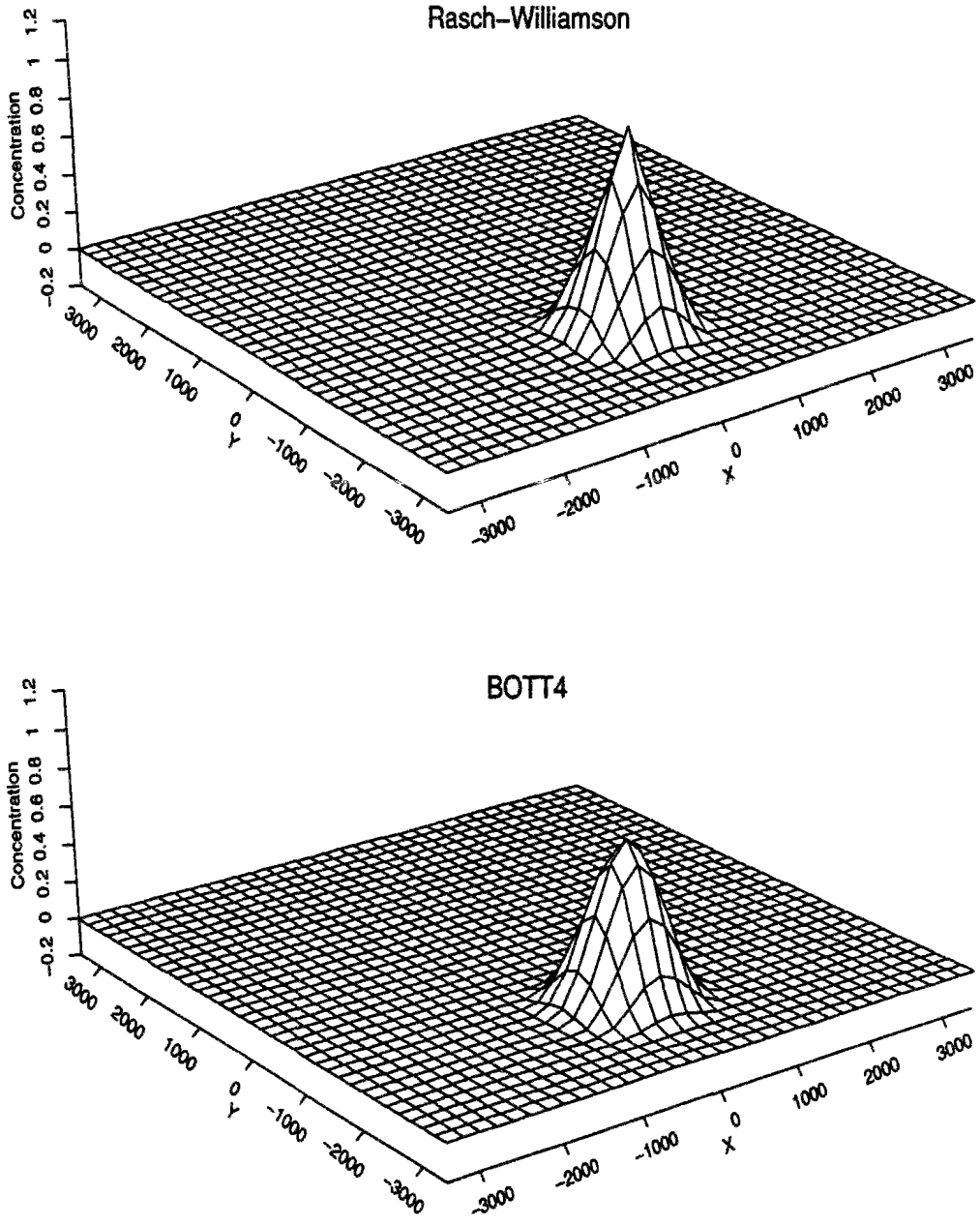


Figure 7.18: Cone-hill : RASCH & BOTT4 solutions after a full revolution (Δt = 30s)

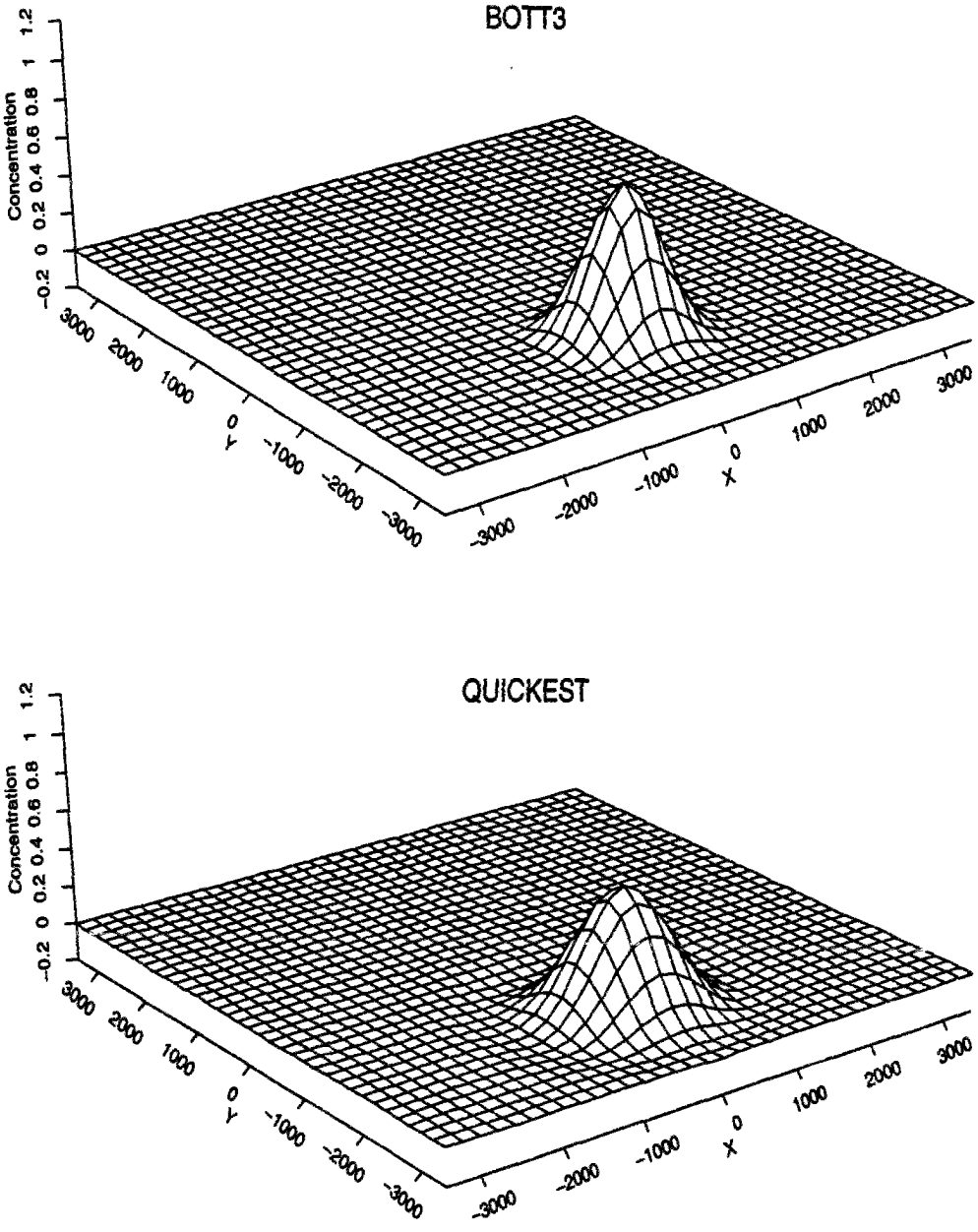


Figure 7.19: Cone-hill : BOTT3 & QUICKEST solutions after a full revolution ($\Delta t = 30$s)

7.3.2 Anisotropic diffusion of gauss-hill : Tests B

This second test case allows us to examine the performance of the different schemes when there exists some moderate physical diffusion. It is also an opportunity for checking the accuracy of the algorithm solving the diffusion step.

This test turns out to be less severe than the preceding one : the errors levels are significantly lower, as can be observed in figures 7.20 and 7.21. The best scheme for this test case proves to be the Holly-Preismann one, followed by Rasch-Williamson, then BOTT4, BOTT3 and lastly QUICKEST.

RASCH displays once again degradation for the smallest time step. On the other hand, as the transported distribution becomes progressively wider and smoother due to diffusion, HOLLY maintains good results even when large time steps are used.

In the range of time steps which ensure the stability of flux-form methods, BOTT4 could be preferred to RASCH because, while it introduces more damping, it better preserves the shape of the transported scalar field (cf figures 7.22 to 7.24).

Yet, once again, the backward characteristic methods offer more opportunity to reach a satisfactory solution than can the flux-form methods in the frame of their stability constraints.

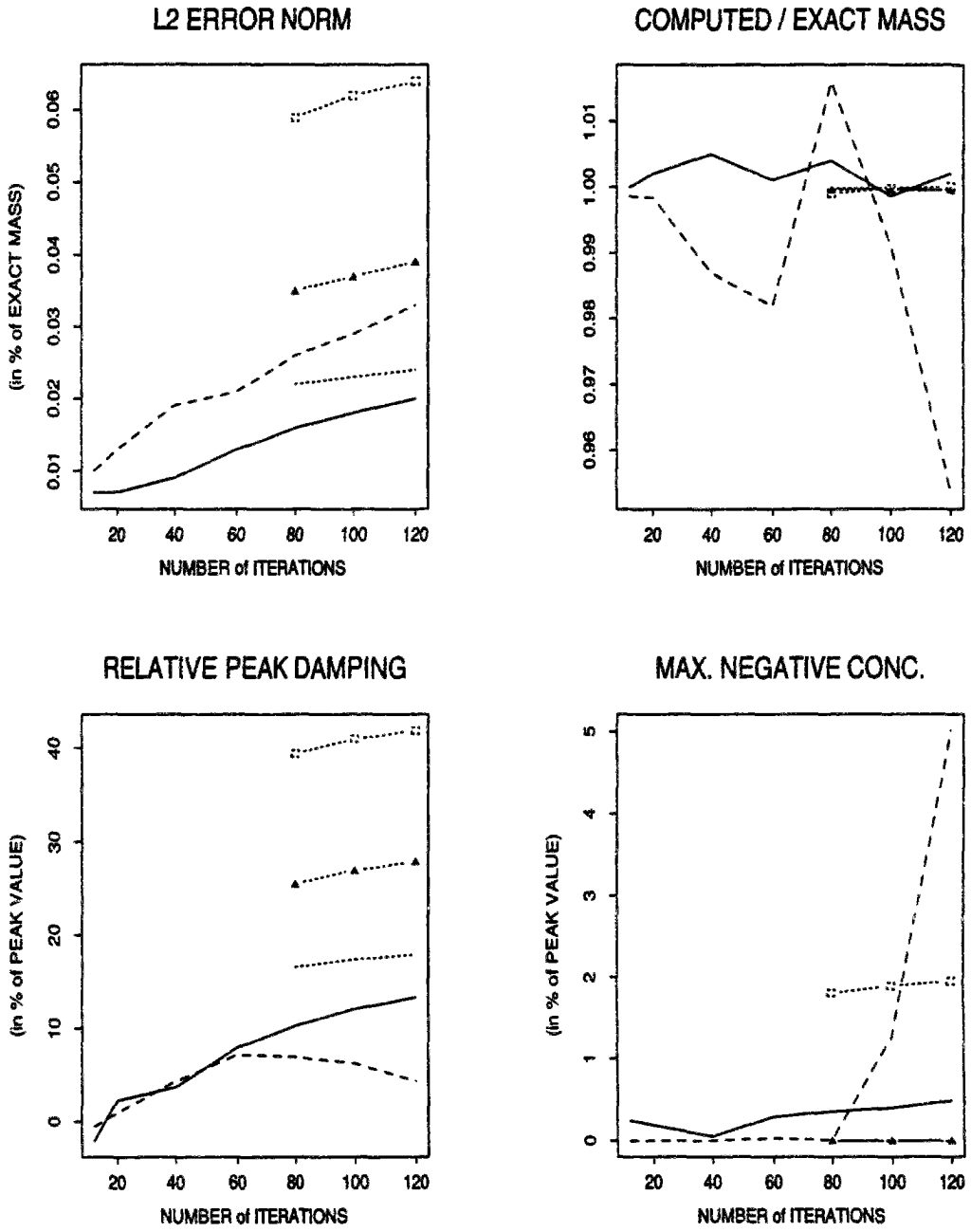


Figure 7.20: Error measures for anisotropic diffusion test (part 1)

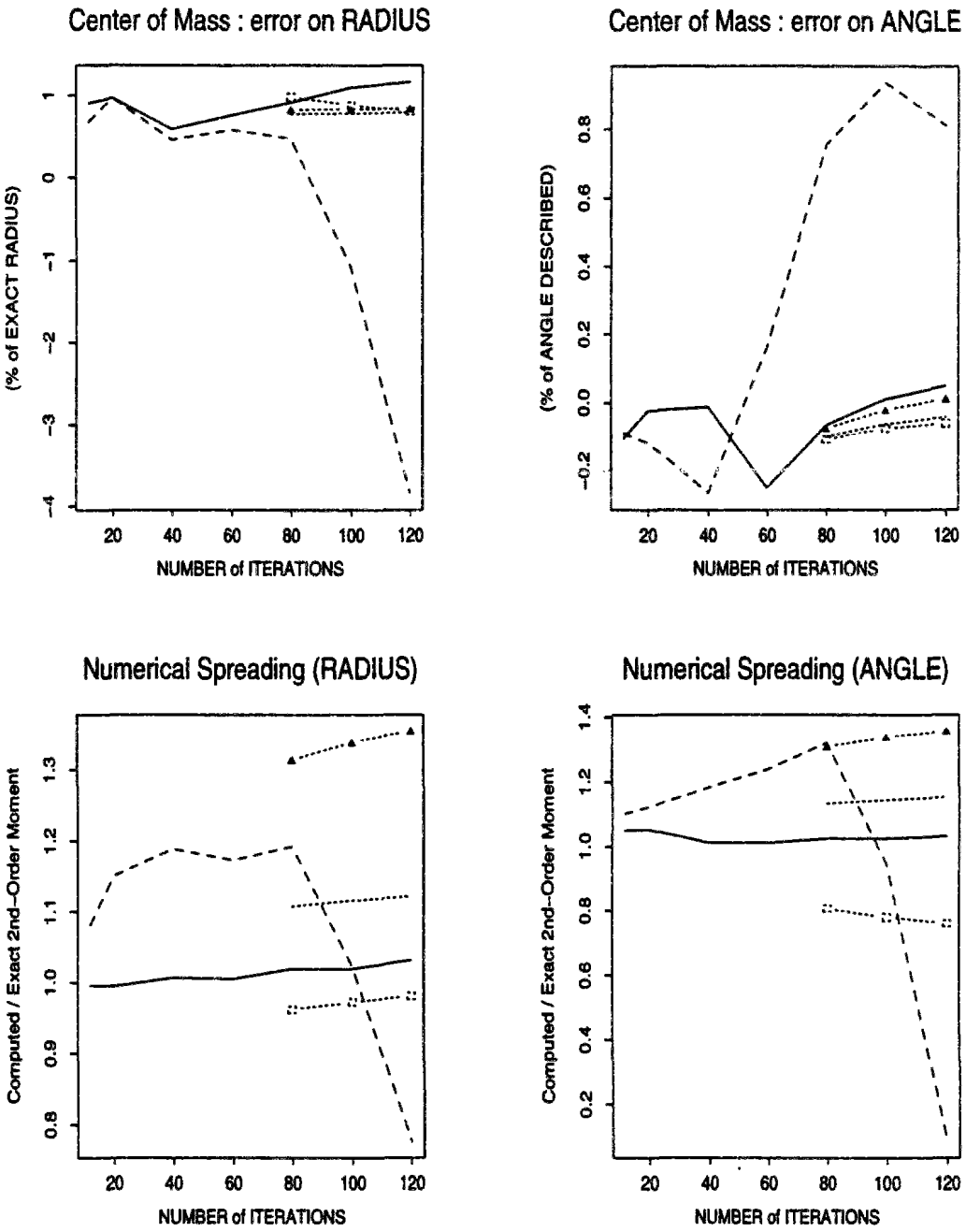


Figure 7.21: Error measures for anisotropic diffusion test (part 2)

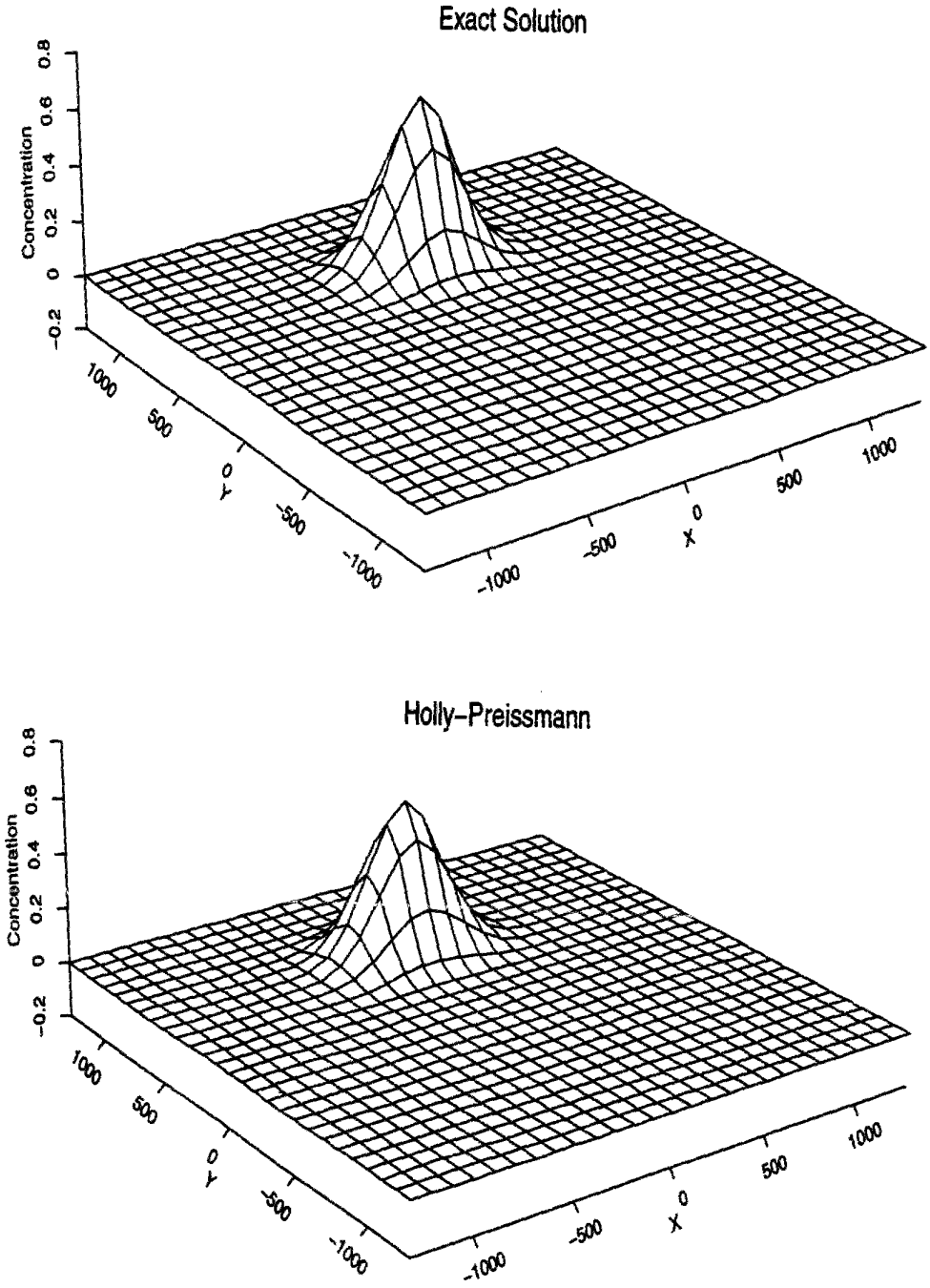


Figure 7.22: Anisotropic diffusion : Exact & HOLLY solutions for $\Delta t = 150s$

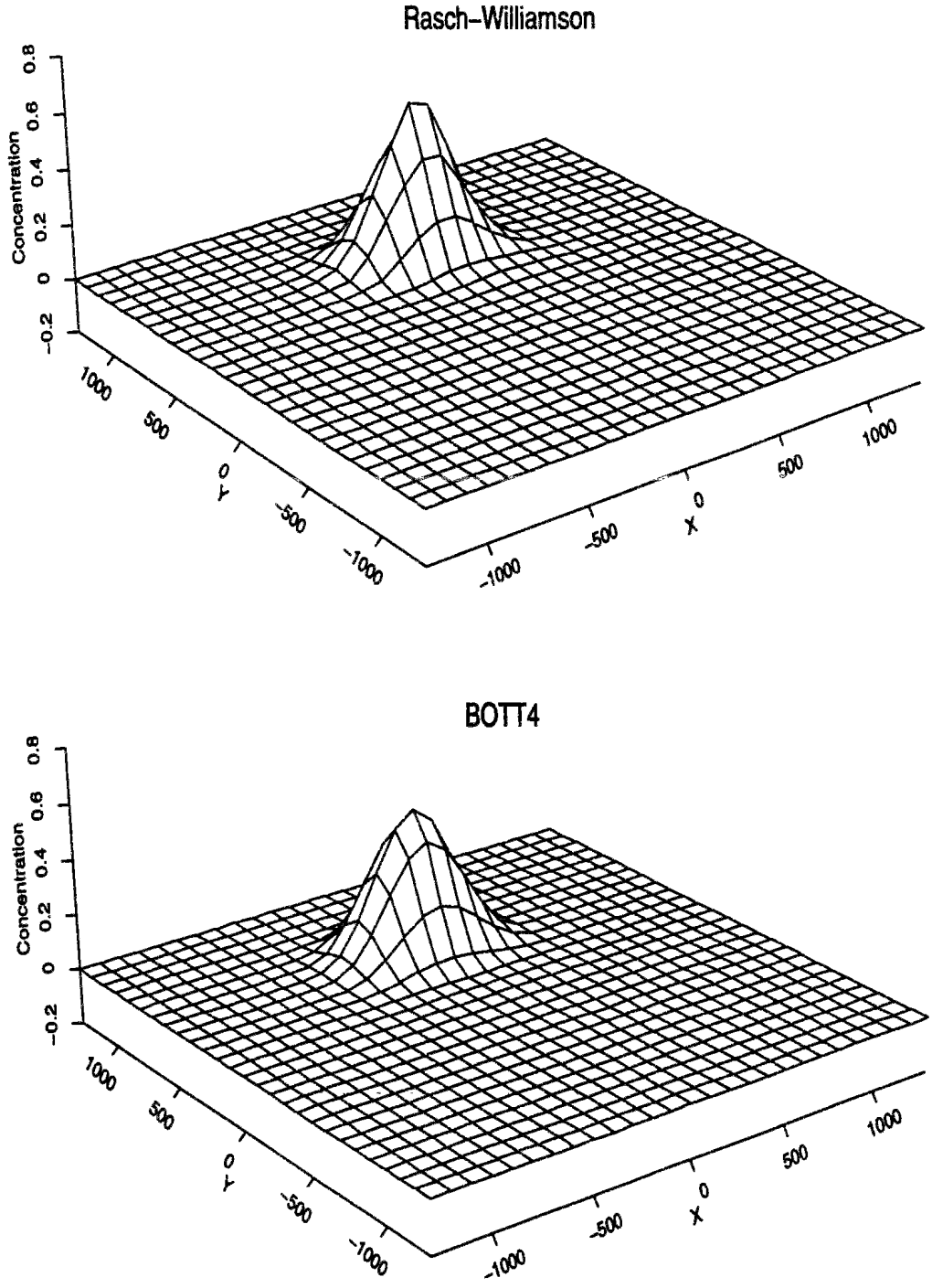


Figure 7.23: Anisotropic diffusion : RASCH & BOTT4 solutions for $\Delta t = 150$s

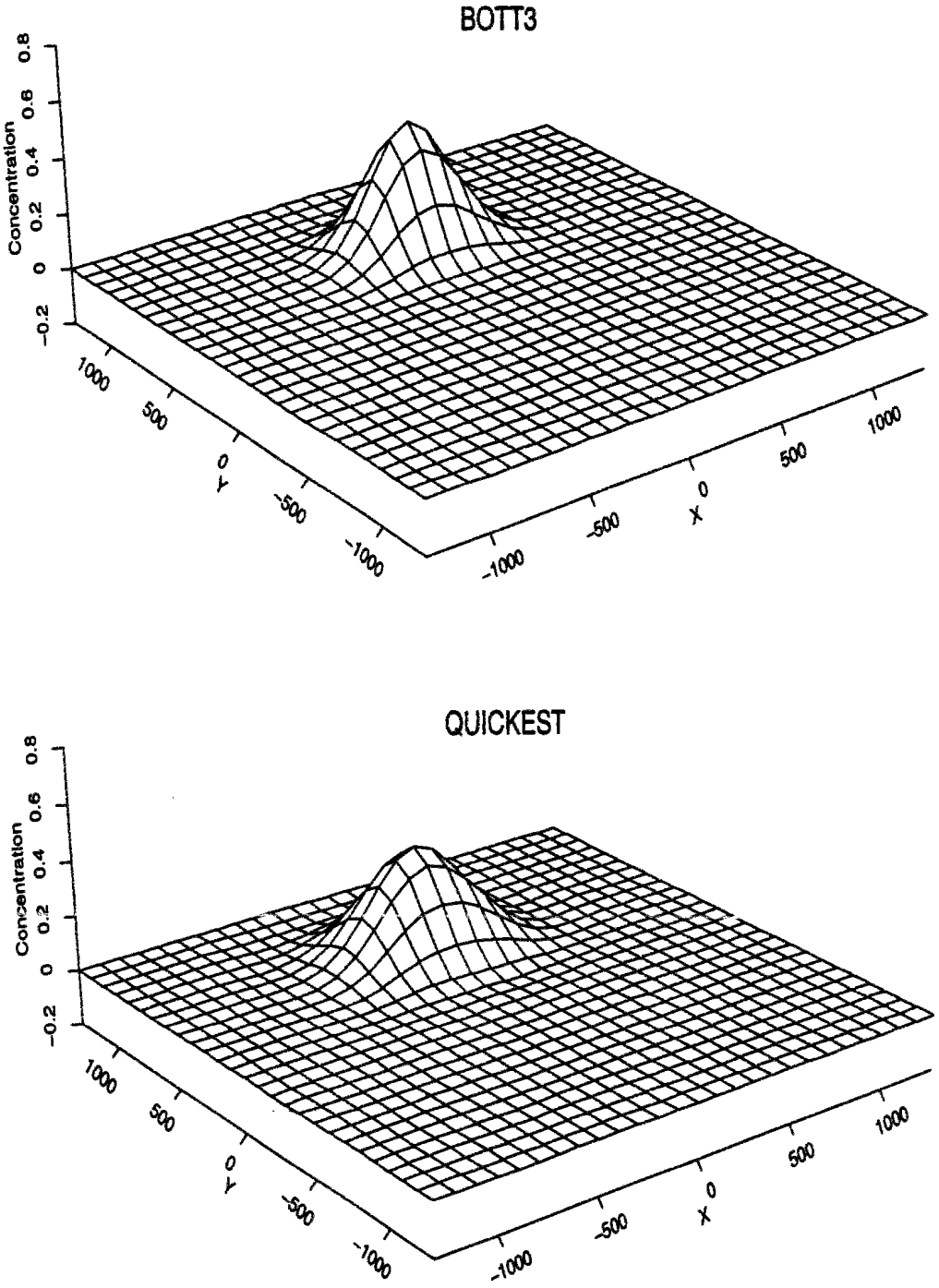


Figure 7.24: Anisotropic diffusion : BOTT3 & QUICKEST solutions for $\Delta t = 150$s

7.4 Computer time requirements

The CPU time requirements for each scheme have been assessed as for the one-dimensional situation, namely by averaging the running times of several simulations ($\simeq$ 20). We consider versions of the schemes designed to deal with non-uniform and unsteady flow and diffusivity fields and able to apply to variable grid spacing.

Notably, this time, we have been effectively including in the Holly-Preissmann scheme the computation of the corrective terms which appear in the advection equations governing the first-order spatial derivatives of the transported scalar (cf Appendix E.1.3). Different trials led us to compute the corrective terms with the help of a bicubic interpolator (cf Appendix F.4).

In the case of the backward characteristics methods, the times indicated for the backtracking stage are obtained when breaking the trajectory into two segments within each mesh cell, which proved to be accurate enough.

Results are summed up in table 7.6. The indicated times refer to the treatment of one grid node per time step :

1. The time required for backtracking the particles' trajectory is slightly more important for the Holly-Preissmann than for the Rasch-Williamson scheme. This is due to time spent in calling the subroutine which performs the integration along the trajectory of the corrective terms in the scalar derivatives equations.

2. The interpolation time is much more important. This large difference does not stem from the fact that three variables (C, C_x and C_y) instead of one are interpolated at the characteristic foot but comes from the time spent in evaluating the corrective terms (which accounts for 75 % of the time consumed by this stage of the scheme).

3. The diffusion stage is the same for four of the five schemes. As it diffuses three variables instead of one only, the Holly-Preissmann solution of the diffusion step is more costly. Yet, as some computations are made only once for the three variables (notably the evaluation and inversion of the linear systems originated by the discretization of the diffusion operator), the cost is not multiplied by three but only doubled.

4. The flux-form finite difference methods appear to be significantly cheaper than the backward characteristic methods, at least as far as the transport of one scalar only is concerned. Indeed, the application of RASCH to the simultaneous transport of two scalars already turns out to be less expensive than the use of BOTT4 (319 μs instead of 352 μs), as the computation of the trajectory needs to be performed only once for the two variables.

5. The Holly-Preissmann scheme is undoubtedly the less economical. Contrary to what occurs

for the Rasch-Williamson scheme, its extension to multiple variable transport does not result in significant savings as its most expensive stage is not the backtracking one but the interpolation stage.

Table 7.6: CPU time requirements for two-dimensional case

Scheme	Advection stage (μs)		Ratio vs. QUICKEST	Diffusion stage (μs)	Global Time	Ratio vs. QUICKEST
QUICKEST	21.2		**1.0**	50.3	72.5	**1.00**
BOTT3	61.9		**2.9**	50.3	113	**1.56**
BOTT4	123.1		**5.8**	50.3	176.2	**2.43**
	Backtracking	Interpolation				
RASCH	136.6	40.7	**8.3**	50.3	230.1	**3.17**
HOLLY	152.2	152.8	**14.4**	108	416.3	**5.74**

7.5 Conclusions on the advection-dispersion solution

We propose hereafter the conclusions we have been drawing from the performed series of one- and two-dimensional tests.

1. Most of our test cases deal with narrow initial sources displaying sharp concentration gradients. In such case, **when advection is dominant, best results can generally be obtained through the use of backwards characteristic methods**, notably because these methods are free from stability constraints and can thus be applied on a wider range of time steps.

   This conclusion stands only for methods which use a high-order interpolator at the origin of each characteristic line. The Hermit cubic polynomial is such an interpolator, provided it is combined with an accurate evaluation of the first-order scalar derivatives. Both the method suggested by Holly and Preissmann (Holly & Preissmann, 1977; Holly & Usseglio-Polatera, 1984; Usseglio-Polatera, 1988) and the Akima-derivative estimate introduced by Rasch and Williamson (Williamson & Rasch, 1989; Rasch & Williamson, 1990) prove to be adequate.

   The above conclusion in favour of backward characteristics method should be adjusted according to the planned applications. Indeed, in real field situations, we shall need to deal with different families of problems. The first one concerns the short-term fate of point

sources in the receiving water body : in the first stages of the pollutant cloud development, an adequate representation of the transport is essential. The second one concerns more general water quality problems, where the transport is but one of the phenomena which govern the evolution of some chemical species more evenly distributed than a point source, and where we are possibly mainly interested in medium-term and long-term trends. The first category of applications has features similar to the academic tests we performed and should thus be treated with backward characteristics methods. On the other hand, as regards the second kind of problems, we should not discard the use of such simpler methods as the QUICKEST one.

2. In practical applications we may find either uniform grids, either continuously varying grids or lastly grids made of the juxtaposition or nesting of coarse and fine grids. In the latter case, the "transition area" (where the otherwise uniform grid spacing evolves between coarse and fine values) has a more limited extent than in the second case. Uniform grids are frequently easier to handle but the use of variable grids or a combination of different uniform grids may be more suited to the exploitation of unevenly distributed field data or to the description of local specificities.
   As far as we have been testing it, the various schemes are not very sensitive to grid non-uniformities and could thus be applied rather indifferently to variable or constant discretizations. Yet, it is worth noting that the use of an uniform grid induce drastic savings as regards the cost of high-order flux form methods (cf section 6.6).

   Similarly, the treatment of boundary conditions (cf section 6.5) does not seem to affect significantly the schemes accuracy. In the case of backward characteristics methods (except the Holly-Preissmann one), open boundary conditions are particularly straightforward.

3. In contrast to their poor reputation **backward characteristic methods are not significantly more costly than high-order flux-form methods**. The CPU time consumed by the treatment of one node per time step exceeds indeed the corresponding time required by the tested flux-form methods. However, characteristic methods can be applied with larger time steps, which reduces the number of operations necessary to reach the end of the simulation. Besides, apart from the Holly-Preissmann method, characteristic methods allow significant savings when dealing with several variables, since the most expensive stage, namely the estimation of particle trajectory, needs to be performed only once. Let us consider for instance the simultaneous solution of the two-dimensional flow and transport equations : at least three variables (one scalar concentration and the two components of the flow velocity) undergo an advection-dispersion process. Thus, for this problem, the interest in using a backward characteristic method is obvious.

   These conclusions concerning the computational efficiency of characteristic methods would of course be all the more confirmed should we compare these schemes to still higher-order

flux-form methods (which, on the other hand, could prove more accurate than the methods tested so far). Yet, we should mention that a promising efficient approach to the solution of advection-dispersion has been suggested by (Leonard & Niknafs, 1991) : it consists of switching to different flux-form methods according to the local features of the transported scalar field (importance of gradients and curvature). Thus, high-order methods are applied only when there are truly needed while the cheapest methods are being used over most of the computational domain. Application of this methodology should enhance the benefits linked to the use of flux-form methods. However, some further work is required in order to design adequate "switching criteria" in most practical applications.

4. Lastly, it is worth noting that the best flux-form methods are those which use positiveness-enforcing limiters (see for more details Appendix F.1). The extension of such techniques to the transport of non-positive scalars is consequently not straightforward. As previously mentioned (cf sec. 4.2.5), one possibility is to add to the scalar field a large positive constant, deal with the modified distribution and then substract the constant (Smolarkiewicz & Clark, 1986). Yet, as most limiters are not linear, the solution could prove dependent on the chosen constant. To which extent ? This should be investigated ...

   On the other hand, backward characteristic methods deal similarly with positive and non-positive scalar fields.

In conclusion to the above remarks, **we finally advocate the use of backward characteristic methods to deal with the solution of advection-dispersion problems whenever advection is the dominant process** (for diffusion-dominated problems, nearly any scheme behaves fairly !). **The Rasch-Williamson algorithm, which relies on the Hermit cubic interpolator and the Akima derivatives estimate, achieves,** in our opinion, **the best balance between accuracy and computational efficiency** for multidimensional applications.

We are fully aware that this algorithm is not perfect. First, it is not quite conservative. Secondly, it eventually yields inadequate results for a few specific combinations of temporal and spatial discretizations for reasons which are still unclear. In one dimensional situations, the extra cost induced by the use of the Holly-Preissmann method is affordable and this scheme could be preferred as, while introducing slightly more damping, it ensures a better mass preservation and is somewhat more "predictable".

Ultimately, our choice has been influenced by the applications we are planning. Indeed, our surface flow and transport model has been developed in the frame of a multidisciplinary research program concerning the impact of combined sewer storm overflows on a receiving river of medium importance (the Seine River which flows through Paris). Consequently, we shall need to focus on the short-term, near-field fate of narrow point sources in unsteady situations and

be especially interested in the forecast of extreme, critical concentrations (e.g. linked to oxygen depletion). We decided to take the risk to work with a non-conservative algorithm as, on the other hand, it proved to preserve extremely well peak values. The Rasch-Williamson algorithm forms thus the core of the advection part of our model.

7.6 Résumé français : "Cas-tests bidimensionnels"

L'étude de cas-tests monodimensionnels nous a d'ores et déjà permis d'apprécier les performances relatives des schémas sélectionnés. Il est néammoins indispensable de compléter cette analyse par des problèmes multi-dimensionnels. Nous nous limiterons au cas bidimensionnel, compte tenu de notre cadre de travail (la Seine).

Un coup d'oeil à la bibliographie indique que les tests les plus couramment pratiqués sont : la rotation de distributions de taille limitée (de forme gaussienne, conique, en cosinus), avec ou sans diffusion; le transport de gaussiennes dans des écoulements cisaillés; le transport de fronts de concentration dans un écoulement à 45 degrés (par rapport aux axes de calcul); la distorsion de champs de concentration dans des écoulements composés de la juxtaposition de circulations en cellule. La dernière situation est typique des écoulements observés en météorologie et ne possède pas de solution analytique. Des trois autres tests, le plus instructif est le premier, comme nous l'avons vérifié lors de travaux antérieurs (Simon, 1990a). On pourrait penser que le second test est caractérisé par des écoulements plus proches de ce qu'on observe en rivière que dans le premier test (cf figure 7.1). Cependant, il ne met en jeu que des champs de vitesse unidirectionnels et uniformes dans le sens du courant, ce qui est beaucoup moins complet. Par ailleurs, il débouche sur le même classement relatif des schémas les uns par rapport aux autres. Pour rester bref, **nous nous ne présenterons donc que les résultats du test le plus riche d'enseignements, à savoir la rotation, avec ou sans diffusion, de distributions de taille limitée. Nous ne nous intéresserons qu'aux 4 schémas les plus performants en monodimensionnel, à savoir RASCH, HOLLY, BOTT4 et BOTT3 ainsi qu'au schéma le plus économique, c'est à dire QUICKEST.**

Les conditions des tests pratiqués sont les suivantes (cf section 7.2) :

- L'écoulement décrit une rotation dans le sens inverse des aiguilles d'une montre. On évalue les solutions numériques après une révolution complète de la distribution de concentrations.
- La première série de tests (tests A) concerne une advection pure. Dans la seconde (tests B), il existe une diffusion anisotropique (2 fois plus forte dans le sens du courant que perpendiculairement) mais l'advection est toujours dominante.
- Les essais d'advection pure sont pratiqués sur une grille uniforme, puis sur une grille variable (pas d'espace compris dans une fourchette de ± 25 % par rapport à leur moyenne, variation d'au plus 2.3 % entre deux mailles adjacentes). Dans le cas de la diffusion anisotropique, le maillage est constant.
- Pour les tests A, on utilise 2 types de distribution initiale, une gaussienne et un cône. Dans les deux cas, la plus grande largeur de la distribution est en gros de 8 mailles de calcul. Pour les tests B, la distribution initiale est gaussienne. Elle s'inscrit dans une ellipse de 8.5 mailles de long, 6 mailles de large. A la fin d'une rotation, les axes de l'ellipse se sont allongés à 11 et $\simeq$8 mailles respectivement.

- Compte tenu de leur condition de stabilité (nombres de Courant locaux inférieurs à 1), les schémas QUICKEST, BOTT3 et BOTT4 ne sont applicables que pour une gamme de pas de temps assez limitée. Ainsi, pour les tests A, on peut utiliser tous les schémas pour des pas de temps tels que la rotation soit complétée en 100 à 300 itérations. Avec les méthodes aux caractéristiques on descend ensuite jusqu'à 6 itérations. Pour les tests B, la limite commune d'application correspond à une rotation en 80 itérations; pour les méthodes aux caractéristiques on a exploré le comportement jusqu'à 12 itérations.

Advection pure Les résultats de la série de tests A sont exposés en section 7.3.1. Nous indiquons d'abord le comportement général des mesures d'erreur :

- **Les résultats sont similaires que la distribution initiale soit une gaussienne ou un cône.** Ils sont cependant un peu meilleurs dans le second cas. Ceci est sans doute dû au fait que le cône est une source un peu plus large.
- **Les résultats sont peu affectés par le changement de discrétisation spatiale.**
- **Les performances de QUICKEST, BOTT3 et BOTT4 sont peu sensibles au choix du pas de temps, dans la gamme testée. Les erreurs de RASCH sont des fonctions décroissantes du pas de temps.**
- **Par contre, le comportement des erreurs de HOLLY n'est pas parfaitement monotone : elles augmentent brutalement pour le (ou les deux) plus grands pas de temps.** Ce phénomène est dû à la méthode d'estimation des dérivées nodales nécessaires à l'application de l'interpolateur bicubique hermitien. Ces dérivées sont en effet calculées en résolvant une équation de transport, qui diffère de l'équation des concentrations par un terme source, lequel doit être intégré le long des trajectoires. Plus le pas de temps est grand, plus la longueur des trajectoires est grande et le calcul du terme source délicat ...jusqu'à ce qu'il finisse par fausser l'algorithme ! Différentes façons de contrôler cette erreur sont examinées en annexe F.4.

Détaillons maintenant les résultats pour chaque mesure d'erreur :

1. **Mesure globale des écarts entre solutions numérique et exacte** Cette mesure est donnée par la norme L2 de la différence entre les deux solutions. L'erreur de QUICKEST est en gros 2 fois celle de BOTT3, qui a son tour vaut 2 fois celle de BOTT4. L'erreur de RASCH devient inférieure à celle de BOTT4 quand le pas de temps est tel qu'on effectue moins de 150 itérations pour une révolution. Les erreurs les plus faibles sont systématiquement engendrées par HOLLY (sauf pour le plus grand pas de temps).
2. **Perte de masse** Elle est négligeable (dans le cas des schémas conservatifs, elle est induite uniquement par le traitement des conditions limite, toutes les frontières étant ouvertes) sauf pour RASCH, où elle excède 5 % pour plus de 150 itérations.
3. **Amortissement du maximum** Le meilleur schéma est RASCH (moins de 10 % d'erreur), suivi de HOLLY (de 10 à 20 % d'amortissement). Les erreurs engendrées par BOTT4, BOTT3 et

QUICKEST sont stables, atteignant respectivement autour de 25, 40 et 60 %. Ceci va de pair avec l'apparition de diffusion numérique (élargissement à la base des distributions transportées).

4. **Concentrations négatives** Elles sont modérées pour QUICKEST (moins de 2 % de la valeur du maximum de concentration) et HOLLY (de 1 à 4 %). Pour RASCH, leur amplitude est au plus de 2 % du pic tant que le nombre d'itérations est inférieur à 150, mais grimpe à 10 % pour 300 itérations !

5. **Déphasage** Le déphasage est faible, sauf pour RASCH au plus petit pas de temps.

En résumé, **seuls les schémas aux cartéristiques permettent de résoudre correctement ce test, moyennant un choix judicieux du pas de temps. Dans la gamme de pas de temps qui leur est autorisée** (pour des raisons de stabilité) **les autres algorithmes engendrent un amortissement inacceptable du pic de concentration.** Le schéma de RASCH est celui qui préserve le mieux la valeur de ce pic. Cependant, il déforme un peu la distribution transportée quand celle-ci est gaussienne, en lui donnant une forme plus conique. Le schéma HOLLY amortit plus le pic mais est meilleur du point de vue de la conservation de la masse et de la forme des distributions.

Diffusion anisotropique (section 7.3.2) Grâce à l'introduction d'une faible diffusion, cette série de tests se révèle un peu moins sévère que la précédente. **Par ordre de performance décroissante, les schémas se classent comme suit : en premier HOLLY, puis RASCH, BOTT4, BOTT3 et QUICKEST.**

1. RASCH a encore une fois un mauvais comportement pour le plus petit pas de temps testé. Par contre, la diffusion adoucissant les gradients de concentration, étalant la distribution initiale, le comportement de HOLLY aux plus forts pas de temps n'est plus dégradé.

2. Dans la gamme de pas de temps où il est applicable, on peut préférer l'emploi de BOTT4 à celui de RASCH car, quoiqu'il amortisse un peu plus le pic de concentration, il préserve mieux (triangularise moins) la forme de la distribution.

3. Là encore, du fait de leur plus grand domaine d'utilisation, les méthodes aux caractéristiques permettent bien mieux que les autres d'accéder à un résultat numérique satisfaisant.

Coûts informatiques (section 7.4) Les coûts de chaque schéma ont été estimés comme dans le cas monodimensionnel, sur la moyenne d'une vingtaine de simulations correspondant au cas d'advection combinée avec une diffusion anisotropique :

1. **Les consommations relatives en temps CPU, par rapport au schéma le plus économique, à savoir QUICKEST, sont de 1.6, 2.4, 3.2 et 5.7 pour BOTT3, BOTT4, RASCH et HOLLY respectivement** dans le cas du transport d'une seule variable.

2. **Le coût nettement plus important du schéma d'Holly-Preissmann est dû au fait qu'on est amené à résoudre une équation de transport pour les dérivées premières des concentrations selon chaque direction de l'espace** (on rappelle que la connaissance des dérivées nodales est nécessaire à la mise en oeuvre de l'interpolateur hermitien).

   D'une part, on se retrouve à travailler avec 3 variables au lieu d'une. D'autre part, l'équation de transport des dérivées diffère de l'équation des concentrations par des termes sources qui doivent être intégrés le long des trajectoires. Ces termes sources sont fonction du gradient du champ de vitesses. Dans les tests monodimensionnels, l'écoulement était uniforme et par conséquent les termes sources nuls ...et leur calcul immédiat ! Dans le cas de l'écoulement en rotation, non uniforme, le poids de leur calcul est évident : il explique en particulier que la résolution de l'étape d'advection coûte 75 % plus en terme de temps CPU pour HOLLY que pour RASCH (alors que les coûts étaient les mêmes en monodimensionnel).

   La structure des coûts n'est pas la même pour les deux schémas. Dans RASCH, la remontée de la caractéristique représente 80 % de l'étape d'advection, 50 % uniquement pour HOLLY. Ainsi, si on advecte deux variables au lieu d'une le temps CPU (advection seule) augmentera d'environ 20 % avec RASCH, de 50 % avec HOLLY. Les économies d'échelle sont nettement plus importantes avec le premier schéma qu'avec le second !

3. **L'étape de diffusion est la même pour 4 des schémas, à savoir QUICKEST, BOTT3, BOTT4 et RASCH.** Elle est bien sûr plus coûteuse pour HOLLY (diffusion des dérivées nodales également). Cependant, comme certaines phases du calcul sont communes (i.e. le calcul des systèmes linéaires issus de la discrétisation de l'opérateur de diffusion), le coût n'est que doublé (et non triplé).

4. Pour compléter le panorama, nous indiquons les coûts CPU relatifs (toujours par rapport à QUICKEST) pour des problèmes multivariables et dans l'ordre pour BOTT3, BOTT4, RASCH et HOLLY : 1.6, 2.4, 2.2 et 4.6 avec 2 variables; 1.6, 2.4, 1.9 et 4.3 avec 3 variables. On constate que **dès que l'on travaille avec plus de 2 variables, RASCH devient plus économique que BOTT4.**

Conclusion générale des tests mono- et bidimensionnels (section 7.5)

1. La plupart des tests que nous avons pratiqués concerne le **transport de sources de petite taille**, où les **gradients de concentration** sont **marqués**, dans des situations où l'**advection** est **dominante**. Ces cas reflètent nos préoccupations en ce qui concerne la Seine (injection ponctuelle de charges trés polluantes). **Dans ce cadre, c'est l'utilisation de méthodes aux caractéristiques qui permet d'atteindre les résultats les plus précis, pourvu qu'on applique un interpolateur cubique (schémas d'Holly-Preissmann et de Rasch-Williamson).** Ceci est possible notamment parce que les méthodes aux caractéristiques ont un domaine d'application plus vaste (pas de condition de stabilité contraignant le choix du pas de temps).

2. **Cependant, pour des applications où le transport n'est pas un phénomène majeur** (ex : évolution à moyen ou long terme d'un composé non-conservatif, donc soumis également à des réactions biogéochimiques, et distribué initialement un peu partout dans le milieu étudié), **on pourra utiliser avec profit des algorithmes moins sophistiqués, QUICKEST par exemple.**

3. Les différents algorithmes testés se sont tous révélés robustes quant au mode de discrétisation adopté (grille de calcul uniforme ou non). On notera cependant que pour certains d'entre eux (BOTT4 notamment, BOTT3 dans une moindre mesure), l'utilisation d'un pas d'espace constant permet de réduire considérablement le temps calcul.

4. De même, tous les algorithmes traitent correctement les conditions limite.

5. **Contrairement à leur mauvaise réputation, les schémas aux caractéristiques ne sont guère plus coûteux que les les méthodes conservatives d'ordre 3 et plus** (c.a.d. telles que le polynôme approximant le champ de concentration pour le calcul des flux advectifs est d'ordre 3 et plus). Leur coût *individuel*, i.e. pour le calcul des concentrations en *un* noeud pour *un* pas de temps, est généralement plus élévé. Cependant, avec les schémas aux caractéristiques on peut avoir recours à des pas de temps significativement plus grands, ce qui diminue d'autant le nombre de calculs à effectuer pour boucler une simulation. Par ailleurs, dès qu'on transporte plusieurs variables, on effectue des économies d'échelle spectaculaires (sauf avec HOLLY) puisque l'étape de remontée des caractéristiques est commune à toutes les substances transportées.

6. Enfin, on notera que parmi les méthodes conservatives testées, les meilleures sont sans conteste celles qui incluent l'utilisation d'un limiteur qui permet de préserver la positivité. L'extension de ces techniques au transport de variables qui ne sont pas systématiquement positives (par exemple, pour calculer l'advection des composantes de la vitesse) est loin d'être claire.

En conclusion, **au vu des remarques ci-dessus, et compte tenu des applications en Seine envisagées, nous estimons que le schéma qui offre le meilleur "rapport qualité/prix" est la méthode aux caractéristiques de Rasch-Williamson.** Nous sommes bien conscients que ce schéma n'est pas parfait. Notamment, il lui arrive d'induire des pertes de masse non négligeables. Cependant, il préserve remarquablement bien la valeur des pics de concentration.
Si nous avions été intéressés au premier chef par des applications monodimensionnelles au lieu d'applications bidimensionnelles, nous aurions probablement retenu le schéma d'Holly-Preissmann, qui, un peu plus dissipatif, est en revanche presque parfaitement conservatif. Le fait qu'il soit plus gourmand en temps CPU est en effet supportable en monodimensionnel.

Part III

Modelling of depth – averaged flows

Chapter 8

Resolution of the shallow-water equations

8.1 Mathematical formulation

The usual model for describing depth-averaged flows relies on the St-Venant equations. Its derivation has been presented in chapter 2, in which we have been discussing too the relevance of this approach (cf section 2.4.5). We recall hereafter its expression in cartesian coordinates and the related notations.

The three dependent variables in the St-Venant equations may either be the free surface elevation and both components of the depth-averaged velocity or the free surface elevation and the unit-width discharges in each direction. The corresponding sets of equations are respectively eq. 8.1 to 8.3 in the first case, eq. 8.5 to 8.7 in the second case. As we shall be mainly interested in modelling domains of limited extent and/or fluvial flows, we neglected both the Coriolis effect and the wind surface stress. Following Benque (Benque *et al.*, 1982), we distinguis the different processes governing mass and momentum conservation.

With $\left(\zeta, \vec{U} \right)$ variables, the St-Venant equations read :

$$\frac{\partial \zeta}{\partial t} + \mathrm{div} h\vec{U} = 0 \tag{8.1}$$

$$\frac{\partial u}{\partial t} + \vec{U}.\nabla u + g\frac{\partial \zeta}{\partial x} + f_b u = S_x \tag{8.2}$$

$$\underbrace{\frac{\partial v}{\partial t}}_{\mathrm{I}} + \underbrace{\vec{U}.\nabla v}_{\mathrm{II}} + \underbrace{g\frac{\partial \zeta}{\partial y}}_{\mathrm{III}} + \underbrace{f_b v}_{\mathrm{IV}} = \underbrace{S_y}_{\mathrm{V}} \tag{8.3}$$

Alternatively, equation 8.1 may be developed so that it appears as an advection term similar to

what is found in eq. 8.2 and 8.3 :

$$\frac{\partial \zeta}{\partial t} + \vec{U}.\nabla h + h\,\mathrm{div}\vec{U} = 0 \tag{8.4}$$

With $\left(\zeta, \vec{Q}\right)$ variables, we have now :

$$\frac{\partial \zeta}{\partial t} + \mathrm{div}\vec{Q} = 0 \tag{8.5}$$

$$\frac{\partial Q_x}{\partial t} + \frac{\partial uQ_x}{\partial x} + \frac{\partial vQ_x}{\partial y} + gh\frac{\partial \zeta}{\partial x} + f_b Q_x = S'_x \tag{8.6}$$

$$\underbrace{\frac{\partial Q_y}{\partial t}}_{\mathrm{I}} + \underbrace{\frac{\partial uQ_y}{\partial x} + \frac{\partial vQ_y}{\partial y}}_{\mathrm{II}} + \underbrace{gh\frac{\partial \zeta}{\partial y}}_{\mathrm{III}} + \underbrace{f_b Q_y}_{\mathrm{IV}} = \underbrace{S'_y}_{\mathrm{V}} \tag{8.7}$$

In the above equations,

- ∇ and div denote respectively the gradient and divergence operators;
- ζ (m) denotes the free surface elevation above a reference level;
  let z_b be the bottom elevation : the water-depth h is defined by $h = \zeta - z_b$;
  We are considering a fixed bottom which does not undergo any erosion or deposition process : consequently z_b is constant and $\frac{\partial \zeta}{\partial t} = \frac{\partial h}{\partial t}$.
- $\vec{U}$ (m/s) denotes the flow depth-averaged velocity, whose components along the $x-$ and $y-$ directions respectively are u and v;
- $\vec{Q}$ (m^2/s) denotes the unit discharge, whose components along the $x-$ and $y-$ directions respectively are $Q_x = hu$ and $Q_y = hv$;
- g ($\mathrm{m.s}^{-2}$) is the local gravity acceleration (usual value 9.81 $\mathrm{m.s}^{-2}$);
- S_x, S_y, S'_x and S'_y denote diffusion operators which complete development is given in 2.4.5. We have for instance

$$S_x = \frac{1}{h}\left[\frac{\partial}{\partial x}\left(hD_{xx}\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(hD_{yy}\frac{\partial u}{\partial y}\right) + \frac{\partial}{\partial x}\left(hD_{xy}\frac{\partial u}{\partial y}\right) + \frac{\partial}{\partial y}\left(hD_{yx}\frac{\partial u}{\partial x}\right)\right] \tag{8.8}$$

  where D_{xx}, D_{yy}, $D_{xy}\,(= D_{yx})$ denote appropriate dispersion coefficients.
- f_b (s^{-1}) is the so-called "friction factor", defined so that the bed shear stress $\vec{\tau}_b$ equates $\rho f_b \vec{Q}$ (ρ being the volumetric mass of the fluid) (cf 2.1.3). As previously mentioned (cf

2.1.3 & 2.4.5), it is usually expressed with the help of the Chezy or Strickler formula, namely

$$f_b = g\frac{\sqrt{u^2+v^2}}{C_h^2 h} = g\frac{\sqrt{Q_x^2+Q_y^2}}{C_h^2 h^2} \tag{8.9}$$

$$f_b = g\frac{\sqrt{u^2+v^2}}{K_s^2 h^{4/3}} = g\frac{\sqrt{Q_x^2+Q_y^2}}{K_s^2 h^{7/3}} \tag{8.10}$$

where the Chezy (C_h) or Strickler (K_s) coefficients are empirically related to the bed rugosity.

The five groups of terms mentioned in equations 8.1 to 8.7 have the following physical interpretation :

I Local flow acceleration (i.e. local variation of momentum with time).

II Transport of momentum by advection. (these terms are also called inertial terms)

III Conservation of mass (eq. 8.1 and 8.5) and momentum changes (eq. 8.2, 8.3, 8.6 and 8.7) due to wave propagation.

IV Momentum sink due to bed friction.

V Horizontal dispersion of momentum (due both to turbulence and vertical non-uniformities).

8.2 Physical analysis of the equations

Before dealing with the numerical solution of these flow equations, it may prove useful to analyse them briefly, mainly in order to estimate which are, according to the most salient features of the flow, the predominant terms among groups **II** to **V**. These predominant terms will naturally require our greatest care.

We shall follow the approach suggested in (Goutal, 1987), which consists first of rewriting the equations in non-dimensional form. For that purpose, we introduce different velocity, length and time scales : let V be a typical velocity modulus (for instance of the flow average velocity), H a typical water-depth (the mean one), L a typical length (related to the dimensions of the studied domain or to a characteristic wavelength). Given V and L we can define a typical time scale $T = L/V$.

From now on, the superscript $\tilde{}$ refers to dimensional quantities or operators.

The dimensional coordinates are defined as $\tilde{x} = \frac{x}{L}$ and $\tilde{y} = \frac{y}{L}$

from which we deduce the following relations between the differential operators and their dimensional form :

$$\widetilde{\mathrm{div}} = L\,\mathrm{div}, \quad \tilde{\nabla} = L\nabla, \quad \tilde{\Delta} = L^2\Delta$$

(where symbols div, ∇, Δ denote respectively the divergence, gradient and Laplacian operators).

The new time coordinate is $\tilde{t} = \dfrac{Vt}{L}$ and $\dfrac{\partial}{\partial\tilde{t}} = \dfrac{L}{V}\dfrac{\partial}{\partial t}$

The new dependent variables are :

$$\tilde{Z} = \frac{\zeta}{H}, \quad \tilde{h} = \frac{h}{H}, \quad \tilde{U} = \frac{\vec{U}}{V}$$

For the sake of simplicity we shall assume that the diffusion of momentum is isotropic and uniform and that terms S_x (eq. 8.2) and S_y (eq. 8.3) reduce respectively to $\nu\Delta u$ and $\nu\Delta v$ (where ν denotes the diffusivity). For the time being, we also denote $\vec{F}_f$ the friction force $\vec{F}_f = f_b\vec{U}$. Then, with the new coordinates and variables, equations 8.1 to 8.3 become respectively :

$$\frac{\partial\tilde{Z}}{\partial\tilde{t}} + \widetilde{\mathrm{div}}\left(\tilde{h}\tilde{U}\right) = 0 \tag{8.11}$$

$$\frac{\partial\tilde{U}}{\partial\tilde{t}} + \left(\tilde{U}.\tilde{\nabla}\right)\tilde{U} + \frac{gH}{V^2}\tilde{\nabla}\tilde{Z} - \frac{\nu}{VL}\tilde{\Delta}\tilde{U} = -\frac{L\vec{F}_f}{V^2} \tag{8.12}$$

In equation 8.12, there appear three dimensional products which determine the weight of the propagation, diffusion of momentum and friction terms with respect to the advective terms. These are respectively :

- for the propagation term, $p_1 = \dfrac{gH}{V^2}$

  Let us introduce the well-known Froude number $F_r = \dfrac{V}{\sqrt{gH}}$ which corresponds to the ratio of flow velocity to wave propagation celerity. We have simply $p_1 = 1/F_r^2$.

- for the diffusion term, $p_2 = \dfrac{\nu}{VL}$. The inverse of p_2 is analogous to a Reynolds number.

- for the friction term, p_3, whose expression we obtain by developing $\dfrac{L\vec{F}_f}{V^2}$ under the assumption that the friction is expressed with the help of the Strickler formula (cf eq. 8.10). Thus,

  $$\vec{F}_f = \frac{g\|\vec{U}\|}{K_s^2 h^{4/3}}\vec{U} = \frac{gV^2}{K_s^2 H^{4/3}}\frac{\|\tilde{U}\|\tilde{U}}{\tilde{h}^{4/3}}$$

  and the magnitude of the friction term is controlled by product

  $$p_3 = \frac{gL}{K_s^2 H^{4/3}}$$

(Goutal, 1987) studies both marine and coastal systems.

- In the first case, we deal with large and deep domains. (Goutal, 1987) suggests that their typical scales are respectively $H = 50$ to 150 m, $L = 50$ to 500 km and $V = 1$ to a few m/s. In such cases, the Froude number ranges between 10^{-1} and 10^{-2}. Thus, p_1 belongs to interval [100, 10000] : the propagation terms appear to be far more important than the convective ones.

  For a typical value $K_s = 50$ and the different possible combinations of L and H within the ranges indicated above, p_3 should vary approximately between 0.5 and 10. A closer look would be required in order to see if bottom friction influence can be completely discarded with respect to that of the propagation term. This is most probably the case in the deepest parts of the marine system. (nb : for such large systems however, it is no longer advisable to neglect the Coriolis force and wind induced surface stress in the flow equations, as their influence can quite outweigh that of the bottom friction).

  For a ν value of a few m.s^{-2}, p_2 is about 10^{-4} : horizontal diffusion of momentum has a very limited weight with respect to all the other terms in 8.12.

  In cases where the propagation is dominant, eq. 8.12 simplifies into

$$\frac{\partial \tilde{\vec{U}}}{\partial \tilde{t}} + \frac{1}{F_r^2} \tilde{\nabla} \tilde{Z} = -\frac{L \vec{F}_f}{V^2} \tag{8.13}$$

  Should we linearize eq. 8.11 (with $\tilde{h} \simeq 1$), the system made of 8.11 and 8.12 can further be manipulated. With appropriate differentiations and substitutions, we finally obtain (Goutal, 1987) that the original variables ζ and $\vec{U}$ satisfy :

$$\frac{\partial^2 \zeta}{\partial t^2} - gH \operatorname{div}(\nabla \zeta) = \ldots \tag{8.14}$$

$$\frac{\partial^2 \vec{U}}{\partial t^2} - gH \nabla \left(\operatorname{div} \vec{U} \right) = \ldots \tag{8.15}$$

  (the right-hand side terms of the above equations depend on the various sink and source terms, namely bed and surface stress, Coriolis force ...and have not been detailed for the sake of simplicity) ζ and $\vec{U}$ obey thus a similar wave equation, whose solutions propagate at celerity $\sqrt{gH}$.

- In coastal systems, the typical scales are fairly different : H would now range between 0.1 and 10 m, L between a hundred meters and a few kilometers, V between ten or so cm/s and 1 or 2 m/s. F_r would now lie between 0.1 and 2; it tends to increase as the flow depth diminishes. Consequently in coastal systems we can find again areas where the propagation term is the more important but also shallow zones where advection is definitely predominant over propagation. For water-depths of about one meter or less, p_3 has the same order of magnitude as p_1 so that the driving forces in the flow (i.e advection) are chiefly balanced by the dissipative bed friction. A correct treatment of these unfortunately non-linear terms will then be crucial.

In addition, we can say that in most rivers (e.g the Seine river), the flow is characterized by low Froude numbers (except possibly for flood conditions in small rivers, very steep channels, dam break, hydraulic discontinuities such as jumps, weirs, ...) and the dominant terms are the propagation and friction terms.

As pointed out in (Goutal, 1987), the relative influence of convection and propagation could also be appraised with the help of the method of characteristics (see (Abbott, 1979) (chapter 3), (Cunge *et al.*, 1980a; Pochat, 1989)). This method (which is not to be confused with backward characteristic methods applied to the solution of advection-dispersion equations !!) applies to one-dimensional nearly horizontal flows. It can be shown that such flows are characterized by invariants (the *Riemann invariants* of the fluid motion). These invariants travel across the fluid along two families of characteristic lines defined by celerities $c^{\pm} = u \pm \sqrt{gh}$. This result can be generalized to two-dimensional flows, provided that the momentum diffusion term is neglected. It appears that, according to the Froude number value, the celerities c^+ and c^- (and thus the transmission of informations relative to flow features) depend more or less on the wave celerity $\sqrt{gh}$ (low F_r) or on the flow velocity u (high F_r). These situations correspond to a dominance of propagation and convection respectively.

8.3 Outline of proposed models

Many finite difference approaches to the solution of the shallow-water equations have been suggested. The main distinctions between them are related to the following points :

- The St-Venant equations couple depth and velocity components. In order to reduce the complexity of the system, it may prove useful to uncouple these equations. Several strategies may fulfill that purpose.

- The St-Venant equations are non-linear. Solving them requires us to linearize them in some manner (cf appendix D and notably part D.5). Different methods are available for doing that.

- Similarly, splitting the equations may simplify their resolution. However, different kinds of splitting may be contemplated.

  Some numerical techniques are driven only by the desire to reduce the multidimensional problem to a sequence of one-dimensional problems, which discretization and solution can be handled more easily and efficiently. Examples of such techniques are various alternate direction (ADI) implicit methods (cf general introduction in appendix D.4.1 and examples of application in (Leendertse, 1970; Weare, 1979; Fairweather & Navon, 1980; Abbott *et al.*, 1981; Stelling *et al.*, 1986).)

Alternatively, process splitting may be favoured. It is based on a recognition of the different physical processes involved in the governing equations : the equations are split up into fractional steps (Yanenko, 1968), each one dealing with a single basic phenomenon. Numerical methods best suited to the physical and numerical features of each phenomenon can then be applied within each fractional step. Schemes falling in that category have been introduced notably by (Benque *et al.*, 1982; Dan N'Guyen, 1988; Wilders *et al.*, 1988). Space splitting can be used within each step.

Finally, some approaches blend approximately in the same proportion space and process splitting (Szymkiewicz, 1993).

- The different approaches to the uncoupling and splitting of the equations recommend different spatial discretizations. The dependent variables can indeed be defined on separate, intertwined grids (Benque *et al.*, 1982; Wilders *et al.*, 1988) : this is referred to as the use of staggered grids. Alternatively, variables can be defined at the same grid nodes (Dan N'Guyen, 1988; Dan N'Guyen, 1993; Szymkiewicz, 1993).

  Each approach has its pros and its cons, as summed up by (Szymkiewicz, 1993). The use of staggered grids can improve the accuracy of the spatial approximation to partial differences. On the other hand, it can be more complicated to handle and care must be exercised in order to maintain the consistency of boundary conditions applied at shifted locations.

We have been developing two models, each one corresponding to a different choice of dependent variables : the first model uses the free surface elevation and velocity components, the second one free surface elevation and the unit-width discharges. The governing equations are thus, in the first case, eq. 8.4, 8.2 and 8.3, in the second case, eq. 8.5 to 8.7. Both models are based first of all on process-splitting : the operators corresponding to advection (terms II), diffusion (terms V) and propagation-friction terms (groups III and IV) are dealt with successively.

In the following, Δt denotes the time step and superscripts n and $n+1$ refer to the values of the dependent variables at times t and $t+\Delta t$ respectively.

8.3.1 Advection step

8.3.1.1 Formulation $\left(\zeta, \vec{U} \right)$ (depth-velocity)

The relevant equations for this step have the form :

$$\frac{\partial f}{\partial t} + \vec{U} . \nabla f = 0 \quad \text{with } f = h, u, v \tag{8.16}$$

Equation 8.16 is solved between times t and $t + \Delta t$ with initial conditions $f(t) = f^n$. The corresponding solution at time $t + \Delta t$ is denoted $f^{n+1/3}$.

A proper resolution of the transport equation is obviously essential when advection is the dominant physical process governing the dependent variables. While the advective terms may have sometimes little influence on the hydraulic variables evolution (as they can be outweighted by friction or propagation terms, cf section 8.2), they are generally important in controlling the fate of dissolved species. The more the transported scalar distribution displays spatial variability, sharp gradients, the more challenging is the resolution of the transport equation. An example of such a tricky situation is the computation of the short-term fate of pollutants discharged by variable point-sources scattered all over the studied domain.

The preceding part of this report (chapters 4 to 7) has been devoted to the analysis of various methods for the solution of the advection-dispersion equation (ADE). We shall recall hereafter its main conclusions (cf section 7.5).

A first approach to the solution of the ADE is the use of backward characteristic methods (sec. 4.1.2). The performance of these methods depends on a proper backtracking of the fluid particles (which can be achieved rather straightforwardly with Runge-Kutta algorithms, cf appendix E.1.1) and on the application of an adequate spatial interpolator at the origin of each characteristic line. Use of a bicubic interpolator can result in very accurate forecasts, even in advection-dominated problems (cf chapters 6 and 7). These methods are rather expensive but can be run for a wide range of time steps, as they have no stability constraint.

Another approach is the use of eulerian finite difference methods. Their application is restricted to time steps which satisfy the Courant-Friedrich-Levy criterion (advective Courant number less than one). There, the estimation of advected fluxes (conservative form of the ADE) or of the gradient of the dependent variables (non-conservative form of the ADE) is generally based on a polynomial approximation of the transported scalar field, derived from its values at computational grid nodes. The use of a low-order approximation (e.g. upwind method) induces numerical damping, which is acceptable only in diffusion-dominated problems. Use of higher-order approximations consistently reduces the numerical dissipation. The improvement over the upwind method is already important with schemes such as QUICKEST or TAKACS, based on a quadratic approximation, and which involve only a moderate amount of extra computational effort. However, the price to pay for this larger precision is a greater tendency to instability. Indeed, approximations of order 2 and above do not preserve the monotonicity of the discrete data set. Thus, in sharp gradient areas, nonsensical over- or undershoots (e.g. negative concentrations) may be observed in the numerical solution. (*nb : such a phenomenon is also observed with the bicubic interpolator but less frequently and to a lesser extent*) If there is no diffusion or any other process acting to smooth out these errors, spurious wiggles develop and may lead

to a complete break down of the algorithm. In fact, eulerian methods which match the accuracy of characteristic methods are those which combine the use of high-order approximations and positiveness-enforcing limiters which control the estimation of advected fluxes (sec. 4.2.4 & 4.2.5). Unfortunately, the extension of such limiters to non-positive scalar fields (the velocity components, for instance) is not straightforward. Besides, as soon as the problem involves the transport of several variables, backward characteristic methods become no more costly than high-order eulerian methods. Indeed, in characteristic methods, the most expensive stage is by far the computation of the particle trajectories (cf sections 6.6 & 7.4) : if the different advected variables are defined at the same locations (i.e. on the same computational grid), this stage is common to them all, which results in significant savings.

In short, we favour the use of backward characteristic methods to deal with the advection-dispersion equation governing dissolved compounds. Bearing in mind that the flow model whose development we are presently discussing is bound to be run in combination with a transport model for chemical species, it appears that it is computationally efficient to solve its advection step with the same backward characteristic methods we use for geochemical variables. Indeed, as mentioned above, the costly stage which consists of particle backtracking is shared by all variables (by the way, this implies the use of a non-staggered grid) :

- The spatial interpolator applied at the origin of each characteristic is, for chemical species, the Hermite cubic interpolant (sec. 4.1.2.1), first-order derivatives at grid nodes being computed according to the Akima estimate (Williamson & Rasch, 1989; Rasch & Williamson, 1990).

- The same interpolator may be chosen for the hydraulic variables. However, as the advective terms are frequently less important within the flow equations, such degree of precision as provided with the bicubic interpolator is not always needed for the hydraulic variables. Besides, in some cases, for instance when computing flow in shallow areas, or dealing with flooding and drying situations, it is more essential to preserve the monotonicity of flow variables and overall the depth positiveness than to avoid dissipation (Goutal, 1987). A simple shape and positiveness preserving interpolator is the bilinear one. Consequently, the model has been written so that the user may choose between the bilinear and bicubic interpolators as regards the treatment of hydraulic variables.

- Finally, we have been adding a last possibility for the solution of the advection step : the simple upwind method. Application of this algorithm should be restricted to cases when the flow model is run alone (i.e. some of the test cases presented hereafter) and when the advective terms have limited influence on the flow evolution. Then, as we shall see later on, it results in considerable CPU time savings, despite the time step limitations that must be enforced in order to satisfy the Courant-Friedrich-Levy criterion.

We shall conclude by delivering a few words of warning : in spite of the fact that the solution technique suggested above has proven to be efficient and robust, it is still quite dependent on an adequate spatial and temporal discretization.

- When using the upwind method, the need to constrain the time step is obvious. On the other hand, backward characteristics methods are stable whatever the chosen time step. However, backtracking the fluid particles amounts theoretically to solve equations of the form :

$$\frac{\partial \vec{X}}{\partial t} = \vec{U}\left(\vec{X}, t\right) \tag{8.17}$$

  Yet, as we do not know the value of the velocity components at the next time step, we solve instead :

$$\frac{\partial \vec{X}}{\partial t} = \vec{U}^n\left(\vec{X}\right) \tag{8.18}$$

  If the chosen time step is too important with respect to the temporal variability of the velocity field, the solutions of 8.17 and 8.18 may thus be far apart.

- Secondly, if the computational grid is too coarse to provide a correct description of the depth and velocity fields, the spatial interpolation applied at each characteristic origin will provide a poor estimation of the advected variables, even if a bicubic interpolator is used.

8.3.1.2 Formulation $\left(\zeta, \vec{Q}\right)$ (depth-discharge)

The relevant equations are this time :

$$\frac{\partial f}{\partial t} + \frac{\partial uf}{\partial x} + \frac{\partial vf}{\partial y} = 0 \quad \text{with } f = Q_x, Q_y \tag{8.19}$$

This conservative form of the transport equation may be solved using an eulerian finite difference flux-form method, such as those discussed in previous section. However, another alternative has been suggested in (Benque *et al.*, 1982). Let us develop eq. 8.19 for variable Q_x. We obtain :

$$h\left(\frac{\partial u}{\partial t} + \vec{U}.\nabla u\right) + u\left(\frac{\partial h}{\partial t} + \mathrm{div}\vec{Q}\right) = 0 \tag{8.20}$$

As the flow field satisfies the continuity equation (eq. 8.5), the second-term above disappears and 8.20 turns out to have the same form as eq. 8.16. The solution suggested by (Benque *et al.*, 1982) is the following : compute the velocities $u^{n+1/3}$ and $v^{n+1/3}$ solution of 8.20, then compute the advected unit-width discharges by $Q_x^{n+1/3} = h^n u^{n+1/3}$ and $Q_y^{n+1/3} = h^n v^{n+1/3}$. It is straightforward to check that the difference between $\vec{Q}^{n+1/3}$, estimated as indicated above, and the ideal $\vec{Q}^*$ which would be the solution of eq. 8.19 is :

$$\vec{Q}^* - \vec{Q}^{n+1/3} = \left(\Delta t \frac{\partial h}{\partial t}\right).\vec{U} \simeq \left(h^{n+1} - h^n\right).\vec{U}^n$$

At this stage of the computation h^{n+1} is unknown so that it is not possible to correct the estimate $\vec{Q}^{n+1/3}$. Intuitively, it appears that the relevance of the discrepancy between estimate $\vec{Q}^{n+1/3}$ and exact $\vec{Q}^*$ depends on the relative importance of spatial (velocity) and temporal (depth) gradients. Generally, this cannot be evaluated a priori. Yet, it can be controlled on-line throughout the simulation. Lastly, we may notice that, if the model is run to determine an equilibrium steady-state flow, the error involved in the computation of the advected unit-width discharges should be vanishing.

(Dan N'Guyen, 1988) has been proposing another approach. He simplifies eq. 8.19 into :

$$\frac{\partial f}{\partial t} + \vec{U}.\nabla f = 0 \quad \text{with } f = Q_x, Q_y \tag{8.21}$$

thus discarding the influence of terms $Q_x \,\mathrm{div}\vec{U}$ and $Q_y \,\mathrm{div}\vec{U}$ respectively. Let $h^{n+1/3}$ be the solution of eq. 8.16 for $f = h$ (i.e. the advected depth). By a development of eq. 8.19, it can be shown that the difference between its exact solution $\vec{Q}^*$ and $\vec{Q}^{n+1/3}$ satisfying 8.21 is :

$$\vec{Q}^* - \vec{Q}^{n+1/3} \simeq \left(h^{n+1} - h^{n+1/3} \right).\vec{U}^n$$

Consequently, even in case of steady-state flow, this error does not disappear, unless the water-depth is uniform over the computational domain.

Once the problem of solving eq. 8.19 has been transformed into the problem of solving equation 8.16, either for the velocity components or for the unit-discharges (Dan N'Guyen's simplification), the methods suggested in conclusion of the previous section 8.3.1.1 can be applied.

8.3.2 Diffusion step

This step concerns only the velocity or discharge components.

According to expression 8.8, the working equations of this stage for the "depth/velocity" formulation have the form :

$$\begin{aligned}\frac{\partial f}{\partial t} = \frac{1}{h}\Big[&\frac{\partial}{\partial x}\left(hD_{xx}\frac{\partial f}{\partial x}\right) + \frac{\partial}{\partial y}\left(hD_{yy}\frac{\partial f}{\partial y}\right) \\ &+ \frac{\partial}{\partial x}\left(hD_{xy}\frac{\partial f}{\partial y}\right) + \frac{\partial}{\partial y}\left(hD_{yx}\frac{\partial f}{\partial x}\right)\Big]\end{aligned} \tag{8.22}$$

where f refers to both u and v.

When working with a "depth/discharge" formulation, the expressions used for S'_x and S'_y are somewhat simpler than their respective counterparts S_x and S_y (Benque *et al.*, 1982) so that the governing equations of the diffusion step read :

$$\frac{\partial f}{\partial t} = \left[\frac{\partial}{\partial x}\left(D_{xx}\frac{\partial f}{\partial x}\right) + \frac{\partial}{\partial y}\left(D_{yy}\frac{\partial f}{\partial y}\right) + \frac{\partial}{\partial x}\left(D_{xy}\frac{\partial f}{\partial y}\right) + \frac{\partial}{\partial y}\left(D_{yx}\frac{\partial f}{\partial x}\right)\right] \tag{8.23}$$

where f refers to Q_x or Q_y.

Equations 8.22 or 8.23 are solved between time t and $t + \Delta t$ with initial conditions $f(t) = f^{n+1/3}$. Their solution is marked out with superscript $n + 2/3$.

The solution technique is a semi-implicit, alternate direction method. The corresponding implicitation parameter is denoted θ. Spatial derivatives are developed into finite differences according to the classical Crank-Nicholson approach. The resulting finite difference equation approximates the original partial difference equation with a second-order error in space when the grid is uniform, with a first-order error otherwise. Full details about the method, which boils down to the solution of a series of tridiagonal linear systems, can be found in appendix E (sections E.2.1 & E.2.2). The models also include the possibility of dealing explicitly with the diffusion step. This more economical option has stability requirements (cf E.2) and should preferably be used only when the influence of the diffusion terms is small. Indeed, the truncation error of the finite difference approximations to the diffusion step equations is a first-order error in time unless $\theta = 0.5$. If the diffusive terms are important, it is then advisable to approximate them as accurately as possible, which leads a choice of θ close to 0.5.

8.3.3 Propagation-friction step

(a) Formulation $\left(\zeta, \vec{U}\right)$ (depth-velocity)

The governing equations of this stage are the remains of the continuity (eq. 8.4) and momentum equations (eq. 8.2 and 8.3) :

$$\frac{\partial h}{\partial t} + h\,\mathrm{div}\vec{U} = 0 \tag{8.24}$$

$$\frac{\partial \vec{U}}{\partial t} + g\nabla\zeta + f_b\vec{U} = 0 \tag{8.25}$$

These are solved between t and $t + \Delta t$ with initial conditions $f(t) = f^{n+2/3}$ (nb : $f = u$, v, h and $h^{n+2/3} = h^{n+1/3}$). This means that the time derivatives in eq. 8.24 and 8.25 are approximated as

$$\frac{\partial h}{\partial t} = \frac{h^{n+1} - h^{n+1/3}}{\Delta t}$$

$$\frac{\partial \vec{U}}{\partial t} = \frac{\vec{U}^{n+1} - \vec{U}^{n+2/3}}{\Delta t}$$

By summing eq. 8.16, 8.22, 8.24 and 8.25, it is straightforward to check that the solutions of this step effectively correspond to the new state of hydraulic variables at time $t + \Delta t$.

(b) Formulation $\left(\zeta, \vec{Q} \right)$ (depth-discharge)

The continuity equation has not been split and the governing equations read this time :

$$\frac{\partial \zeta}{\partial t} + \mathrm{div}\vec{Q} = 0 \tag{8.26}$$

$$\frac{\partial \vec{Q}}{\partial t} + g h \nabla \zeta + f_b \vec{Q} = 0 \tag{8.27}$$

These need to be solved between t and $t + \Delta t$, the time derivatives corresponding respectively to :

$$\frac{\partial \zeta}{\partial t} = \frac{\zeta^{n+1} - \zeta^n}{\Delta t}$$

$$\frac{\partial \vec{Q}}{\partial t} = \frac{\vec{Q}^{n+1} - \vec{Q}^{n+2/3}}{\Delta t}$$

As regards the solution of the propagation step, we proceed as follows :

1. Eq. 8.24 and 8.25 (respectively eq. 8.26 and 8.27) are developed in a semi-implicit fashion, introducing implicitation parameter γ ($0 \leq \gamma \leq 1$).

2. First, as done in (Benque *et al.*, 1982; Wilders *et al.*, 1988; Dan N'Guyen, 1988; Dan N'Guyen, 1993), we uncouple equation 8.24 and 8.25 (resp. eq. 8.26 and 8.27).

   - Eq. 8.25 (respectively eq. 8.27) is linearized so that the unknown variable $\vec{U}^{n+1}$ (resp. $\vec{Q}^{n+1}$) is expressed as a function of past values of the variables, ζ^n, $\vec{U}^n$ (resp. $\vec{Q}^n$), intermediate values $\vec{U}^{n+2/3}$ (resp. $\vec{Q}^{n+2/3}$) and variable ζ^{n+1}.
   - This allows us to eliminate $\vec{U}^{n+1}$ (resp. $\vec{Q}^{n+1}$) from eq. 8.24 (resp. eq. 8.26) which becomes an equation in the single unknown ζ. This equation bears some resemblance to a Poisson equation. Through space-splitting, its solution can be reduced to the problem of solving one-dimensional tridiagonal linear systems along each coordinate line.

3. Once ζ^{n+1} is computed, it is back-substituted into 8.25 (respectively 8.27) in order to update the velocity (discharge).

The procedure outlined above will now be fully detailed.

8.4 Solution of the propagation-friction step

8.4.1 Development of the equations

8.4.1.1 Formulation $\left(\zeta, \vec{U}\right)$ (depth-velocity)

Let δz denote the increment of the free surface between t and $t + \Delta t$: as the bottom elevation is fixed, $\delta z = \zeta^{n+1} - \zeta^n = h^{n+1} - h^n$. Similarly, we define $\delta\vec{U} = \vec{U}^{n+1} - \vec{U}^n$. In order to develop eq. 8.24 and 8.25, we introduce the implicitation parameter γ.

Eq. 8.24 reads now :

$$\delta z - \left(h^{n+1/3} - h^n\right) + \gamma\,\Delta t\left(h\,\mathrm{div}\vec{U}\right)^{n+1} + (1-\gamma)\,\Delta t\left(h\,\mathrm{div}\vec{U}\right)^n = 0 \qquad (8.28)$$

Introducing δz and $\delta\vec{U}$, the non-linear product $\left(h\,\mathrm{div}\vec{U}\right)^{n+1}$ is linearized into :

$$h^n\,\mathrm{div}\vec{U}^n + h^n\,\mathrm{div}\delta\vec{U} + \delta z\,\mathrm{div}\vec{U}^n$$

(cross-product of increments δz and $\delta\vec{U}$ is neglected). Consequently, eq. 8.24 finally becomes :

$$\left(1 + \gamma\,\Delta t\,\mathrm{div}\vec{U}^n\right)\delta z + \gamma\,\Delta t\,h^n\,\mathrm{div}\delta\vec{U} = \left(h^{n+1/3} - h^n\right) - \Delta t\left(h\,\mathrm{div}\vec{U}\right)^n \qquad (8.29)$$

The semi-implicit development of eq. 8.25 is :

$$\vec{U}^{n+1} - \vec{U}^{n+2/3} + g\gamma\,\Delta t\,\nabla\zeta^{n+1} + g\,(1-\gamma)\,\Delta t\,\nabla\zeta^n + \gamma\,\Delta t\left(f_b\vec{U}\right)^{n+1} + (1-\gamma)\Delta t\left(f_b\vec{U}\right)^n = 0$$

The non-linear product $\left(f_b\vec{U}\right)^{n+1}$ is linearized on the basis of the "frozen coefficient" approximation (Wilders *et al.*, 1988). This means we set simply $\left(f_b\vec{U}\right)^{n+1} \simeq f_b^n\vec{U}^{n+1}$. Thus, eq. 8.25 can be reduced to :

$$(1 + \gamma\,f_b^n\,\Delta t)\,\delta\vec{U} + g\gamma\,\Delta t\,\nabla\delta z = \left(\vec{U}^{n+2/3} - \vec{U}^n\right) - \Delta t\left(f_b\vec{U}\right)^n - g\,\Delta t\,\nabla\zeta^n \qquad (8.30)$$

Eq. 8.29 and 8.30 correspond to approximations of 8.24 and 8.25 respectively, which are second-order accurate in time if $\gamma = 1/2$, first-order accurate otherwise. For the sake of brevity, the right-hand side terms of equations 8.29 and 8.30 will now be denoted $\mathcal{H}$ and $\vec{\mathcal{W}}$ respectively.

Eq. 8.30 allows us to compute the divergence of $\delta\vec{U}$. A subsequent substitution into eq. 8.29 yields :

$$\left(1 + \gamma\,\Delta t\,\mathrm{div}\vec{U}^n\right)\delta z - g\,(\gamma\,\Delta t)^2\,h^n\,\mathrm{div}\left(\frac{1}{1+\gamma\,f_b\,\Delta t}\nabla\,\delta z\right) = \mathcal{H} - (\gamma\Delta t)\,h^n\,\mathrm{div}\frac{\vec{\mathcal{W}}}{1+\gamma\,f_b\,\Delta t}$$

Further manipulation of the right-hand side term finally leads to :

$$\delta z - \frac{g\,(\gamma\,\Delta t)^2\,h^n}{1+\gamma\,\Delta t\,\mathrm{div}\vec{U}^n}\,\mathrm{div}\left(\frac{1}{1+\gamma f_b^n \Delta t}\,\nabla\,\delta z\right) = \frac{1}{1+\gamma\,\Delta t\,\mathrm{div}\vec{U}^n}\left(h^{n+1/3} - h^n - \Delta t\,h^n\,\mathrm{div}\frac{\vec{B}}{1+\gamma f_b^n \Delta t}\right) \quad (8.31)$$

with

$$\vec{B} = (1-\gamma)\,\vec{U}^n + \gamma\,\vec{U}^{n+2/3} - g\,(\gamma\Delta t)\,\nabla\zeta^n$$

In case the spatial variation of the factor $1/(1+\gamma\,f_b\,\Delta t)$ may be neglected, 8.31 becomes :

$$\delta z - \frac{g\,(\gamma\,\Delta t)^2\,h}{(1+\gamma\,f_b\,\Delta t)\,(1+\gamma\,\Delta t\,\mathrm{div}\vec{U})}\,\Delta\,(\delta z) = \frac{1}{1+\gamma\,\Delta t\,\mathrm{div}\vec{U}}\left(h^{n+1/3} - h^n - \frac{\Delta t\,h}{1+\gamma\,f_b\,\Delta t}\,\mathrm{div}\vec{B}\right) \quad (8.32)$$

where the symbol Δ alone stands for the Laplace operator. This is similar to a Poisson equation.

8.4.1.2 Formulation $\left(\zeta, \vec{Q}\right)$ (depth-discharge)

The semi-implicit development of equations 8.26 and 8.27 are respectively :

$$\delta z + \gamma\,\Delta t\,\mathrm{div}\vec{Q}^{n+1} + (1-\gamma)\,\Delta t\,\mathrm{div}\vec{Q}^n = 0 \quad (8.33)$$

$$\vec{Q}^{n+1} - \vec{Q}^{n+2/3} + g\,\gamma\,\Delta t\,(h\nabla\zeta)^{n+1} + g\,(1-\gamma)\,\Delta t\,(h\nabla\zeta)^n + \gamma\,\Delta t\left(f_b\vec{Q}\right)^{n+1} + (1-\gamma)\Delta t\left(f_b\vec{Q}\right)^n = 0 \quad (8.34)$$

The non-linear products $(h\nabla\zeta)^{n+1}$ and $\left(f_b\vec{Q}\right)^{n+1}$ in eq. 8.34 are respectively linearized as $h^n\nabla\zeta^{n+1}$ and $f_b^n\vec{Q}^{n+1}$ so that eq. 8.34 reduces to :

$$(1+\gamma\,f_b^n\,\Delta t)\,\vec{Q}^{n+1} + g\,(\gamma\,\Delta t)\,h^n\nabla\delta z = \vec{Q}^{n+2/3} - (1-\gamma)\,\Delta t\left(f_b\vec{Q}\right)^n - g\,\Delta t\,h^n\,\nabla\zeta^n \quad (8.35)$$

We proceed as for the "depth/velocity" formulation : relation 8.35 allows us to compute the divergence of $\vec{Q}^{n+1}$, which we then substitute into eq. 8.33.

We find that δz should satisfy equation :

$$\delta z - g\,(\gamma\,\Delta t)^2\,\mathrm{div}\left(\frac{h^n}{1+\gamma\,f_b^n\,\Delta t}\,\nabla\,\delta z\right) = -\Delta t\,\mathrm{div}\,\frac{\vec{B}}{1+\gamma\,f_b^n\,\Delta t} \quad (8.36)$$

with

$$\vec{B} = (1-\gamma)\,\vec{Q}^n + \gamma\,\vec{Q}^{n+2/3} - g\,(\gamma\Delta t)\,(h\,\nabla\zeta)^n$$

Neglecting the spatial variations of factor $1/(1+\gamma f_b \Delta t)$ leads to solve for δz equation :

$$\delta z - \frac{g(\gamma \Delta t)^2}{1+\gamma f_b^n \Delta t} \mathrm{div}(h^n \nabla \delta z) = -\frac{\Delta t}{1+\gamma f_b^n \Delta t} \mathrm{div}\vec{B} \quad (8.37)$$

Dan N'Guyen (Dan N'Guyen, 1988) further simplifies 8.37 into :

$$\delta z - \frac{g(\gamma \Delta t)^2 h^n}{1+\gamma f_b^n \Delta t} \Delta \delta z = -\frac{\Delta t}{1+\gamma f_b^n \Delta t} \left[\mathrm{div}\left((1-\gamma)\vec{Q}^n + \gamma \vec{Q}^{n+2/3} \right) - g(\gamma \Delta t)(h \Delta \zeta)^n \right] \quad (8.38)$$

8.4.2 Validity of the developed equations

Apart from the model suggested by Dan N'Guyen (Dan N'Guyen, 1988; Dan N'Guyen, 1993), other models rely on the approach presented in sections 8.4.1.1 and 8.4.1.2, namely deriving an equation whose only unknown is the free-surface elevation : equations 8.31, 8.32, 8.36, 8.37 have the same degree of complexity as eq. 8.38 used by Dan N'Guyen and are somewhat simpler than equations obtained by (Benque *et al.*, 1982) and especially by (Wilders *et al.*, 1988).

In short, the ability of equations 8.31 and 8.36 to describe faithfully the evolution of the free surface depends mainly on whether the time step is correctly chosen. An inadequate choice of the time step may lead the model either to diverge either to produce a solution with seemingly acceptable features but which bears little resemblance to the true flow field.

First, the time step controls the magnitude of the truncation error between the original working equations of the propagation step and their semi-implicit developments. This truncation error is also dependent on the choice of implicitation parameter γ. Ensuring stability of the computation generally requires one to pick γ strictly superior to its "optimal" value 0.5 which allows the approximation of the propagation step equations to be second-order accurate in time. Consequently, the error is most often a first-order function of Δt. It cannot be quantified a priori, due to the complexity of the working equations. It will of course be all the more important in the case of quickly varying flows.

Secondly, equations 8.31 or 8.36 are derived by **linearizing** the semi-implicit developments of the propagation step equations. In fact, as explained in appendix D.5, there are basically two approaches to the problem of solving non-linear systems. The simplest one consists of linearizing the system with the help of values of the variables at the preceding time-level. This is how most classic and also some recent (Dan N'Guyen, 1988) ADI methods proceed; this is what we have done in sections 8.4.1.1 and 8.4.1.2. Then, the linearized system is solved in one single stage. The second approach consists of approximating iteratively the system and its related solution

by applying methods such as quasi-Newton or gradient methods (Benque *et al.*, 1982; Wilders *et al.*, 1988; Szymkiewicz, 1993).

When models are used to compute an equilibrium steady-steate flow, the linearization error is naturally vanishing as we converge towards the solution. **On the other hand, when dealing with unsteady flows, we may intuitively expect the first approach (i.e. direct, one-step linearization) to be valid only below some critical time step beyond which the applied linearization becomes too crude.** (This limitation may worsen limitations already owed to the treatment of irregular boundaries or bathymetries). On the contrary, it has been proved that the second kind of algorithms keeps on providing accurate results for large time steps (Benque *et al.*, 1982; Wilders *et al.*, 1988) ... but at the expense of an increase in the adequate number of sub-iterations.

Because of the repetition of sub-iterations, the second group of methods are more costly than the first group of methods for going from one time level to the next. On the other hand, they can be run with bigger time steps, thus decreasing the number of time levels needed within a whole simulation. Since solving the propagation step is just one stage of the computation, the costs ratio per time step is not simply proportional to the number of sub-iterations performed, but inferior to it. The effectiveness of iterative methods depend thus on the balance between time consumed in solving advection and diffusion stages and time devoted to multiple resolutions of the propagation operators. The required number of sub-iterations is problem dependent. (Benque *et al.*, 1982) for instance state that for the CYTHERE model it is controlled by the magnitude of C_p the local Courant number for wave propagation (i.e. calculated from the wave celerity) defined as

$$C_p = \sqrt{g\,h}\,\Delta t\sqrt{\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2}}$$

where Δx and Δy are the gridsize respectively in the $x-$ and $y-$directions. Only a few iterations are needed when $C_p \leq 10$ but typically 15 to 50 iterations are required for $20 \leq C_p \leq 40$. Consequently, the relative efficiency (from a computational point of view) of one-step and sub-iterative methods is problem-dependent too ... and can be difficult to guess "a priori" !

8.4.3 Alternatives for the solution of the free-surface increment equation

In summary, equations derived for the free-surface increment involve generally a two-dimensional differential operator which appears to be some perturbation of the Laplacian operator with non-constant coefficients. Some authors deal directly with this multidimensional system, e.g. (Wilders *et al.*, 1988) who make use of a preconditioned Conjugate Gradients method to solve it (cf general description in D.3.1). However, most researchers apply space-splitting. Different approaches are possible. They belong mostly to the category of alternate direction

implicit methods, first introduced by (Leendertse, 1970) and since then widely used (Weare, 1979; Fairweather & Navon, 1980; Stelling *et al.*, 1986), or alternatively rely on an approximate factorization of the operator, as illustrated in (Dan N'Guyen, 1988). There, equation 8.38 is rewritten as

$$\left(I - \frac{g\,(\gamma\,\Delta t)^2\,h}{1+\gamma\,f_b\,\Delta t}\,\frac{\partial^2}{\partial x^2} \right) \circ \left(I - \frac{g\,(\gamma\,\Delta t)^2\,h}{1+\gamma\,f_b\,\Delta t}\,\frac{\partial^2}{\partial y^2} \right) \delta z \;=\; \mathcal{B} \tag{8.39}$$

and is then solved in two steps :

$$\left(I - \frac{g\,(\gamma\,\Delta t)^2\,h}{1+\gamma\,f_b\,\Delta t}\,\frac{\partial^2}{\partial x^2} \right) \delta z^* \;=\; \mathcal{B} \tag{8.40}$$

$$\left(I - \frac{g\,(\gamma\,\Delta t)^2\,h}{1+\gamma\,f_b\,\Delta t}\,\frac{\partial^2}{\partial y^2} \right) \delta z \;=\; \delta z^* \tag{8.41}$$

Each step involves only one-dimensional (tridiagonal) linear systems.

The above factorization is accurate if and only if term

$$\left(\frac{g\,(\gamma\,\Delta t)^2\,h}{1+\gamma\,f_b\,\Delta t} \right)^2 \frac{\partial^4\,(\delta z)}{\partial x^2 \partial y^2}$$

remains negligible with respect to terms involving the second-order space derivatives of the free surface elevation. As indicated in (Dan N'Guyen, 1993), this term is approximately proportional to the fourth power C_p^4. Thus, instability and accuracy loss may be expected for too large C_p. Dan n'Guyen (Dan N'Guyen, 1993) points out that, as the friction factor appears in the denominator of this truncation error term, increased friction should play a significant part in alleviating the inaccuracies eventually raised by the truncation.

Besides, the criteria

$$\left(\frac{g\,(\gamma\,\Delta t)^2\,h}{1+\gamma\,f_b\,\Delta t} \right)^2 \frac{\partial^4\,(\delta z)}{\partial x^2 \partial y^2} \ll \frac{g\,(\gamma\,\Delta t)^2\,h}{1+\gamma\,f_b\,\Delta t}\,\frac{\partial^2 \delta z}{\partial x^2} \quad \text{and} \quad \left(\frac{g\,(\gamma\,\Delta t)^2\,h}{1+\gamma\,f_b\,\Delta t} \right)^2 \frac{\partial^4\,(\delta z)}{\partial x^2 \partial y^2} \ll \frac{g\,(\gamma\,\Delta t)^2\,h}{1+\gamma\,f_b\,\Delta t}\,\frac{\partial^2 \delta z}{\partial y^2}$$

can be rewritten as

$$\frac{g\,(\gamma\,\Delta t)^2\,h}{1+\gamma\,f_b\,\Delta t}\,\frac{\partial^2 \zeta}{\partial y^2} \ll \zeta \quad \text{and} \quad \frac{g\,(\gamma\,\Delta t)^2\,h}{1+\gamma\,f_b\,\Delta t}\,\frac{\partial^2 \zeta}{\partial x^2} \ll \zeta \tag{8.42}$$

This highlights the fact that the validity of the factorization is also dependent on the state of the free surface : should it display too strong a curvature, the factorization may not be advisable.

Schemes relying on straightforward directional separation (ADI methods) have been shown to suffer from some typical shortcomings : when the time step is too large, the directional separation may induce spurious polarisations along the axes of discretization or may result in an inexact handling of the boundary conditions. Such troubles are particularly salient when dealing with a computational domain whose geometry or bathymetry is irregular : they have notably

been discussed by (Weare, 1979; Benque *et al.*, 1982; Stelling *et al.*, 1986; Wilders *et al.*, 1988). Both the theoretical analysis provided in (Weiyan, 1992) (chapter 9) and the various benchmark tests performed in (Dan N'Guyen, 1988; Dan N'Guyen, 1993) or (Appere, 1988) suggest that approximate factorization has a domain of application significantly wider and a greater accuracy than classic space splitting methods.

In order to eliminate the error introduced by directional splitting (and, subsequently, to allow for the use of bigger time steps), it is possible to use the technique of sub-iteration. This means that an iterative process is embedded within each time step.

- An example of iterative improvement of the ADI approach is provided in (Benque *et al.*, 1982) where the δz equation is rewritten as :

$$(I - \mathcal{L}_x - \mathcal{L}_y) \delta z = \mathcal{B}_x + \mathcal{B}_y \tag{8.43}$$

where $\mathcal{L}_x$ and $\mathcal{L}_y$ denote respectively differential operators along the $x-$ and $y-$ directions only and $\mathcal{B}_x$ and $\mathcal{B}_y$ are similarly decompositions of the right-hand side term according to $x-$ and $y-$ directions. Then, (Benque *et al.*, 1982) solve iteratively 8.43 using an alternating direction operator with coordination. At each iteration, the following one-dimensional systems are solved :

$$(I - \mathcal{L}_x) \delta z_1 = \mathcal{B}_x - p \tag{8.44}$$

$$(I - \mathcal{L}_y) \delta z_2 = \mathcal{B}_y + p \tag{8.45}$$

I denotes the identity operator and p the coordination term (a Lagrange multiplier). The problem now consists of finding the value of p for which $\delta z_1 = \delta z_2 = \delta z$. p may for instance be initialized to its value at the end of the previous time step. Whatever the chosen initialization, p is then optimized using the associated gradient $\delta z_1 - \delta z_2$ (cf appendix D.4.2). This approach of *decomposition with coordination* retains some of the simplicity of classic alternate direction methods as it involves only the solution of one-dimensional systems. Its performance appears both to outweigh those of ADI algorithms (Benque *et al.*, 1982) and to match those of more complex methods which do not apply space splitting (Wilders *et al.*, 1988).

- Similarly, the methodology to improve approximate factorization methods is indicated in (Giles, 1989), based on the example of the Navier-Stokes equations. These can be expressed in the following mathematical form :

$$\frac{\partial U}{\partial t} + \frac{\partial F}{\partial x} + \frac{\partial G}{\partial y} = 0 \tag{8.46}$$

where U denotes the flow velocity and F and G momentum fluxes accounting for convection, viscosity and pressure effects, ... (cf 2.1.1). Eq. 8.46 can be approximated in a fully

implicit fashion by using backward Euler time discretization :

$$\frac{U^{n+1}-U^n}{\Delta t} + \frac{\partial F^{n+1}}{\partial x} + \frac{\partial G^{n+1}}{\partial y} = 0 \tag{8.47}$$

Eq. 8.47 stands as a non-linear system of equations for U at the new time-level $n+1$. The next step is to linearize it, writing :

$$F^{n+1} \simeq F^n + A\,\Delta U \quad \text{where } A \equiv \frac{\partial F}{\partial U} \text{ and } \Delta U = U^{n+1} - U^n$$

Linearizing G in a similar manner yields the following scheme :

$$\left(I + \Delta t \frac{\partial}{\partial x} A + \Delta t \frac{\partial B}{\partial y} \right) \Delta U = -\Delta t \frac{\partial F^n}{\partial x} - -\Delta t \frac{\partial G^n}{\partial y} \tag{8.48}$$

Applying approximate factorization amounts to rewrite eq. 8.48 as

$$\left(I + \Delta t \frac{\partial}{\partial x} A \right) \circ \left(I + \Delta t \frac{\partial}{\partial y} B \right) \Delta U = -\Delta t \frac{\partial F^n}{\partial x} - -\Delta t \frac{\partial G^n}{\partial y} \tag{8.49}$$

The factorization error $\left(\Delta t^2 \frac{\partial}{\partial x} A \frac{\partial}{\partial y} B\, \Delta U \right)$ can be reduced as follows :

- The process begins by setting $\Delta U = 0$
- The next step is to calculate δU using the following factorized discrete equation

$$\left(I + \Delta t \frac{\partial}{\partial x} A \right) \circ \left(I + \Delta t \frac{\partial}{\partial y} B \right) \delta U = -\Delta t \frac{\partial}{\partial x}(F^n + A\Delta U) - \Delta t \frac{\partial}{\partial y}(G^n + B\Delta U) \tag{8.50}$$

- The third step consists of defining an improved value of ΔU by

$$\Delta U^{\text{new}} = \Delta U^{\text{old}} + \delta U$$

One then can iterate by repeating the last two steps as often as one wishes. Iterating only once gives the standard factorization algorithm. Iterating many times converges to the true solution to original system 8.48. It has been reported that, in practice, iterating three times eliminates most of the factorization error (Giles, 1989).

Obviously, iterative processes aiming at eliminating directional splitting errors can be combined with iterative linearization of the operator.

Considering its computational efficiency, we have decided to apply the factorization method suggested in (Dan N'Guyen, 1988), (Dan N'Guyen, 1993).

The proper resolution of the propagation step can be checked as indicated below.
Once δz has been computed and $\vec{U}^{n+1}$ (respectively $\vec{Q}^{n+1}$) consequently updated, we may back-substitute $\vec{U}^{n+1}$ (resp. $\vec{Q}^{n+1}$) into the original semi-implicit development of the continuity

equation (respectively relations 8.28 and 8.33). Continuity is satisfied for increment $(\delta z)^*$ so that,

$$\left(1 + \gamma \Delta t \, \mathrm{div}\vec{U}^{n+1} \right) (\delta z)^* = h^{n+1/3} - h^n - \Delta t \, h^n \left(\gamma \mathrm{div}\vec{U}^{n+1} + (1-\gamma)\mathrm{div}\vec{U}^n \right) \quad (8.51)$$

or ("depth/discharge" formulation) :

$$(\delta z)^* = -\gamma \, \Delta t \, \mathrm{div}\vec{Q}^{n+1} - (1-\gamma) \, \Delta t \, \mathrm{div}\vec{Q}^n \quad (8.52)$$

The occurence of significant discrepancies between $(\delta z)^*$ and the computed δz means that either linearization or directional separation was improper, or that spurious numerical errors, round-offs, occurred when solving the one-dimensional systems produced by the decomposition. **The resolution accuracy may be improved by the combined implementation of an iterative elimination of the factorization errors and of an iterative linearization of the governing equations**. That topic will be discussed in next section.

Table 8.1: Summary of proposed flow models

Formulation "depth/velocity"	**Formulation "depth/unit discharge"**
$h^n, \vec{U}^n \Rightarrow h^{n+1/3}, \vec{U}^{n+1/3}$ **Advection stage**	$\vec{Q}^n \Rightarrow \vec{Q}^{n+1/3}$
Governing equation : $\frac{\partial f}{\partial t} + \vec{U}^n.\nabla f = 0$ with $f = h, u, v$ Solution : 3 options - Backward characteristic method with bicubic or bilinear interpolator - Upwind method	Governing equation : $\frac{\partial f}{\partial t} + \frac{\partial uf}{\partial x} + \frac{\partial vf}{\partial y} = 0$ with $f = Q_x, Q_y$ developed into : $\frac{\partial f}{\partial t} + \vec{U}^n.\nabla f = 0$ with $f = u, v$ Solution : - Advection of velocity components as indicated in left column - Computation of advected discharges by : $Q_x^{n+1/3} = h^n u^{n+1/3}$ and $Q_y^{n+1/3} = h^n v^{n+1/3}$
$\vec{U}^{n+1/3} \Rightarrow \vec{U}^{n+2/3}$ **Diffusion stage**	$\vec{Q}^{n+1/3} \Rightarrow \vec{Q}^{n+2/3}$
Semi-implicit development of the diffusion operator Spatial discretization by Crank-Nicholson method Directional splitting into tridiagonal one-dimensional systems	
$h^{n+1/3}, \vec{U}^{n+2/3} \Rightarrow h^{n+1}, \vec{U}^{n+1}$ **Propagation/friction stage**	$h^n, \vec{Q}^{n+2/3} \Rightarrow h^{n+1}, \vec{Q}^{n+1}$
(a) Semi-implicit development and linearization of the remains of momentum equations (nb : $\delta z = \zeta^{n+1} - \zeta^n = h^{n+1} - h^n$)	
$(1 + \gamma f_b^n \Delta t)\, \vec{U}^{n+1} + g\,(\gamma \Delta t)\, \nabla \delta z =$ $\vec{U}^{n+2/3} - (1-\gamma)\Delta t \left(f_b \vec{U}\right)^n - g \Delta t\ (\nabla \zeta)^n$	$(1 + \gamma f_b^n \Delta t)\, \vec{Q}^{n+1} + g\,(\gamma \Delta t)\, h^n \nabla \delta z =$ $\vec{Q}^{n+2/3} - (1-\gamma)\Delta t \left(f_b \vec{Q}\right)^n - g \Delta t\ (h \nabla \zeta)^n$
(b) Substitution into the continuity equation (or its remainder)	
$\delta z - \frac{g\,(\gamma \Delta t)^2 h^n}{1 + \gamma \Delta t \,\mathrm{div} \vec{U}^n} \mathrm{div} \left(\frac{\nabla \delta z}{1 + \gamma f_b^n \Delta t} \right) = \mathcal{B}$	$\delta z - g\,(\gamma \Delta t)^2\ \mathrm{div} \left(\frac{h^n \nabla \delta z}{1 + \gamma f_b^n \Delta t} \right) = \mathcal{B}$
$\left(1 + \gamma \Delta t \,\mathrm{div} \vec{U}^n\right) \mathcal{B} = h^{n+1/3} - h^n$ $-\Delta t\, h^n \,\mathrm{div} \left[\vec{B} / (1 + \gamma f_b^n \Delta t) \right]$	$\mathcal{B} = -\Delta t \,\mathrm{div} \left[\vec{B} / (1 + \gamma f_b^n \Delta t) \right]$
$\vec{B} = (1-\gamma)\vec{U}^n + \gamma \vec{U}^{n+2/3} - g\,(\gamma \Delta t)\, \nabla \zeta^n$	$\vec{B} = (1-\gamma)\vec{Q}^n + \gamma \vec{Q}^{n+2/3} - g\,(\gamma \Delta t)\,(h \nabla \zeta)^n$
(c) Factorization of the above equation into the product of one-dimensional operators Solution of the resulting tridiagonal systems by double-sweep method	

8.5 Possible modifications of the propagation step solution

As mentioned above, inaccuracies in the solution of the propagation step may have different causes :

1. **The equations solved at this stage are inappropriate** because the applied linearization is not acceptable.

2. **The applied factorization is inadequate.**

3. **The one-dimensional systems resulting from factorization are not correctly solved.**

Let us deal first with the last point.
The systems considered here are tridiagonal. The related matrices are perturbations (induced by grid non-uniformity, depth and friction factor variability) of the following :

$$M = \begin{bmatrix} . & . & 0 & 0 & . & . & . & . \\ -\lambda & 1+2\lambda & -\lambda & 0 & . & . & . & . \\ 0 & -\lambda & 1+2\lambda & -\lambda & 0 & . & . & . \\ 0 & 0 & -\lambda & 1+2\lambda & -\lambda & 0 & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & 0 & -\lambda & 1+2\lambda & -\lambda & 0 \\ . & . & . & . & 0 & 0 & . & . \end{bmatrix}$$

where $\lambda = g\,(\gamma\Delta t)^2\,h\,/\,(1+\gamma\,f_b\,\Delta t)\,/\,(\Delta l)^2$ (Δl denoting the mesh size, either in the $x-$ or $y-$ direction). They are ill-conditioned for $\lambda \gg 1$ which occurs when propagation Courant numbers are large. In such cases, application of a standard double-sweep method will lead to very large amplification of machine round-off errors. The alternative is either to reduce the time step in order to reduce λ, or to turn to more clever inversion methods : several of them are introduced for instance in (Press *et al.*, 1989). Consequently, this kind of difficulty can be overcome with little pain ... and we shall no longer dwell on it.

Both the adequacy of the applied linearization and the magnitude of the factorization error can also be controlled by decreasing the time step. Yet, we followed another track : it consists of developing an alternative treatment of the propagation step, in which the linearization of the governing equations is corrected within a *limited* cycle of sub-iterations and where the factorization error can eventually be reduced.

8.5.1 Iterative Version - Formulation $\left(\zeta, \vec{U} \right)$ (depth-velocity)

Let $\vec{U}^{[q]}$ and $\delta z^{[q]}$ be the estimates of $\vec{U}^{n+1}$ and $\delta z^{n+1} = \zeta^{n+1} - \zeta^n$ after iteration number q of the propagation-friction step (nb : $\vec{U}^{[0]} = \vec{U}^n$ and $\delta z^0 = 0$). Superscript $[q]$ refers to all quantities calculated with the help of these estimates. What equations do the next estimates $\vec{U}^{[q+1]}$ and $\delta z^{[q+1]}$ satisfy ?

The development of equation 8.31 relied on the linearization of two products, namely $\left(h \, \mathrm{div}\vec{U}\right)^{n+1}$ (in the remainder of the continuity equation) and $\left(f_b \vec{U}\right)^{n+1}$ (in the remainder of the momentum equations).

The friction term was linearized as $f_b^n \, \vec{U}^{n+1}$. At iteration $q+1$ a straightforward improvement of this development consists of writing :

$$f_b^{n+1} \, \vec{U}^{n+1} \simeq f_b^{[q]} \, \vec{U}^{[q+1]} \tag{8.53}$$

Now, let us consider the other term. Till now, we had been writing :

$$\begin{aligned} h^{n+1} \, \mathrm{div}\vec{U}^{n+1} = (h^n + \delta z) \, \mathrm{div}\vec{U}^{n+1} &= h^n \, \mathrm{div}\vec{U}^{n+1} + \delta z \, \mathrm{div}\vec{U}^{n+1} \\ &\simeq h^n \, \mathrm{div}\vec{U}^{n+1} + \delta z \, \mathrm{div}\vec{U}^n \end{aligned}$$

A straightforward modification is to approximate (at iteration $q+1$) as follows :

$$h^{n+1} \, \mathrm{div}\vec{U}^{n+1} \simeq h^n \, \mathrm{div}\vec{U}^{[q+1]} + \delta z^{[q+1]} \, \mathrm{div}\vec{U}^{[q]} \tag{8.54}$$

A simpler method would be to take

$$h^{n+1} \, \mathrm{div}\vec{U}^{n+1} \simeq h^{[q]} \, \mathrm{div}\vec{U}^{[q+1]} \tag{8.55}$$

In that case, at the first iteration, the equation governing the free-surface increment is not equivalent to 8.31.
In the following pages, we summarize the governing equations obtained for sub-iteration $q+1$ of the propagation step, according to the choice of approximation 8.54 or 8.55.

The sub-iteration cycle stops whenever $\left|\delta z^{[q+1]} - \delta z^{[q]}\right|$, $\left|u^{[q+1]} - u^{[q]}\right|$, $\left|v^{[q+1]} - v^{[q]}\right|$ are all under some threshold value ϵ or for $q = q_{\max}$.

- We have chosen $\epsilon = 0.0005$, as used by (Wilders *et al.*, 1988) : this means that we allow the solution to be computed with an uncertainty of about 1 mm as regards surface elevation and water depth and 1 mm/s as regards flow velocity. This appears to be quite reasonable in the frame of most practical applications, where the measurements eventually available to check the model results would barely achieve this level of precision.

- $q_{\max}$ has been set to 5 : in the test cases discussed hereafter, it was observed that convergence is generally achieved after 2 to 3 iterations at the utmost.

Iterative "depth/velocity" formulation - Version 1

Initialisation : $\delta z^{[0]} = 0,\ h^{[0]} = h^n,\ \vec{U}^{[0]} = \vec{U}^n,\ f_b^{[0]} = f_b^n$

1. **Equation satisfied by the free surface increment :**

$$\delta z^{[q+1]} - \frac{g\,(\gamma\,\Delta t)^2\,h^n}{1+\gamma\,\Delta t\,\mathrm{div}\vec{U}^{[q]}}\,\mathrm{div}\left(\frac{1}{1+\gamma f_b^{[q]}\Delta t}\,\nabla\,\delta z^{[q+1]}\right) = \frac{\mathcal{B}^{[q+1]}}{1+\gamma\,\Delta t\,\mathrm{div}\vec{U}^{[q]}} + \mathcal{C}^{[q+1]} \tag{8.56}$$

with

$$\mathcal{B}^{[q+1]} = \mathcal{H} - \gamma\,\Delta t\,h^n\,\mathrm{div}\left[\frac{\vec{W}}{1+\gamma f_b^{[q]}\Delta t}\right] \tag{8.57}$$

where quantities $\mathcal{H}$ and $\vec{W}$ are independent of the sub-iteration and equate respectively :

$$\mathcal{H} = h^{n+1/3} - h^n - (1-\gamma)\,\Delta t\,h^n\,\mathrm{div}\vec{U}^n \tag{8.58}$$

$$\vec{W} = \vec{U}^{n+2/3} - (1-\gamma)\,\Delta t\,f_b^n\,\vec{U}^n - g\,\Delta t\,\nabla\zeta^n \tag{8.59}$$

and $\mathcal{C}^{[q+1]}$ is the term accounting for the correction of the factorization error.

2. **Correction of the factorization error**

The complete expression of $\mathcal{C}^{[q+1]}$ is quite complicated. It may be simplified in several ways, according to the respective magnitude and variability of friction factors and velocity divergences. For instance, in a frictionless situation it boils down to :

$$\mathcal{C}^{[q+1]} = \frac{g^2\,(\gamma\,\Delta t)^4\,h^n}{1+\gamma\,\Delta t\,\mathrm{div}\vec{U}^{[q]}}\,\frac{\partial^2}{\partial x^2}\left(\frac{h^n}{1+\gamma\,\Delta t\,\mathrm{div}\vec{U}^{[q]}}\,\frac{\partial^2\delta z^{[q]}}{\partial y^2}\right) \tag{8.60}$$

This correction may eventually be neglected ($\mathcal{C}^{[q+1]} = 0$).

3. **Updating of the velocity**

$$\left(1+\gamma f_b^{[q]}\Delta t\right)\vec{U}^{[q+1]} = \vec{W} - g\gamma\,\Delta t\,\nabla\delta z^{[q+1]} \tag{8.61}$$

Iterative "depth/velocity" formulation - Version 2

Initialisation : $\delta z^{[0]} = 0$, $h^{[0]} = h^n$, $\vec{U}^{[0]} = \vec{U}^n$, $f_b^{[0]} = f_b^n$

1. **Equation satisfied by the free surface increment :**

$$\delta z^{[q+1]} - g\,(\gamma\,\Delta t)^2\,h^{[q]}\,\mathrm{div}\left(\frac{1}{1+\gamma f_b^{[q]}\Delta t}\,\nabla\,\delta z^{[q+1]}\right) = \mathcal{E}^{[q+1]} + \mathcal{F}^{[q+1]} \tag{8.62}$$

with

$$\mathcal{E}^{[q+1]} = \mathcal{H} - \gamma\,\Delta t\,h^{[q]}\,\mathrm{div}\left[\frac{\vec{W}}{1+\gamma f_b^{[q]}\Delta t}\right] \tag{8.63}$$

where quantities $\mathcal{H}$ and $\vec{W}$ are defined as in relations 8.58 and 8.59 respectively, and $\mathcal{F}^{[q+1]}$ is the term accounting for the correction of the factorization error.

2. **Correction of the factorization error**

The complete expression of $\mathcal{F}^{[q+1]}$ is complex. In a frictionless case it simplifies into :

$$\mathcal{F}^{[q+1]} = g^2\,(\gamma\,\Delta t)^4\,h^{[q]}\,\frac{\partial^2}{\partial x^2}\left(h^{[q]}\,\frac{\partial^2 \delta z^{[q]}}{\partial y^2}\right) \tag{8.64}$$

It may eventually be neglected.

3. **Updating of the velocity**

It proceeds as in Version 1 (equation 8.61), namely :

$$\left(1 + \gamma f_b^{[q]}\Delta t\right)\vec{U}^{[q+1]} = \vec{W} - g\,\gamma\,\Delta t\,\nabla\delta z^{[q+1]}$$

Iterative "depth/unit discharge" formulation

Initialisation : $\delta z^{[0]} = 0,\ h^{[0]} = h^n,\ \zeta^{[0]} = \zeta^n,\ \vec{Q}^{[0]} = \vec{Q}^n,\ f_b^{[0]} = f_b^n$

1. **Equation satisfied by the free surface increment :**

$$\delta z^{[q+1]} - g\,(\gamma\Delta t)^2\ \mathrm{div}\left(\frac{h^{[q]}}{1+\gamma f_b^{[q]}\Delta t}\nabla\,\delta z^{[q+1]}\right) = -\Delta t\ \mathrm{div}\,\frac{\vec{B}^{[q+1]}}{1+\gamma f_b^{[q]}\Delta t} + \mathcal{G}^{[q+1]} \tag{8.65}$$

with

$$\vec{B}^{[q+1]} = (1-\gamma)\,\vec{Q}^n + \gamma\,\vec{Q}^{n+2/3} + \gamma\,(1-\gamma)\,\Delta t\left(f_b^{[q]} - f_b^n\right)\vec{Q}^n - g\,(\gamma\Delta t)\left(h^n + \gamma\delta z^{[q]}\right)\nabla\zeta^n \tag{8.66}$$

2. **Correction of the factorization error**

When it is not neglected, the corrective term $\mathcal{G}^{[q+1]}$ is estimated by :

$$\mathcal{G}^{[q+1]} = g^2\,(\gamma\Delta t)^4\,\frac{\partial}{\partial x}\left[\frac{h^{[q]}}{1+\gamma f_b^{[q]}\Delta t}\,\frac{\partial^2}{\partial x\partial y}\left(\frac{h^{[q]}}{1+\gamma f_b^{[q]}\Delta t}\,\frac{\partial\delta z^{[q]}}{\partial y}\right)\right] \tag{8.67}$$

3. **Updating of the discharge**

$$\begin{aligned}\left(1+\gamma f_b^{[q]}\Delta t\right)\vec{Q}^{[q+1]} &= \vec{Q}^{n+2/3} - (1-\gamma)\,\Delta t\left(f_b\vec{Q}\right)^n \\ &- g\,(1-\gamma)\,\Delta t\,(h\nabla\zeta)^n - g\gamma\Delta t\,(h\nabla\zeta)^{[q+1]}\end{aligned} \tag{8.68}$$

8.5.2 Iterative Version - Formulation $\left(\zeta, \vec{Q} \right)$ (depth-discharge)

As previously, superscript $[q]$ denotes estimates issued from sub-iteration number q.

Equation 8.36 was obtained after linearizing products $(h\nabla\zeta)^{n+1}$ and $\left(f_b\vec{Q}\right)^{n+1}$ respectively as $h^n\nabla\zeta^{n+1}$ and $f_b^n\vec{Q}^{n+1}$. At sub-iteration $q+1$, these are replaced by $h^{[q]}\nabla\zeta^{[q+1]}$ and $f_b^{[q]}\vec{Q}^{[q+1]}$.

The features of the resulting iterative loop are summarized in the preceding page.

8.5.3 Control of the iterative loop

The need for solving iteratively the propagation step, and then the performance of the iterative loop, can be appraised by surveying the evolution of errors $\phi = \left| (\delta z)^* - \delta z^{[q]} \right|$, $(\delta z)^*$ being defined by 8.51 or 8.52 (section 8.4.3). In most applications, the number of grid nodes is large. It would be cumbersome to analyze the difference value at each node. Its magnitude can be evaluated with the help of synthetic indicators, for instance its upper ($\phi_{\max}$) and lower limits and its mean ($\overline{\phi}$). It may also prove useful to record where errors reach their peak. If maximum errors are systematically observed in the same area, there is probably something to do about it, for example modifying the discretization in this spot ... The iterative loop appears to be relevant when final $\overline{\phi}$ and $\phi_{\max}$ turn out to be significantly lower than after the first sub-iteration.

As we are using a very simple method to adapt the linearization of the flow equations, we may expect occasionally to get into trouble. Indeed, if the time step is too large, so that the first estimates of the dependent variables at time $(n+1)\,\Delta t$ (i.e the outcome of the first sub-iteration) is quite erroneous, the iterative loop may fail to converge.

Finally, in spite of the improvement it can bring, the sub-iteration may still prove insufficient with final $\phi_{\max}$ and $\overline{\phi}$ being important with respect to the typical depth scales. In such cases, a reduction of the time step becomes inavoidable.

8.6 Treatment of boundary conditions

Until now we have been dealing with interior nodes only. In order to achieve the description of these models, we need now to present the treatment of boundary nodes. We already mentioned (cf section 3.7) that it constitutes one of the tricky points to deal with and that this task can notably be complicated by the application of process and space splitting to the governing equations.

Considering the organization of our models and the hyperbolic or parabolic nature of the

different operators involved, it appears that in order to solve the different stages we require :

- for the **advection stage**
  - **Formulation** $\left(\zeta, \vec{U} \right)$ **(depth-velocity)**
    the knowledge of $\vec{U}^n$ and ζ^n on any inflow boundary at any computational time t^n;
  - **Formulation** $\left(\zeta, \vec{Q} \right)$ **(depth-discharge)**
    the knowledge of $\vec{Q}^n$ on any inflow boundary
- for the **diffusion stage**, some condition about the diffused velocities or discharges (resp. $\vec{U}^{n+2/3}$ and $\vec{Q}^{n+2/3}$) on every boundary, at any time step;
- for the **propagation stage**, some condition about ζ on every boundary, at any time step.

These are more conditions than would be required to ensure the well-posedness of the physical problem at hand (cf section 3.7.1), because of the introduction of intermediate variables.

We shall have to deal with two kinds of problems :

- First, when sufficient boundary conditions are not supplied by data, what approximate conditions can we devise which are consistent with the physical problem at hand ?
- Secondly, once an appropriate boundary condition has been selected, how can we implement it numerically, without jeopardizing the stability and accuracy of our models ?

8.6.1 Open boundaries

(Weiyan, 1992) (chapter 3) provides a useful summary of which types of conditions are commonly used in practical applications. It depends on which problems are tackled. For instance, radiative boundary conditions are often used in marine applications (Orlanski, 1976; Blumberg & Kantha, 1985; Arnold, 1987; Bills & Noye, 1987). In subcritical flows in an open channel, we find most frequently that unit-width discharges are specified at the inflow boundary and that the outflow boundary condition consists either of a water level hydrograph, or of some rating curve. (Weiyan, 1992) insists on the fact that, depending on which condition is chosen, the open boundary must be located carefully, preferably in areas where the dependent variables are not likely to display great space-variation. Otherwise, any disturbance brought by the approximate boundary conditions may expand quickly.

It is far beyond the scope of this paper to deliver a complete review of open boundary conditions. We shall only present, as an example, what we have been using in fluvial applications.

Inflow boundary condition on water level Let us consider an inflow boundary parallel to the $y-$ direction. The tangential flow is assumed to be null, the flow normal to the boundary (unit-discharge Q_x) is specified. As mentioned above, we need an additional condition on the water level. It can be devised by using either the continuity equation or the momentum equations :

1. The continuity equation reads $\frac{\partial \zeta}{\partial t} + \mathrm{div}\vec{Q} = 0$ whose explicit development allows us to write :

$$\zeta^{n+1} = \zeta^n - \Delta t \, \mathrm{div}\vec{Q}^n \tag{8.69}$$

   Due to the fact that this development is explicit and that consequently, $\vec{Q}^{n+1}$ does not play a part in the estimation of ζ^{n+1}, we may expect trouble when the flow is quickly varying.

2. In most fluvial applications, the diffusive terms are far from being dominant. When they are neglected, the conservation of momentum along the $x-$direction reads :

$$\frac{\partial Q_x}{\partial t} + \frac{\partial u Q_x}{\partial x} + \frac{\partial v Q_x}{\partial y} + g\,h\,\frac{\partial \zeta}{\partial x} + f_b\,Q_x = 0 \tag{8.70}$$

   At the inflow boundary, equation 8.70 is developed semi-implicitly as for the interior nodes :

$$\frac{Q_x^{n+1} - Q_x^n}{\Delta t} + \frac{\partial u^n Q_x^n}{\partial x} + g\,h^n \left(\gamma \frac{\partial \delta z}{\partial x} + \frac{\partial \zeta^n}{\partial x} \right) + f_b^n \left(\gamma Q_x^{n+1} + (1-\gamma) Q_x^n \right) = 0 \tag{8.71}$$

   Sometimes, the advective terms may also be neglected (cf sec. 8.2). Otherwise, the only available possibility is to approximate them with the help of one-sided differences. It appears that if the gradients of discharges and velocities in the vicinity of the boundary are important, and if the advective Courant number is large, such approximation is inadequate.

   Assuming that the advective terms are correctly estimated, since the temporal evolution of Q_x is known, relation 8.71 allows us to estimate the evolution of the free surface gradient $\partial(\delta z)/\partial x$ at the open boundary. Besides, the conservation of momentum along the $y-$direction implies that $\partial \zeta / \partial y$ is null along the inflow boundary.

   $\partial(\delta z)/\partial x$ can be discretized as $[\ (\delta z)_2 - (\delta z)_1\ ]/(x_2 - x_1)$ (with subscripts 1 and 2 referring respectively to the inflow boundary and to the next section downstream). This provides us with a boundary equation for δz, which is substituted to general equations 8.36, 8.37 or 8.38 of the propagation step, which apply to interior nodes of the computational domain. We proceed similarly for "depth/velocity" models.

   This treatment of inflow conditions on the water level appears to be more consistent with the algorithm used for interior nodes than the first treatment suggested.

Inflow boundary conditions for the advection step Let us consider first a "depth/unit discharge model". There, the advection step concerns only the unit-width discharge. It is customary to assume at inflow boundary nodes that $Q_x^{n+1/3} = Q_x^{n+1}$ and $Q_y^{n+1/3} = Q_y^{n+1}$. When the characteristic line associated with an interior node exits the computational domain, the value of the discharges at the intersection of the characteristic and the inflow boundary is estimated both by spatial and temporal interpolation (based on the prescribed values for Q_x and Q_y) and is then affected to the interior node.

In case we deal with a "depth/velocity" model, things are somewhat more complicated. Assuming we choose the first option above (relation 8.69) we have got an estimate of ζ^{n+1} and h^{n+1} at the boundary. Since Q_x^{n+1} is given, the inflow velocities at the next time step can be evaluated. At the inlet, we then set simply $f^{n+1/3} = f^{n+1}$ for $f = u, v, h$.
Now, if the second option (eq. 8.71) is used, h^{n+1} is unknown at this stage of the computation. If the dimensional analysis demonstrates that advective terms are negligible with respect to propagation and friction terms, we can assume that at the inlet $f^{n+1/3} = f^n$ for $f = u, v, h$. Another, "mixed", condition is to postulate instead that $h^{n+1/3} = h^n$, $u^{n+1/3} = Q_x^{n+1}/h^n$ and $v^{n+1/3} = Q_y^{n+1}/h^n$.

Open boundary conditions for the diffusion step This step concerns only the unit-width discharges or the velocity components. Diffusion of momentum is usually neglected at open boundaries, so that we have $f^{n+2/3} = f^{n+1/3}$ for $f = Q_x, Q_y$ or $f = u, v$.

Outflow boundary conditions for the propagation step Water level is specified at the outlet. Use of complementary relation 8.69 or 8.71 at the inlet and of land boundary conditions we shall discuss in next section provide the necessary and sufficient conditions to solve the propagation-step equation governing the evolution of the free surface elevation. It is then necessary to update the velocities or the unit discharges. At the outflow boundary, we use for that purpose the same relation as for interior nodes, except that all spatial derivatives intervening in that formula are necessarily one-sided.

8.6.2 Fixed land boundaries

Possible conditions Let us consider a land boundary as sketched in figure 8.1. There can be no tranfer of mass or energy through this impervious limit. Consequently, velocities normal to the boundary should be zero. Assuming that the diffusive terms are negligible, this also implies (because of momentum conservation) that the normal gradient of water level is null. Different conditions can be used as regards the tangential velocity :

1. It can be left free. This means that once water levels have been computed, notably by using the zero normal gradient condition in order to modify the propagation step governing equations at closed boundary nodes, tangential velocities are computed normally, with the formula applicable to interior nodes.

2. It can be constrained in several ways (it is worth noting that then the problem become over-determined).

   (a) On walls, as the water is a viscous fluid, its velocity is theoretically null. Tangential velocity too should be set to zero. This is termed a "non-slip" condition. However, in order to describe accurately the velocity profile along the wall, the choice of a non-slip boundary condition must be combined with a local refinement of the computational grid (Mary, 1982; Stelling & Wang, 1984; Weiyan, 1992). In many applications, the grid size is much too large and a "perfect-slip" boundary condition is used instead.

   (b) A "perfect-slip" condition consists of assuming that the normal gradient of the tangential velocity is zero.

   (c) Another, intermediate, option is to take into account the fact that, in turbulent flows, the velocity profile in the wall boundary layer follows a logarithmic profile.

   Finally, as mentioned in (Stelling & Wang, 1984), the possible closed boundaries conditions on the tangential velocity can be summarized by the following formula :

$$(1-\alpha)\,U_t + \alpha\,\delta l\,\frac{\partial U_t}{\partial n} = 0 \tag{8.72}$$

   where U_t denotes the tangential velocity, n the direction normal to the boundary, δl is a length related to the distance between the closed boundary and the nearest computational point and α denotes a slip parameter. Picking α between 0 and 1 allows to describe every condition intermediate between the non-slip ($\alpha = 0$) and perfect slip ($\alpha = 1$) situation : when α differs from 0 or 1, the velocity obeys a logarithmic law. If we apply a first order discretization to the normal derivative in equation 8.72, the interpretation of the α parameter is straightforward : the tangential velocity on a solid boundary node is simply α times the value of this component at the nearest interior node.

   The α parameter is a priori time and space dependent. α should especially depend on the wall roughness which influences the development of the boundary layer. (Mary, 1982) proposed a way to estimate the α parameter, based on the prior estimation of a typical roughness height. He used empirical formulations (Carlier, 1986) allowing to link this roughness height to the overall Strickler rugosity coefficient.

Lastly, land boundaries can have dead angles. At such points, velocity is usually set to zero. By the way, this implies that local gradients of the water level in both directions are zero.

Now that we have mentioned the possible solid-walls conditions, let us see how we can implement them numerically.

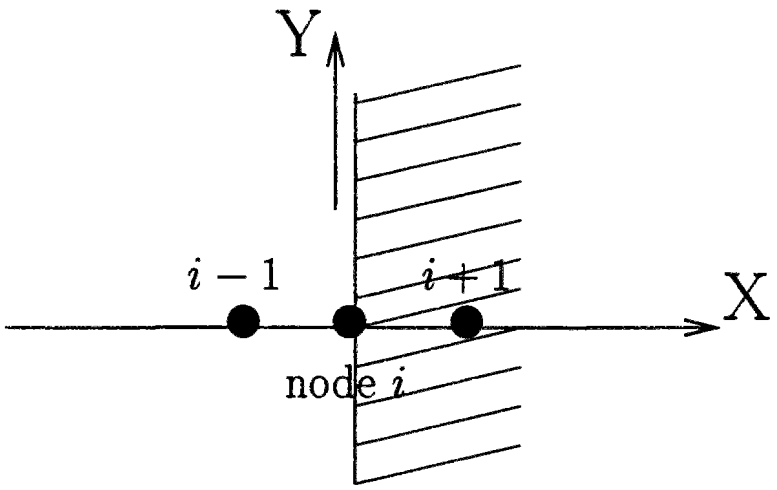


Figure 8.1: Land boundary

Diffusion step Let us consider the boundary in figure 8.1.

Considering the remains of the momentum equations (eq. 8.25 or 8.27) and their semi-implicit developments at the propagation step, it appears that, as $u = 0$ (respectively $Q_x = 0$) and $\frac{\partial \zeta}{\partial n} = 0$ at any time, we have also $u^{n+2/3} = 0$ (respectively $Q_x^{n+2/3} = 0$).

There should be no diffusion across a solid wall. Thus, we compute $v^{n+2/3}$ ($Q_y^{n+2/3}$) by applying only diffusion in the $y-$direction to the advected velocity (unit-discharge) $v^{n+1/3}$ ($Q_y^{n+1/3}$).

Propagation step The problem consists of introducing condition $\frac{\partial \zeta}{\partial n} = 0$ at closed boundaries. This implies that $\frac{\partial \delta z}{\partial n} = 0$.

Once eq. 8.31 or 8.36 have been factorized, the resulting equations along the $x-$direction read :

$$\left[I - \kappa \frac{\partial}{\partial x} \left(\mu \frac{\partial}{\partial x} \right) \right] \delta z = \mathcal{B} \tag{8.73}$$

where I denotes the identity operator, κ and μ are quantities depending on the time step, implicitation parameter, local depths and friction factors, ... and $\mathcal{B}$ depends on past, advected or diffused values of the dependent variables. Discretizing 8.73 at an interior node requires the use of its two closest neighbours along the $x-$direction.

A first option for a closed boundary node consists of substituting straightforwardly to 8.73

the zero gradient condition. If this gradient is approximated with a first order-error, we would simply write, in a situation similar to figure 8.1 :

$$\delta z_i - \delta z_{i-1} = 0 \tag{8.74}$$

If a higher-order approximation of the normal gradient is used, this will involve more computational nodes. Then, some manipulations of the resulting matrices are required in order to reduce them to a tridiagonal form to which the double-sweep method can apply.

We have also been implementing a second option, which is described below.

First, quantity $\frac{\partial}{\partial x}\left(\mu \frac{\partial \delta z}{\partial x}\right)$ develops as $\frac{\partial \mu}{\partial x}.\frac{\partial \delta z}{\partial x} + \mu.\frac{\partial^2 \delta z}{\partial x^2}$

At closed boundary node i, the first product above disappears. Let us consider fictitious node $i+1$ which is a mirror image, with respect to the wall, of node $i-1$, the interior node closest to i (cf fig. 8.1). Let Δx denotes the local grid size $x_i - x_{i-1}$. Second-order approximations of the first and second order derivatives of δz at node i are thus respectively :

$$\frac{\delta z_{i+1} - \delta z_{i-1}}{2\Delta x} \quad \text{and} \quad \frac{\delta z_{i+1} + 2\delta z_i - \delta z_{i-1}}{(\Delta x)^2}$$

Nullity of the normal gradient leads to $\delta z_{i+1} = \delta z_{i-1}$ so that

$$\frac{\partial^2 \delta z}{\partial x^2}_{|i} \simeq \frac{2\,(\,\delta z_{i-1} - \delta z_i)}{(\Delta x)^2}$$

This approximation can be used in order to discretize eq. 8.73 at the boundary node. The resulting system is still tridiagonal and can easily be inverted.

Open channel simulation - Applied boundary conditions

1. **Known boundary conditions** (inflow boundary assumed to be parallel to Oy) :
   - At the inflow boundary, normal discharge (Q_x) is prescribed, tangential discharge (Q_y) is assumed to be zero.
   - Water level (ζ or h) prescribed at the outflow boundary.
2. **Complementary condition on ζ at the inflow boundary** obtained
   - either by using the continuity equation (option 1) : $\dfrac{\partial \zeta}{\partial t} = -\Delta t \, \mathrm{div}\vec{Q}$
   - or by using the conservation of momentum relation (option 2) :

$$\frac{\partial \zeta}{\partial x} = -\frac{1}{g\,h}\left[\frac{\partial Q_x}{\partial t} + \frac{\partial u Q_x}{\partial x} + f_b Q_x\right]$$

3. **Land boundary conditions**
   - Boundary parallel to the $x-$ or the $y-$ direction
     (a) Normal velocity (U_n) set to zero
     (b) Normal gradient of water level set to zero : $\dfrac{\partial \zeta}{\partial n} = 0$
     (c) Tangential velocity U_t is either free, either constrained by condition

$$(1-\alpha)\,U_t + \alpha\,\delta l\,\frac{\partial U_t}{\partial n} = 0$$

     where α is a "slip" parameter, to be tuned.
   - Dead angle : $\vec{U} = 0$ and $\nabla \zeta = 0$

Open channel simulation - Numerical treatment of boundary conditions

1. **Advection step** Boundary conditions are necessary at the inflow boundary

   - Model based on a "depth/unit discharge" formulation.
     Set $Q_x^{n+1/3} = Q_x^{n+1}$ and $Q_y^{n+1/3} = Q_y^{n+1}$.
   - Model based on a "depth/velocity" formulation. Use

     (a) if ζ^{n+1} has been estimated through option 1 above,
     $h^{n+1/3} = h^{n+1}$, $u^{n+1/3} = Q_x^{n+1}/h^{n+1}$ and $v^{n+1/3} = Q_y^{n+1}/h^{n+1}$

     (b) if ζ^{n+1} is still unknown, and advective terms negligible,
     $h^{n+1/3} = h^n$, $u^{n+1/3} = u^n$ and $v^{n+1/3} = v^n$
     otherwise, $h^{n+1/3} = h^n$, $u^{n+1/3} = Q_x^{n+1}/h^n$ and $v^{n+1/3} = Q_y^{n+1}/h^n$

2. **Diffusion step**

   - Diffusion is neglected at open boundaries :
     $f^{n+2/3} = f^{n+1/3}$ for $f = Q_x$, Q_y or $f = u$, v.
   - Land boundary normal to Ox :
     $u^{n+2/3} = 0$ ($Q_x^{n+2/3} = 0$), diffusion of v (Q_y) in the x-direction neglected.
   - Land boundary normal to Oy :
     $v^{n+2/3} = 0$ ($Q_y^{n+2/3} = 0$), diffusion of u (Q_x) in the y-direction neglected.

3. **Propagation step**

   (a) Modification of the governing equation for δz at closed boundary nodes, either by replacing it directly by the zero normal gradient condition $\dfrac{\partial \delta z}{\partial n} = 0$, or by applying approximation such as
   (i boundary node, $i-1$ its closest interior neighbour) :

$$\frac{\partial}{\partial x}\left(\mu \frac{\partial \delta z}{\partial x}\right) \simeq \mu_i \frac{2\,(\,\delta z_{i-1} - \delta z_i)}{(\Delta x)^2}$$

   (b) Updating of the velocities (or discharges) :

   - At a closed boundary, the normal velocity is null, tangential velocity is updated either by the same formula as for interior nodes, or, if a slip condition is imposed, as a function of the tangential velocity at the closest interior node.
   - At an outflow boundary, the velocity is updated with usual formula (using one-sided finite differences). At the inflow boundary, velocity is calculated so that the prescribed condition on the discharge is satisfied.

8.7 Conclusion

After manipulating a lot of equations throughout this chapter, we :

- proposed two different models for the solution of the flow equations, each one using a different set of dependent variables, each one open to different simplifications (sections 8.3 and 8.4);
- tried to underline (notably in sections 8.4.2 and 8.4.3) what can cause these models to fail producing accurate results;
- suggested possible strategies to improve the models results or range of applicability (section 8.5).

No matter how carefully we might have proceeded in their derivation, it does not warrant our models will perform conveniently in any case. Many parameters control the flow patterns, e.g. the bathymetry and geometry of the studied domain, the rugosity, the extent of the diffusion phenomenon, the boundary conditions, ... Consequently, prior to applying a model to real field situations, it is useful to consider first benchmark tests where either a reference solution or measurements, or some qualitative information about the plausible flow behaviours are available. This allows us to investigate thoroughly the performance of the different parts of the model. A similar methodology was applied in order to evaluate the performance of different advection-diffusion algorithms (cf. part II).

Various tests have been introduced in the literature (Weare, 1979; Benque *et al.*, 1982; Wilders *et al.*, 1988; Galland & Hervouet, 1988; Dan N'Guyen, 1988; Dan N'Guyen, 1993). We shall deal first with two typical applications implying the computation of equilibrium steady-state flows (chapter 9) : this should allow us notably to evaluate which simplifications of the models equations or algorithms are acceptable. Secondly, we shall consider test cases relative to unsteady flows (chapter 10) : this will in particular provide an opportunity for assessing the interest of modifying the solution of the propagation step.

8.8 Résumé français : "Résolution des équations de St-Venant"

Les équations de St-Venant et leur signification physique Comme indiqué en conclusion du chapitre 2, nous avons décidé, compte tenu des applications en Seine envisagées et de leur contexte (*cf* section 2.4.4), de nous en tenir à une représentation bidimensionnelle des écoulements et de décrire ceux-ci grâce aux équations de St-Venant que nous rappelons ci-dessous (section 8.1).

La première inconnue est la cote de la surface libre ζ, ou la hauteur d'eau h (à partir du moment où l'on considère les fonds fixes - ce qui sera notre cas - le choix est indifférent). L'expression des équations diffère ensuite selon que l'on retient comme inconnues les composantes u et v de la vitesse $\vec{U}$ du fluide (moyennée sur la colonne d'eau) ou les composantes (Q_x, Q_y) du débit unitaire $\vec{Q}$ (produit de la vitesse par la hauteur d'eau, c.a.d. débit par unité de largeur).
Avec les variables $\left(\zeta, \vec{Q} \right)$, on aboutit à :

$$\begin{aligned}
\frac{\partial \zeta}{\partial t} + \mathrm{div}\vec{Q} &= 0 \\
\frac{\partial Q_x}{\partial t} + \frac{\partial u Q_x}{\partial x} + \frac{\partial v Q_x}{\partial y} + gh\frac{\partial \zeta}{\partial x} + f_b Q_x &= S'_x \\
\underbrace{\frac{\partial Q_y}{\partial t}}_{\mathrm{I}} + \underbrace{\frac{\partial u Q_y}{\partial x} + \frac{\partial v Q_y}{\partial y}}_{\mathrm{II}} + \underbrace{gh\frac{\partial \zeta}{\partial y}}_{\mathrm{III}} + \underbrace{f_b Q_y}_{\mathrm{IV}} &= \underbrace{S'_y}_{\mathrm{V}}
\end{aligned}$$

La première équation est l'équation de continuité, qui traduit la conservation de la masse. Les deux suivantes expriment la conservation de la quantité de mouvement. On y trouve cinq groupes de termes, dont la signification physique est la suivante :

I Accélération locale de l'écoulement (i.e. variation de la quantité de mouvement en fonction du temps)

II Transport de la quantité de mouvement par l'écoulement "moyen" (termes inertiels)

III Variations de quantité de mouvement dues à la propagation des ondes

IV Dissipation d'énergie par frottement

V Dispersion horizontale de la quantité de mouvement (engendrée à la fois par la turbulence et par la convection différentielle suivant une verticale)

Les variations de I sont dus à la combinaison des actions II à V.

Avec les variables $\left(\zeta, \vec{U} \right)$ on obtient une forme un peu différente des équations, la divergence du débit, notamment, étant développée de façon à accroître la similitude entre les trois équations.

$$\begin{array}{lllll} \frac{\partial \zeta}{\partial t} + \vec{U}.\nabla h + & h\,\mathrm{div}\vec{U} & & = & 0 \\ \frac{\partial u}{\partial t} + \vec{U}.\nabla u + & g\frac{\partial \zeta}{\partial x} & + f_b u & = & S_x \\ \underbrace{\frac{\partial v}{\partial t}}_{\mathrm{I}} + \underbrace{\vec{U}.\nabla v}_{\mathrm{II}} + & \underbrace{g\frac{\partial \zeta}{\partial y}}_{\mathrm{III}} & + \underbrace{f_b v}_{\mathrm{IV}} & = & \underbrace{S_y}_{\mathrm{V}} \end{array}$$

NB : Dans les équations ci-dessus,

- ∇ et div désignent respectivement les opérateurs gradient et divergence; (nb : Δ désignera le Laplacien, c.a.d. la divergence du gradient)
- g est l'accélération de la gravité (soit 9.81 m.s^{-2});
- S_x, S_y, S'_x et S'_y sont des opérateurs de diffusion (développement complet : cf equation 8.8);
- le paramètre f_b (s^{-1}) est le "facteur de frottement" tel que la contrainte de frottement au fond $\vec{\tau}_b$ soit égale à $\rho\, f_b \vec{Q}$ (ρ masse volumique du fluide); f_b est généralement exprimé à l'aide des formules de Chezy ou Strickler, soit respectivement

$$f_b = g\frac{\sqrt{u^2+v^2}}{C_h^2 h} = g\frac{\sqrt{Q_x^2+Q_y^2}}{C_h^2 h^2} \quad \text{et} \quad f_b = g\frac{\sqrt{u^2+v^2}}{K_s^2 h^{4/3}} = g\frac{\sqrt{Q_x^2+Q_y^2}}{K_s^2 h^{7/3}}$$

où les coefficients empiriques de Chezy (C_h) ou Strickler (K_s) traduisent la rugosité du lit.

Avant d'aborder la résolution proprement dite, il est intéressant d'estimer l'importance relative des termes II à V, les termes dominants requérant bien sûr un surcroît d'attention. Pour ce faire, nous utilisons l'analyse adimensionnelle (cf section 8.2), introduisant des échelles caractéristiques respectivement des vitesses, de la dimension (en plan) du domaine de calcul ou des longueurs d'onde de l'écoulement, ainsi que de la profondeur : V, L et H. Le tilde dénote les nouvelles coordonnées temporelles et spatiales et les opérateurs différentiels qui leur sont liés. En supposant que la dispersion est isotrope et uniforme, caractérisée par la diffusivité ν, la conservation de la quantité de mouvement (en partant de sa formulation en variables ζ et $\vec{U}$) s'exprime :

$$\frac{\partial \tilde{U}}{\partial \tilde{t}} + \left(\tilde{U}.\tilde{\nabla}\right)\tilde{U} + \frac{gH}{V^2}\tilde{\nabla}\tilde{Z} - \frac{\nu}{VL}\tilde{\Delta}\tilde{U} = -\frac{L\vec{F}_f}{V^2}$$

($\vec{F}_f$ représente la force de frottement, $\vec{F}_f = f_b\vec{U}$)

En paramétrant le frottement par la formule de Strickler on fait finalement apparaître trois produits adimensionnels qui représentent le poids respectif des termes de propagation, dispersion et frottement par rapport aux termes inertiels (i.e. advectifs), à savoir :

- $p_1 = \frac{gH}{V^2}$ qui peut être exprimé en fonction du nombre de Froude $F_r = \frac{V}{\sqrt{gH}}$, quotient de la vitesse de l'écoulement par la célérité des ondes. On a en effet $p_1 = 1/F_r^2$.
- $p_2 = \frac{\nu}{VL}$ qui se présente comme l'inverse d'un nombre de Reynolds

- $p_3 = \dfrac{g\,L}{K_s^2 H^{4/3}}$ tel que $-L/V^2\,\vec{F}_f = p_3\,\|\tilde{U}\|\,\tilde{U}\,\tilde{h}^{-4/3}$

Considérant des valeurs de H, L, V, ν, K_s typiques des systèmes océaniques, côtiers (Goutal, 1987) et fluviaux, on en déduit que : les termes de propagation sont généralement trés dominants dans le premier cas; dans le second cas, on a une juxtaposition de situations diverses, avec des zones profondes où la propagation domine et les franges littorales, zones couvrantes-découvrantes, où l'équilibre advection/frottement régit tout; dans le troisième cas, l'écoulement est fréquemment contrôlé par propagation et frottement.

Principes de résolution (section 8.3)

Nous avons développé deux types de modèles, qui correspondent à des choix de variables différents : cote de la surface libre et vitesse dans le premier cas; cote et débit unitaire dans le second. Chaque modèle repose néammoins sur la même approche à pas fractionnaires. Les équations sont d'abord éclatées selon les processus physiques en cause : ainsi on prendra en compte successivement les opérateurs différentiels correspondant aux termes advectifs (groupe II), puis à la dispersion de quantité de mouvement(groupe V), enfin ...à tout ce qui reste (propagation, frottement, conservation de la masse ...). On couple à cela (étapes 2 et 3 du calcul) l'éclatement suivant les directions de l'espace. Comme les équations de base sont légèrement différentes, le découpage entre les étapes de calcul est également un peu distinct.
(Par la suite, l'indice n se réferera à l'état des variables au temps t^n, l'indice $n+1$ à l'état au pas de temps suivant $t^{n+1} = t^n + \Delta t$, les indices $n+1/3$ et $n+2/3$ aux états intermédiaires respectivement après prise en compte de l'advection, puis de la dispersion)

Les équations de travail de **l'étape d'advection (section 8.3.1)** sont respectivement, en formulation hauteur-vitesse (sec. 8.3.1.1),

$$\frac{\partial f}{\partial t} + \vec{U}.\nabla f = 0 \quad \text{avec } f = h,\, u,\, v$$

et, en formulation hauteur-débit (sec. 8.3.1.2),

$$\frac{\partial f}{\partial t} + \frac{\partial uf}{\partial x} + \frac{\partial vf}{\partial y} = 0 \quad \text{avec } f = Q_x,\, Q_y$$

- Dans le premier cas, on se retrouve avec la forme classique, non conservative, de l'équation de transport pur d'un soluté, dont nous avons étudié les méthodes de résolution tout au long de la partie II de ce rapport. Nous implanterons donc dans notre modèle la méthode jugée la meilleure dans les cas délicats d'advection dominante et de gradients forts du scalaire transporté, à savoir la méthode aux caractéristiques de Rasch et Williamson (cf 7.5). Cependant, afin de traiter de façon "économique" les cas où le terme advectif n'a

qu'une faible influence, nous implantons de plus la méthode UPWIND (cf chap. 4, section 4.2.2).

- Dans le second cas, deux options ont été essayées.
  - Si on développe l'équation des composantes du débit (Benque *et al.*, 1982), par exemple pour Q_x, on obtient :
$$h\left(\frac{\partial u}{\partial t} + \vec{U}.\nabla u\right) + u\left(\frac{\partial h}{\partial t} + \mathrm{div}\vec{Q}\right) = 0$$
ce qui, compte tenu de l'équation de continuité, se réduit à $\frac{\partial u}{\partial t} + \vec{U}.\nabla u = 0$. Cette équation sur u est identique à ce qu'on obtient directement avec les variables hauteur et vitesse. On la résoud pareillement puis on obtient le débit après prise en compte de l'advection par application de la formule $\vec{Q}^{n+1/3} = h^n \vec{U}^{n+1/3}$.
  - On simplifie l'équation de travail. Ainsi, (Dan N'Guyen, 1988) néglige le terme $\left(\mathrm{div}\vec{U}.\vec{Q}\right)$ ce qui fait que les composantes du débit après advection se retrouvent régies par :
$$\frac{\partial f}{\partial t} + \vec{U}.\nabla f = 0 \quad \text{avec } f = Q_x, Q_y$$

  On peut évaluer théoriquement les erreurs entraînées par l'une et l'autre option. Soit $\vec{Q}^*$ la solution idéale. On a respectivement
$$\vec{Q}^* - \vec{Q}^{n+1/3} \simeq \left(h^{n+1} - h^n\right).\vec{U}^n$$
$$\text{et} \quad \vec{Q}^* - \vec{Q}^{n+1/3} \simeq \left(h^{n+1} - h^{n+1/3}\right).\vec{U}^n$$
  Dans le premier cas, contrairement au second, l'erreur disparaît quand l'écoulement est stationnaire. Dans le second cas, l'erreur n'apparaît a priori négligeable que si la divergence du champ de vitesses est faible, donc l'écoulement proche de l'uniformité.

Une des différences importantes entre les deux types de formulation réside dans le fait que, dans la formulation hauteur-vitesse, la hauteur est également concernée par l'étape d'advection.

Pour l'**étape de dispersion (section 8.3.2)** on a à résoudre un opérateur de diffusion on ne peut plus classique. Pour celà, on a recours à une méthode aux différences finies, semi-implicite, aux directions alternées, qui, compte tenu de la discrétisation adoptée (méthode de Crank-Nicholson) ne met en jeu que des systèmes linéaires de type tridiagonal. Cette étape ne concerne que les valeurs des composantes du débit ou de la vitesse.

L'**étape de propagation-frottement** est celle qui est traitée de la façon la plus originale. Ses équations de travail sont respectivement, en formulation hauteur-vitesse
$$\frac{\partial h}{\partial t} + h\,\mathrm{div}\vec{U} = 0$$
$$\frac{\partial \vec{U}}{\partial t} + g\nabla\zeta + f_b\vec{U} = 0$$

et en formulation hauteur-débit

$$\frac{\partial \zeta}{\partial t} + \mathrm{div}\vec{Q} = 0$$
$$\frac{\partial \vec{Q}}{\partial t} + gh\nabla\zeta + f_b\vec{Q} = 0$$

Dans le premier cas, contrairement au second, l'équation de quantité de mouvement n'est pas la seule à avoir été décomposée process par process : l'équation de continuité a été partiellement traitée lors de l'étape d'advection.

Ces équations doivent être résolues entre t^n et t^{n+1}, les conditions initiales correspondant à l'état des variables après prise en compte des termes advectifs et dispersifs. Ceci implique que les dérivées temporelles sont développées ainsi :

$$\frac{\partial h}{\partial t} = \frac{h^{n+1} - h^{n+1/3}}{\Delta t} \quad \frac{\partial \vec{F}}{\partial t} = \frac{\vec{F}^{n+1} - \vec{F}^{n+2/3}}{\Delta t} \text{ pour } \vec{F} = \vec{U}, \vec{Q} \quad \text{et enfin } \frac{\partial \zeta}{\partial t} = \frac{\zeta^{n+1} - \zeta^n}{\Delta t}$$

Le principe de résolution est le suivant :

1. **On développe tout d'abord les équations de façon semi-implicite**, en introduisant le paramètre d'implicitation γ ($0.5 \leq \gamma \leq 1$), et en faisant apparaître la variable δz, qui est l'incrément de la surface libre entre deux pas de temps. Puisque les fonds sont fixes, $\delta z = \zeta^{n+1} - \zeta^n = h^{n+1} - h^n$.
2. Ensuite, **on découple équation (ou résidu d'équation) de continuité et résidu de l'équation de conservation de la quantité de mouvement.**
   (a) Pour celà, **on linéarise l'équation de quantité de mouvement** ce qui permet d'exprimer l'inconnue $\vec{U}^{n+1}$ (resp. $\vec{Q}^{n+1}$) comme une fonction des valeurs passées ζ^n, $\vec{U}^n$ (resp. $\vec{Q}^n$), intermédiaires $\vec{U}^{n+2/3}$ (resp. $\vec{Q}^{n+2/3}$) et de l'incrément δz.
   (b) Ceci permet d'**éliminer par substitution $\vec{U}^{n+1}$ de l'équation de continuité** (ou de son résidu). **On obtient ainsi une équation dont la seule inconnue est δz.**
3. L'équation régissant δz est **proche d'une équation de Poisson. Elle est résolue après factorisation selon les deux directions de l'espace.** On se ramène ainsi à la manipulation de systèmes linéaires tridiagonaux.
4. Une fois δz estimé, et ζ^{n+1} ainsi déterminé, sa valeur est réinjectée dans l'équation (cf point 2.a) gouvernant $\vec{U}^{n+1}$ (resp. $\vec{Q}^{n+1}$) afin de calculer celle-ci.

L'équation satisfaite par δz est différente dans les deux formulations puisqu'on part de systèmes d'équations différents selon que l'on travaille en vitesse ou en débit unitaire. Sa forme générale est :

$$\delta z - \alpha \mathrm{div}(\beta \nabla \delta z) = \mathcal{M}$$

α et β revêtent des expressions différentes selon les simplifications adoptées.

1. Le premier niveau de simplification concerne aussi bien la formulation hauteur-vitesse que la formulation hauteur-débit. Il consiste à négliger les variations spatiales du terme $1 + \gamma f_b \Delta t$. On peut alors extraire le quotient $1/(1 + \gamma f_b \Delta t)$ de dessous l'opérateur divergence. On conçoit intuitivement que ceci ne sera admissible que lorsque le frottement est faible (f_b petit) et le pas de temps également limité.

   Quand cette simplification est admissible, l'équation à résoudre pour δz en formulation $\left(\zeta, \vec{U} \right)$ devient une équation de Poisson, i.e. $\delta z - \alpha' \Delta \delta z = \mathcal{M}$
   Pour la formulation $\left(\zeta, \vec{Q} \right)$, on obtient une expression un peu plus compliquée, à savoir $\delta z - \alpha' \operatorname{div}(h^n \nabla \delta z) = \mathcal{M}$

2. Le deuxième niveau de simplification ne concerne que la formulation $\left(\zeta, \vec{Q} \right)$. Si l'on néglige les variations spatiales de h^n, comme proposé par (Dan N'Guyen, 1988), on se ramène en effet également à une équation de Poisson.

Finalement, modèles et techniques de résolution suggérés, en formulation $\left(\zeta, \vec{U} \right)$ ainsi que $\left(\zeta, \vec{Q} \right)$, sont résumés dans la table 8.1.

Validité de la méthode proposée La validité de notre approche dépend en premier lieu de **la validité de l'équation obtenue pour l'incrément** δz. En bref, on peut dire que celle-ci **dépendra de la valeur du pas de temps** Δt :

- La précision du développement semi-implicite des équations de continuité et quantité de mouvement dépend en effet de l'**erreur de troncature proportionnelle à** Δt. Si l'on choisit celui-ci trop grand, le modèle peut diverger, ou converger vers une solution plausible ... mais qui ne présente guère de ressemblance avec l'écoulement réel.

- Par ailleurs, **la valeur de** Δt **conditionne la qualité de la linéarisation**, du terme de frottement notamment. La solution la plus simple est de prendre pour le coefficient f_b sa valeur au temps t^n. Dans le cas d'un écoulement transitoire, ceci sera d'autant plus erroné que Δt est grand.

Une autre source d'erreur est liée à la méthode de résolution adoptée pour l'équation contrôlant δz, à savoir la factorisation. La forme factorisée de l'équation diffère de l'originale par un terme de troncature qui est proportionnel à la puissance quatrième du nombre de Courant de propagation, $C_p = \sqrt{gh}\,\Delta t \sqrt{\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2}}$, ainsi qu'à la courbure de la surface libre suivant les deux directions de l'espace. Ce terme de troncature sera donc d'autant plus important que : (i) la célérité des ondes ($\sqrt{gh}$) est forte (ii) le pas de temps est grand (iii) la surface libre est "chahutée".

Erreurs de factorisation et de linéarisation peuvent être réduites moyennant un processus itératif, dont le principe est exposé en fin de section 8.4.3 et la mise en oeuvre

détaillée en sections 8.5.1 et 8.5.2 pour les formulations hauteur-vitesse et hauteur-débit respectivement. On notera qu'il existe deux façons différentes d'itérer l'étape de propagation en formulation $\left(\zeta, \vec{U} \right)$, une seule pour la version $\left(\zeta, \vec{Q} \right)$.
Le contrôle de l'itération (fin de la boucle) dépend de l'écart entre les solutions itérées successives (cf sec 8.5.1). Dans les divers tests pratiqués (cf chapitres suivants), la convergence a été atteinte au pire à l'issue de 3 itérations.
La nécessité de procéder à un cycle de résolution itératif et l'efficacité de ce cycle sont appréciés en fonction de l'erreur commise sur la résolution de l'équation de continuité (sec 8.5.3), plus exactement de la moyenne et du maximum de cette erreur (dont le mode d'évaluation est indiqué en 8.4.3).

Conditions limite On aborde en section 8.6 le problème des conditions limite. Tout d'abord, il convient de choisir celles-ci de façon cohérente avec la physique du phénomène hydraulique. Ensuite, il faut les implanter numériquement, les principaux écueils concernant la définition de conditions limite ad hoc pour les étapes de calcul intermédiaire et leur "éclatement" (i.e. découplage) suivant les directions de l'espace quand l'opérateur différentiel correspondant est éclaté.

Après une revue "générale" des conditions limites plausibles suivant les types d'écoulement (cf sections 8.6.1 et 8.6.2 respectivement pour les frontières ouvertes et fermées), on détaille les conditions applicables aux écoulements fluviaux et leur traitement numérique.

De la nécessité des tests Au cours de ce chapitre nous avons proposé diverses approches relatives à la résolution des équations bidimensionnelles de St-Venant, qui se distinguent les unes des autres par le choix des variables de travail, le degré de simplification admis, le recours éventuel à un cycle de résolution itératif. Un point commun à ces modèles, ou différentes versions de modèles, est néammoins l'adoption d'une technique à pas fractionnaires et le découplage systématique (par directions alternées ou factorisation) suivant les directions de l'espace.

Nous avons attiré au passage l'attention sur les hypothèses sous-jacentes aux différentes simplifications et leur inéluctable limitation. Cependant, nous n'avons guère moyen de juger du bien fondé de telle ou telle version de modèle sans passer à une phase de tests, comme nous l'avons fait dans notre recherche d'algorithmes satisfaisants pour l'advection-diffusion. Ces tests seront détaillés en chapitres 9 et 10 pour les écoulements permanents et instationnaires respectivement.

Chapter 9

Validation on benchmark tests : steady-state flows

9.1 Choice of test cases

As in the case of advection-diffusion solution, we are faced to the difficulty of finding test problems which allow an objective comparison of algorithms and, at the same time, are relevant with respect to the forthcoming applications. Unfortunately, there are few situations where analytical solutions are available or where the plausible flow pattern can be easily guessed. The test cases studied in this chapter and the following one are, apart from the expanding flume test (section 10.3), extracted from a set of benchmark problems proposed by hydraulicians at Électricité de France (Galland & Hervouet, 1988) (a reference in France as regards computational fluid mechanics applied to environment). A few other tests have been proposed in the literature : the case of a straight channel (with a 45 degrees angle with respect to the coordinates axis) (Weare, 1979), the case of a rectangular tidal basin with a deeper S-shaped channel inside (Benque *et al.*, 1982), the case of wave propagation (without friction neither advection) inside another rectangular tidal basin (Dan N'Guyen, 1988). It is also worth noting that various experimental studies have been published about velocity distribution in laboratory flumes (in bends notably). It would surely be interesting to build some test based on such experiences. However, at the time being, we are not aware of an effort similar to the task undertaken by the organizers of the Convection-Diffusion Forum, consisting of a systematic compilation of reference problems and their solution. We thus decided to work with what was available : the most complete inventory of tests was indeed to be found in (Galland & Hervouet, 1988).

The two first tests presented below concern the computation of steady-state equilibrium flows. We selected them because they are the closest to fluvial applications we could find. As could

be expected, they do not allow to investigate every possible trouble. First, they deal with flows for which the diffusion of momentum is neglected : fortunately, this neglect is commonplace in fluvial situations. Secondly, and this is more annoying, they concern unidirectional flows. However, they do present some advantages :

- Both possess an analytical reference solution. This allows us to check at the same time the accuracy and convergence rate of the applied models.

- The fractionary step approach and the propagation step treatment we apply have been first suggested in (Dan N'Guyen, 1988). However, the subsequent tests and full scale applications (Dan N'Guyen, 1988; Dan N'Guyen, 1993) deal mainly with coastal and marine situations where the propagation term is dominant. In the following tests, the situation is more balanced : advection, friction, propagation have roughly the same order of magnitude. Consequently, we shall observ numerical phenomena unnoticed till now.

As regards steady-state flows, it could have been fruitful to work on the 45 degrees channel (there, boundary conditions are tricky to apply in the frame of directional splitting). Yet, we have a similar, but real, problem at hand with the meandering Seine (cf chapter 11), on which we decided to work directly.

We have been proposing in the previous chapter different possibilities for dealing with the advective terms (section 8.3.1), and different versions of the governing equations of the propagation step, corresponding to different degrees of simplification (sec. 8.4.1.1 and 8.4.1.2). For the sake of brevity, we shall hereafter refer to these different options as follows :

- **Formulation "depth/velocity"** : UH2 refers to the model based on complete equation 8.31, UH1 to the model based on equation 8.32.

- **Formulation "depth/discharge"** :

  - The free surface evolution is governed by the equation suggested in (Dan N'Guyen, 1988), i.e. equation 8.38. The corresponding models are respectively denoted DAN0 and QH0, according to the method applied to deal with the advective terms (respectively suggested in (Dan N'Guyen, 1988) or (Benque *et al.*, 1982)).

  - No simplification is applied to the free surface increment equation 8.36 : corresponding models are denoted DAN2 and QH2.

  - Intermediate simplification is applied, as presented in equation 8.37 : corresponding models are DAN1 and QH1.

9.2 Backwater curve calculation

9.2.1 Presentation

One of the most frequent tasks in fluvial hydraulics is to compute surface profiles. The present test is typical of such problems. It consists of computing the features of the equilibrium flow which develops in a sloping channel controlled by an upstream steady-state discharge condition and a downstream condition concerning the free surface elevation (there, it forces the flow depth to be superior to its normal value).

In this specific case (cf table 9.1), the Froude number F_r ranges between 0.13 and 0.4 approximately : the dimensional product p_1 which reflects the relative weight of propagation terms with respect to inertial ones (cf section 8.2) varies between 6 and 60. Considering that the typical length scale of the problem is the channel length, we find also that factor p_3 indicating the relative influence of friction terms ranges between 12 and 30. This indicates that the main forces governing the channel hydraulics are propagation and bed friction, which have roughly the same importance.

The surface profile for a steady flow in a channel with uniform slope and shape satisfies the following equation :

$$\frac{\partial h}{\partial x} = \frac{I - \dfrac{Q_w^2}{K_s^2 R_h^{10/3}}}{I - \dfrac{Q_w^2}{g R_h^3}} \tag{9.1}$$

where R_h denotes the hydraulic radius, which is equal to the ratio of the wet section by the wet perimeter (other notations given in table 9.1). Here, as the channel is much wider than deep, we may assume $R_h \simeq h$. Equation 9.1 can be integrated, for instance by a Runge-Kutta method, if some boundary condition on h is supplied. Thus, we obtain a reference solution which can be compared to model forecasts.

Assuming we have no hint at the correct solution, we obtain the initial surface profile by a crude linear interpolation between the upstream and downstream endpoints of the channel. This initial surface profile appears to be far apart from the exact one :

$$\frac{h_{init} - h_{exact}}{h_{exact}} \simeq 60\% \quad \text{for } x = 3300\,\text{m}$$

Initial velocities are set so that the flow is uniform throughout the reach :

$$u(x) = \frac{Q_w}{h(x)}$$

Consequently, at the beginning of the simulation, while we overestimate the water-depths we

are underestimating drastically the velocity :

$$\frac{u_{init} - u_{exact}}{u_{exact}} \simeq -37.5\% \quad \text{for } x = 3300\,\text{m}$$

These initial flow conditions obviously do not satisfy the St-Venant equations.

Table 9.1: Conditions of backwater calculation test

Physical parameters		
L	channel length (m)	5000
W	channel width (m)	200
I	bottom slope (m/m)	9.10^{-4}
Q_{tot}	upstream flow rate ($m^3.s^{-1}$)	240
Q_w	flow by unit width ($m^2.s^{-1}$)	1.2
K_s	Strickler roughness coefficient	40
h_d	prescribed downstream water-depth (m)	2
h_u	initial upstream water-depth (m)	1
Numerical parameters		
Δx	mesh size in the $x-$ direction (m)	20
Δy	mesh size in the $y-$ direction (m)	50
t_0	simulation initial time (s)	0
T_f	simulation final time (s)	4500 to 6000
Δt	time step (s)	10 while $t \leq 3000$ 5 for $t > 3000$

As there is only one significant direction for wave propagation (the longitudinal one), the relevant propagation Courant number appears to be $C_p = \sqrt{gh}\Delta t/\Delta x$. It is in the range [1.57, 2.21] for $\Delta t = 10$s, in the range [0.78, 1.11] for $\Delta t = 5$s. The Courant number for advection belongs to interval [0.3, 0.6] for $\Delta t = 10$s and interval [0.15, 0.3] for $\Delta t = 5$s.

The upstream boundary condition concerns the flow rate by unit width. In order to solve the problem, it is necessary, both for "depth/velocity" and "depth/discharge" formulations, to prescribe also the free surface elevation at this inflow boundary. An adequate complementary condition is obtained by discretizing semi-implicitly the momentum conservation equations (cf 8.71 in section 8.6.1).

At nodes located on closed boundaries, the condition of zero normal gradient of the water level is substituted directly to the equation governing the free surface increment (cf section 8.6.2) and undergoes a first-order spatial discretization. The tangential velocity obeys a perfect slip condition.

9.2.2 Results

Model results have been stored and surveyed every 500s from $t = 500$ s till $t = 6000$s. The quality of the forecasts has been assessed with the help of various error measures, namely :

- $\overline{\delta Q_x}$ and $\max \delta Q_x$ are respectively the average and maximum **relative error on the value of the flow per unit width.**

$$\delta Q_x = \frac{Q_x^{\text{num}} - Q_w}{Q_w}$$

(nb : $\max \delta Q_x$ has been always observed to occur at the downstream boundary of the channel)

- $\overline{\Delta H}$ and $\max \Delta H$ denote respectively the average and maximum value of the **absolute error on the water depth.**

$$\Delta H = \mid h^{\text{num}} - h^{\text{ex}} \mid$$

(nb : we always observed that, at any location along the channel and any time of the simulation, $h^{\text{num}} \geq h^{\text{ex}}$)

- $\overline{\delta h}$ and $\max \delta h$ denote respectively the average and maximum value of the **relative error on the water depth.**

$$\delta h = \frac{\mid h^{\text{num}} - h^{\text{ex}} \mid}{h^{\text{ex}}}$$

Error measures corresponding to the different models applied are summarized in table 9.2. Let us comment first on the results of the models based on a "depth/unit discharge" formulation :

1. Model QH2 converges between $t = 4000$ and 4500s and gives excellent results. Forecasts appear to be insensitive to the choice of parameter γ in the range [0.5, 1]. They are also quite independent of the algorithm chosen to solve the advection step (upwind method, backward characteristic method combined with bilinear interpolation, or characteristic method combined with bicubic interpolation).

2. Model QH1 has a slower rate of convergence (it occurs between 5500 and 6000 s). Its results are slightly more sensitive to the choice of γ but remain independent of the scheme used to compute the advected discharges. The simplifications adopted in QH1, namely the neglect of spatial variation of factor $1 + \gamma f_b \Delta t$ (which depends on the bottom friction), induce a degradation of the model results, notably as regards the discharge preservation. While remaining moderate, errors are approximately ten times larger than errors given by model QH2.

3. We observe a slight degradation of the results of model QH0 with respect to the results generated by model QH1. However, both models display the same behaviour as regards convergence rate, sensitivity to γ, etc ...

4. The simplification proposed by (Dan N'Guyen, 1988) as regards the advective terms is obviously not advisable. When we compare results of models DAN0 to DAN2 to results yielded by their counterparts based on the treatment of advective terms as suggested in (Benque *et al.*, 1982), namely QH0 to QH2, we observe that, while the errors on the preservation of the unit discharges are similar, errors on the computation of the surface profile are considerably more important. (For instance, errors produced by UH2 are 50 times larger than errors yielded by QH2).

As the trials performed with "depth/discharge" formulations indicated that a correct treatment of factor $1 + \gamma f_b \Delta t$ variability was essential to ensure accurate results, we applied only model UH2. Results raised by this "depth/velocity" formulation are close to results obtained with the best version QH2 but they are achieved with a slower rate of convergence (between $t = 5500$ and 6000s). This time, forecasts appear to be sensitive to the advection step algorithm : upwind method and backward characteristic method with bilinear interpolator produce identical forecasts which are significantly worse than those obtained when the backward characteristic scheme is combined with a bicubic interpolator. This indicates that in this test case, the computation of accurate advected water-depths is more critical than the computation of advected velocities.

In addition to table 9.2, table G.1 (cf appendix G.1) allows a direct comparison of exact and computed water-depths at different locations along the channel. We also mentioned water-depths computed by TELEMAC, as they are reported in (Galland & Hervouet, 1988).

Model forecasts (whatever the version) have turned out to be insensitive to the method applied at the inflow boundary to prescribe the water level (cf alternate option 8.69 in sec. 8.6.1) and, on closed boundaries, to the manner in which the zero gradient condition on the water level is introduced. Letting the tangential velocity be free on closed boundaries does not modify the model outcomes either.

We then investigated results sensitivity to the choice of time step. As regards models based on "depth/unit discharge" formulation, we studied only the model which appeared to achieve the best balance between accuracy and economy, namely QH2 where advection is solved with the upwind method. The maximum allowable time step with the upwind method is 15 s : results obtained for $\Delta t = 10$ then 5, $\Delta t = 10$ and $\Delta t = 15$ s are identical ! Model UH2 based on "depth/unit discharge" formulation is more sensitive to Δt choice as illustrated in table 9.3.

Table 9.2: Error measures : sloping channel test (Δt = 10 then 5s)

Model	t (s)	$\overline{\delta Q_x}$ (%)	max δQ_x (%)	$\overline{\Delta H}$ (mm)	max ΔH (mm)	$\overline{\delta h}$ (%)	max δh (%)
Formulation "depth/discharge" - Advection according to Benque							
QH2	4500	0.04	0.13	0.13	0.50	0.01	0.05
QH2/upwind	4500	0.04	0.13	0.15	0.60	0.01	0.06
QH1	4500	0.44	1.94	1.59	5.40	0.14	0.49
	6000	0.38	1.78	1.40	5.00	0.12	0.45
QH0	4500	0.54	2.18	1.97	6.70	0.17	0.60
	6000	0.46	1.96	1.71	6.10	0.15	0.55
Formulation "depth/discharge" - Advection according to Dan N'Guyen							
DAN2	4500	0.09	0.24	7.01	23.90	0.60	2.11
DAN1	4500	0.53	2.10	8.77	29.50	0.75	2.61
	6000	0.41	1.79	8.35	28.60	0.71	2.52
DAN0	4500	0.65	2.40	9.28	30.90	0.80	2.75
	6000	0.51	2.02	8.76	29.80	0.75	2.63
Formulation "depth/velocity"							
UH2 / bicubic	4500	0.14	0.40	0.51	1.60	0.05	0.15
characteristic	6000	0.05	0.15	0.17	0.70	0.02	0.07
UH2	4500	0.28	0.96	1.04	3.40	0.09	0.32
upwind	6000	0.18	0.71	0.69	2.50	0.06	0.23

Table 9.3: Sensitivity of model UH2 to Δt choice (results at $t = 6000$s)

Δt (s)	$\overline{\delta Q_x}$ (%)	$\max \delta Q_x$ (%)	$\overline{\Delta H}$ (mm)	$\max \Delta H$ (mm)	$\overline{\delta h}$ (%)	$\max \delta h$ (%)
10 for $t \leq 3000$ 5 otherwise	0.05	0.15	0.17	0.70	0.02	0.07
10	0.08	0.28	0.32	1.20	0.03	0.11
15	0.12	0.39	0.47	1.70	0.04	0.16

The CPU time required per time step and per 1000 computational nodes is mainly dependent on the technique chosen to solve the advection stage : $\simeq 0.255$ s when models use the backward characteristic method (whatever the chosen interpolator), 0.095 s when they use the upwind method (*hardware : SUN SPARC Station 10, language : FORTRAN*). Time consumed by the solution of the propagation stage is approximately 0.086 s per time step and per 1000 computational nodes. When the backward characteristic method is applied, three-quarters of the computational effort within the advection step are devoted to the calculation of trajectories.

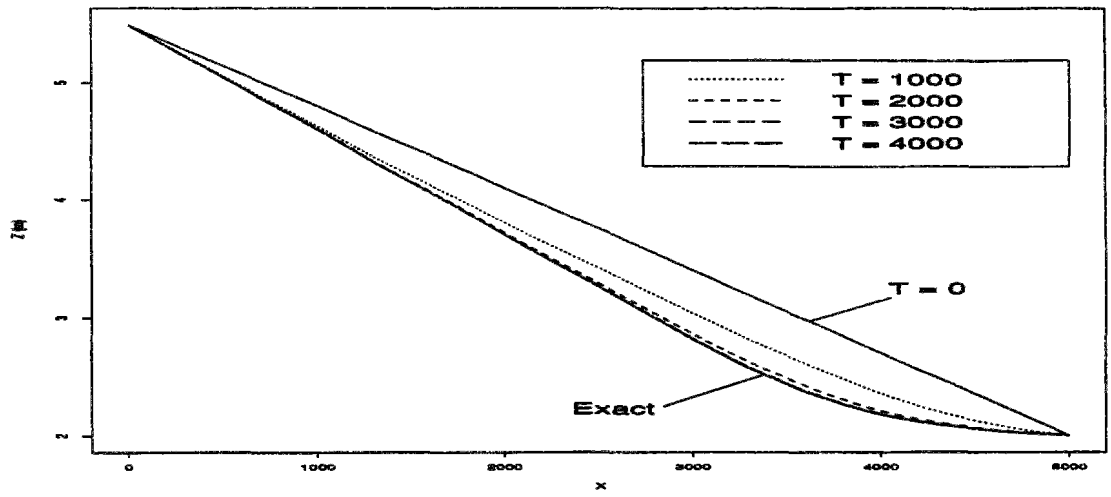


Figure 9.1: Evolution of free surface profile in sloping channel (model QH2)

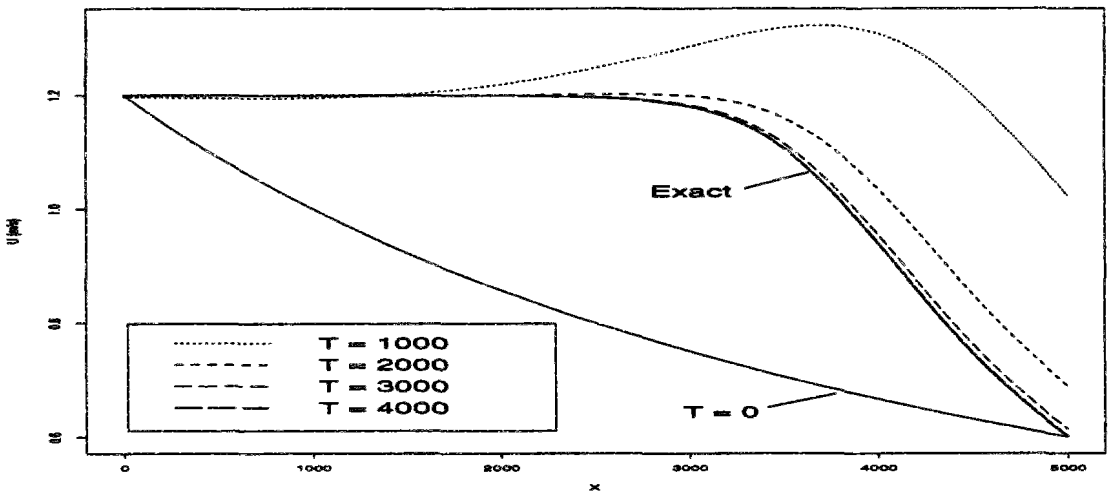


Figure 9.2: Evolution of velocity profile in sloping channel (model QH2)

9.3 Fluvial flow over a sill

9.3.1 Presentation

This second test is also extracted from (Galland & Hervouet, 1988) and has similarly one-dimensional features. It gives some indications about the ability of the model to cope with bathymetric specificities and to represent the balance established between inertial and propagation terms in a frictionless situation.

The test deals with the computation of a steady flow in a channel where there exists a parabolic sill, the bottom being flat otherwise. According to the data (cf table 9.4), the Froude number is about 0.5 - 0.6. Consequently, propagation terms do not outweigh inertial ones (cf sec. 8.2).

When there is no friction, an analytical solution can be derived. Indeed, the specific energy of the flow is preserved. Let Q_w be the flow per unit width prescribed at the upstream inflow boundary and let h_d be the downstream boundary condition on the water-depth. The surface profile satisfies :

$$\frac{Q_w^2}{2\,g\,h^2(x)} + h(x) + z_b(x) = \mathcal{H}_0 = \frac{Q_w^2}{2\,g\,h_d^2} + h_d + z_{b\,d}$$

(nb : we recall that z_b denotes the river bottom elevation and subscript d denotes values at the downstream boundary). Consequently, h can be computed at every location x along the channel by solving merely a third-order equation (this was done with the help of function UNIROOT in mathematical and graphical package Splus).

The initial conditions are : flat free surface profile ($\zeta = 2$ m), uniform unit-discharge ($Q_x = Q_w$, $Q_y = 0$). Upstream boundary conditions on free surface elevation and closed boundary conditions are managed as in the previous test.

The model has been run with different time steps in order to assess its sensitivity with respect to the propagation Courant number. Table 9.5 indicates the tested Δt and gives the corresponding range of advection (C_r) and propagation (C_p) Courant numbers. Once again, as there is only one significant wave propagation direction, we consider $C_p = \sqrt{g\,h}\Delta t/\Delta x$. As the surface profile evolves quickly, the time steps are small. Simulations were performed till the surface profile reached an equilibrium.

Table 9.4: Conditions of sill test

Physical parameters		
L	channel length (m)	20.5
W	channel width (m)	2
z_b	bed elevation (m)	0 if $x \leq 8$ $0.2 - 0.05\ (x-10)^2$ if $8 < x < 12$ 0 if $x \geq 12$
Q_w	flow by unit width ($m^2.s^{-1}$)	$\sqrt{2g} = 4.4294$
Q_{tot}	upstream flow rate ($m^3.s^{-1}$)	8.8589
K_s	Strickler roughness coefficient	∞
h_d	prescribed downstream water-depth (m)	2
Numerical parameters		
Δx	mesh size in the $x-$ direction (m)	0.5 for $x \leq 7$ and $x \geq 13$ 0.2 otherwise
Δy	mesh size in the $y-$ direction (m)	0.2
t_0	simulation initial time (s)	0

Table 9.5: Time steps for sill test

Δt (s)	$\min C_p$	$\max C_p$	$\min C_r$	$\max C_r$
0.2	1.772	4.43	0.886	2.595
0.1	0.886	2.215	0.443	1.298
0.05	0.443	1.108	0.221	0.649
0.01	0.0886	0.221	0.0443	0.130

9.3.2 Results

First we may notice that the simplification proposed by Dan N'Guyen as regards the advective terms must be rejected. Indeed, in that case, the initial hydraulic conditions satisfy the simplified flow equations, so that the hydraulic variables do not depart from this initial, erroneous, state. Since there is no friction, model versions QH2 and QH1 on the one hand, UH2 and UH1 on the other hand are equivalent.

During the simulations, forecasts were stored and error measures computed every 5 s. This allowed survey of the convergence of the models. The error measures are the same as for previous test except that this time, averages $\overline{\delta h}$, $\overline{\delta Q_x}$, $\overline{\Delta H}$ have been computed on area $7 \leq x \leq 13$ which encloses the sill (mean error values taken on the whole computational domain are slightly inferior). These error values are indicated in tables 9.6 to 9.9 (nb : computed water-depths are given in tables G.3 and G.4, appendix G.2).

In summary, the forecasted surface profiles are found to display the following inaccuracies, whose magnitude depends on the time step : phase shift with respect to the exact surface profile, heightening of the surface profile upstream of the sill, occurrence of spurious overshoot downstream the sill. Besides, the sill induces some perturbation of the flow conservation, whose shape, extent and amplitude depend both on the time step and on the tested model.

1. **Rate of convergence**
   Models QH2 and QH0 behave similarly : for all time steps, the free surface profile becomes stable between $t = 55$ and 60 s. The value of the implicitation parameter γ has no influence on their convergence rate. It has neither impact on forecasts yielded by model QH2. Model QH0 is more sensitive to γ value : as γ is set closer to 1, results tend to worsen. This is not much significant for $\Delta t \leq 0.10$ s but it is pretty dramatic for $\Delta t = 0.20$ s.

   Model UH2 converges faster than "depth/unit discharge" models, except for the smallest time step : stabilization occurs between 40 and 45 s for $\Delta t = 0.2$ and 0.1 s, between 45 and 50 s for $\Delta t = 0.05$ s.

2. **Comparison of models QH2 and QH0**
   For this comparison, the advection step was solved in both models with the backward characteristic method combined with a bicubic interpolator.

   Reduction of the time step improves the forecasts of both QH2 and QH0.

   The simplifications embedded in model QH0 induce a degradation of flow conservation, as illustrated in figures 9.4, 9.6, 9.8 and 9.10.

   They also worsen the free surface profile estimation. This is not obvious when considering the error measures (tables 9.6 to 9.9) but it becomes quite plain when looking at

the computed surface profiles (except perhaps for the smallest time step). In particular, on figures 9.3, 9.5 and 9.7, we can observe that profiles computed by QH0 display unacceptable spurious oscillations immediately downstream of the sill. For all time steps except the smallest, the surface profile phase shift is slightly less important with model QH0. On the other hand, both upstream extra elevation and downstream overshoot are always larger and the deepest point of the free surface is not so well approximated as with QH2.

3. **Comparison of models QH2 and UH2**
Once again, we compare the results when advection is solved with the backward characteristic method and bicubic interpolator.

As for model QH2, model UH2 improves steadily as the time step is decreased and is fairly insensitive to the implicitation parameter choice. However, **model QH2 based on a "depth/unit discharge" formulation performs consistently better than model UH2 based on a "depth/velocity" formulation.** For the same time step, error measures relative to the surface profile estimation are approximately twice as big with UH2 as with QH2. Error measures relative to flow preservation are between twenty times ($\Delta t = 0.2$) and twice ($\Delta t = 0.01$) as big : the difference between the two models tends to decrease with the time step.

For $\Delta t = 0.2$, the phase shift is approximately two mesh sizes for QH2, 3 to 4 mesh sizes for UH2. For both models, it is halved when the time step is halved, first to 0.1 then to 0.05s. For $\Delta t = 0.01$s, both QH2 and UH2 forecasted surface profiles are nearly perfect except for the slight overshoot immediately downstream of the sill, at $x = 12.2$ (1.9 and 3.6 mm high for QH2 and UH2 respectively). In fact, this overshoot, which is apparent only on figures 9.7 (3 mm for both models) and 9.9, occurs also for the two largest time steps : because of the greater phase shift plaguing the computed profiles, it is then located further downstream of the sill and it is lower. A closer observation of the profiles highlights in fact that the computed profiles display, in an area one meter long downstream of this overshoot, very slight oscillations around the equilibrium value $\zeta = 2$m. However, these oscillations are minor, they have much less amplitude than with model QH0.

As water-depths computed by the model TELEMAC between $x = 8$ and $x = 13.2$ m were indicated in (Galland & Hervouet, 1988), we could also plot these results. It appears that TELEMAC, which relies also on a "depth/velocity" formulation, has forecasts fairly close to those of UH2, except for $\Delta t = 0.01$: in this case, the TELEMAC results are no better than with $\Delta t = 0.05$ s.

The behaviour of models QH2 and UH2 as regards flow conservation is quite different (apart from the different scale of their errors). Indeed, with QH2, the sill perturbs the flow conservation in very limited areas, corresponding to each extremity of the sill (cf figures 9.4, 9.6 and 9.8). With UH2, it is a different story. In the upstream part of

the channel, the flow is correctly preserved. Then, the flow is underestimated over the upstream half of the sill and, on the contrary, is overestimated over its downstream half. These errors are not symmetric : overestimation is more marked than underestimation; besides, it lingers downstream of the sill. There the flow rate converges once again towards a uniform value, which is systematically slightly superior to the upstream imposed flow rate.

Finally, oscillations displayed by the computed flow rates on figure 9.10 might seem impressive at first glance. However, beware the scale of the drawing : they amount to less than 0.2 % of the upstream discharge. Thus, flow preservation can be deemed excellent for both QH2 and UH2 with $\Delta t = 0.01$ s. The wavelength of these oscillations is about 2 mesh sizes : it seems that we are witnessing here a well known numerical phenomenon which plagues sometimes models using a non-staggered grid, namely the Gibbs oscillations (Roache, 1985).

4. **Influence of the treatment of the advection step**
This was studied both for models QH2 and UH2.

For $\Delta t \geq 0.10$ s, only characteristic methods can be applied as the advection Courant number may exceed 1. Use of the bilinear interpolator instead of the bicubic one increases but slightly the errors on the computed surface profiles. With QH2, there is neither any significant consequence on the computed flow rates. As in the previous test, UH2 is more sensitive than QH2 to the chosen technique of interpolation.

For smaller time steps, the economical upwind method can also be tested. As regards QH2, bilinear interpolation and the upwind method produce nearly indistinguishable results. It is not the case with UH2 (cf table 9.8).

For $\Delta t = 0.05$ and model QH2, all three options for the treatment of the advection step perform nearly as well. On the contrary, UH2 forecasts with the bicubic interpolator are significantly better than with the two other advection solution techniques.

The sensitivity of models to the solution of the advection step is in fact really apparent with the smallest time step $\Delta t = 0.01$ s (cf table 9.9). More time steps are needed till the final convergence time. The errors generated each time the advection step is solved sum up and eventually have a negative influence on the final forecast accuracy : upwind and bilinear methods perform significantly worse than the bicubic interpolator. In fact, as regards surface profile evaluation, they even generate poorer results than with $\Delta t = 0.05$ s.

5. **Computational costs**
The computational costs depend on the solution of the advection step as the solution of the propagation step has the same cost in all model versions (QH2, QH0, UH2). When a

backward characteristic method is applied, the costs depend also on the time step. Indeed, in order to maintain a good precision, the computation of the trajectory is broken into segments according to the number of mesh cells a particle crosses during a single time step (cf appendix E.1.1) : the greater the advective Courant numbers, the more lengthy particle backtracking becomes.

When the backward characteristic method is applied in combination with the bicubic interpolator, the CPU time consumed per time step and per 1000 grid nodes is respectively 0.391 s, 0.318 s, 0.250 s for $\Delta t = 0.2$s, $\Delta t = 0.1$s and $\Delta t = 0.05$ & 0.01s respectively. The solution of the advection step consumes between 80 and 70 % of the total computational time.

When the bilinear interpolator is applied, the savings amount approximately to 0.02 s per time step and per 1000 nodes. When the upwind method is applied, the CPU cost is 0.082 s.

In conclusion, this test confirms that applying simplifications to the governing equations of the propagation step is not advisable.

Further trials, commented upon in appendix G.2, demonstrate that here the main source of troubles lies in the inability of the models to deal perfectly with bathymetric discontinuities inducing sharp gradients of the dependent variables. Dealing with these sharp gradients is a challenge both during the advection step (cf point 4 above) and during the propagation step, whose equations are discretized with centred space differences, ill-suited when the variables display strong curvature.

Once again, model based on a "depth/unit discharge" formulation fares better than model based on a "depth/velocity" formulation. This is probably due to the splitting of the continuity equation and to the fact that advecting water-depths which display strong gradients constitutes an additional source of error when solving the flow equations. Yet, even if forecasts yielded by QH2 are superior, model UH2 results can also be deemed fairly good for time steps $\Delta t \leq 0.05$.

Table 9.6: Error measures - parabolic sill test - $\Delta t = 0.20$s

Model	$\overline{\delta Q_x}$ (%)	max δQ_x (%)	δQ_{av} (%)	$\overline{\Delta H}$ (mm)	max ΔH (mm)	$\overline{\delta h}$ (%)	max δh (%)
QH0	0.72	1.88	0.24	6.8	15.7	0.37	0.83
QH2/bicubic	0.06	0.45	0.05	12.3	25.9	0.67	1.42
QH2/bilinear	0.06	0.45	0.07	12.8	27.1	0.70	1.49
UH2/bicubic	1.25	3.45	0.80	23.0	47.2	1.26	2.67
UH2/bilinear	1.30	3.57	0.84	23.8	48.6	1.29	2.75

Table 9.7: Error measures - parabolic sill test - $\Delta t = 0.10$s

Model	$\overline{\delta Q_x}$ (%)	max δQ_x (%)	δQ_{av} (%)	$\overline{\Delta H}$ (mm)	max ΔH (mm)	$\overline{\delta h}$ (%)	max δh (%)
QH0	0.36	0.97	0.02	4.0	12.1	0.21	0.62
QH2/bicubic	0.06	0.37	0.07	6.1	13.2	0.33	0.69
QH2/bilinear	0.06	0.37	0.06	7.1	15.2	0.39	0.80
UH2/bicubic	0.69	1.79	0.28	12.2	24.9	0.66	1.35
UH2/bilinear	0.79	2.07	0.47	14.0	28.8	0.76	1.58

Table 9.8: Error measures - parabolic sill test - $\Delta t = 0.05$s

Model	$\overline{\delta Q_x}$ (%)	max δQ_x (%)	δQ_{av} (%)	$\overline{\Delta H}$ (mm)	max ΔH (mm)	$\overline{\delta h}$ (%)	max δh (%)
QH0	0.19	0.53	0.04	2.6	9.9	0.14	0.51
QH2/bicubic	0.06	0.28	0.04	3.3	9.2	0.18	0.47
QH2/bilinear	0.06	0.27	0.05	4.7	10.6	0.26	0.56
UH2/bicubic	0.38	0.98	0.12	6.7	15.6	0.36	0.82
UH2/bilinear	0.54	1.40	0.08	9.5	20.1	0.51	1.07
UH2/upwind	0.59	1.54	0.26	10.3	22.6	0.56	1.19

Table 9.9: Error measures - parabolic sill test - $\Delta t = 0.01$s

Model	$\overline{\delta Q_x}$ (%)	$\max \delta Q_x$ (%)	δQ_{av} (%)	$\overline{\Delta H}$ (mm)	$\max \Delta H$ (mm)	$\overline{\delta h}$ (%)	$\max \delta h$ (%)
QH0	0.07	0.20	0.15	0.8	3.5	0.04	0.18
QH2/bicubic	0.05	0.22	0.13	0.9	3.6	0.05	0.18
QH2/bilinear	0.06	0.18	0.06	5.1	12.7	0.28	0.65
UH2/bicubic	0.10	0.45	0.05	1.5	6.8	0.08	0.35
UH2/upwind	0.59	1.60	0.29	10.3	22.7	0.56	1.19

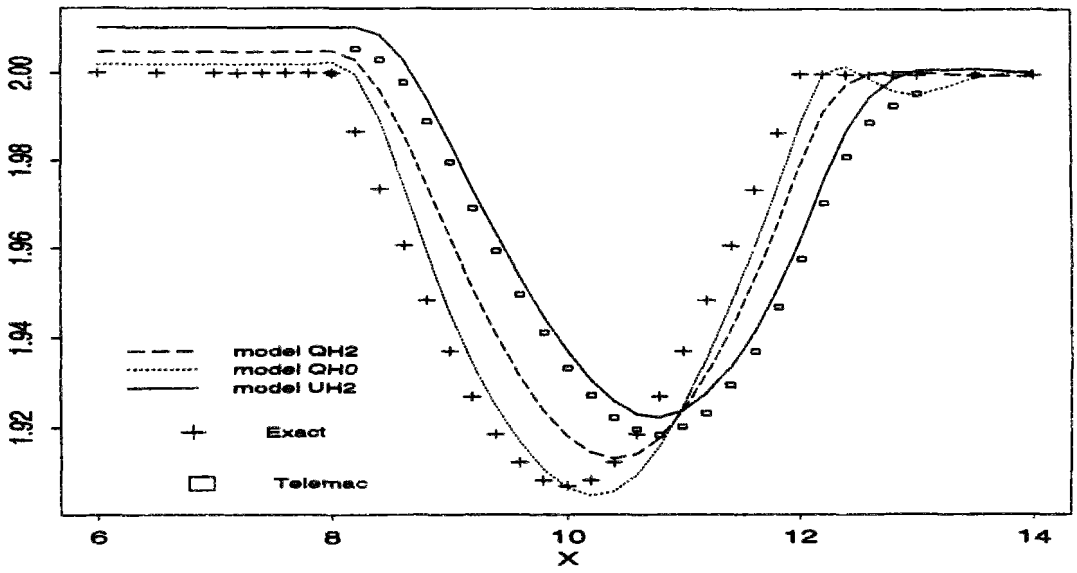


Figure 9.3: Computed free surface elevations / sill test : $\Delta t = 0.20$ s

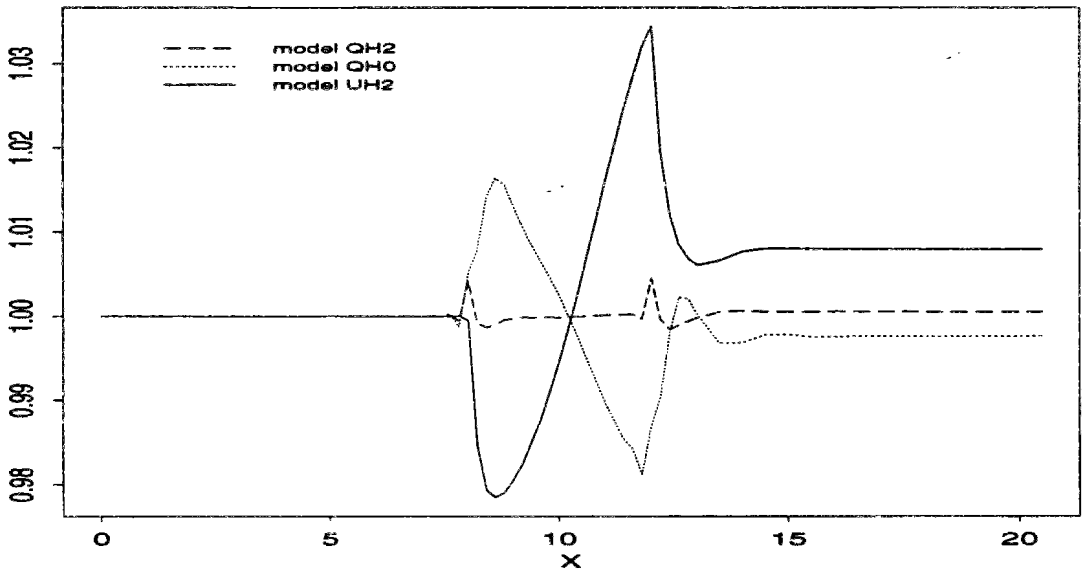


Figure 9.4: Ratios of computed by exact flow rate / sill test : $\Delta t = 0.20$ s

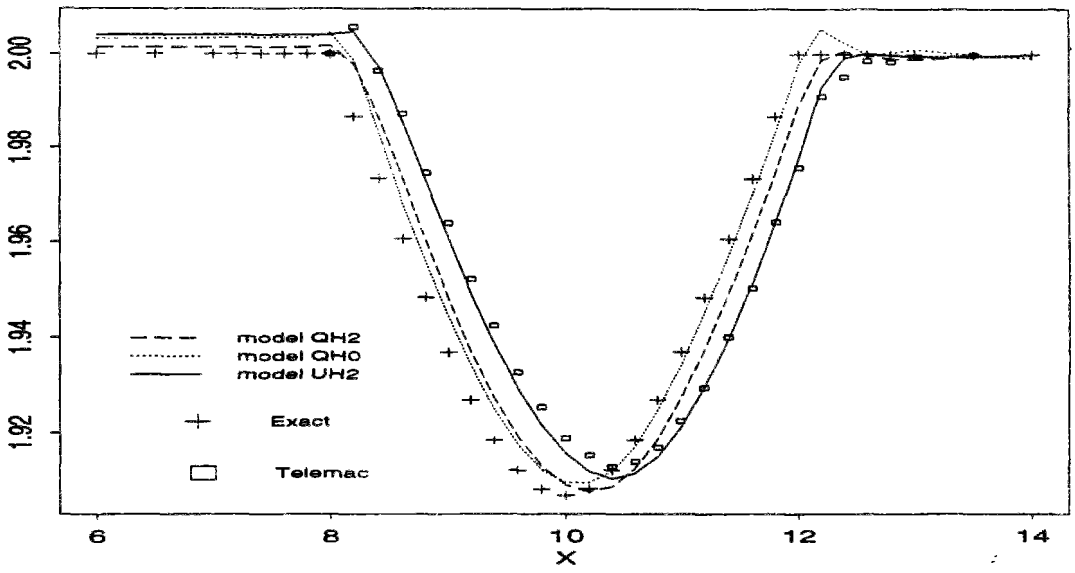


Figure 9.5: Computed free surface elevations / sill test : $\Delta t = 0.10$ s

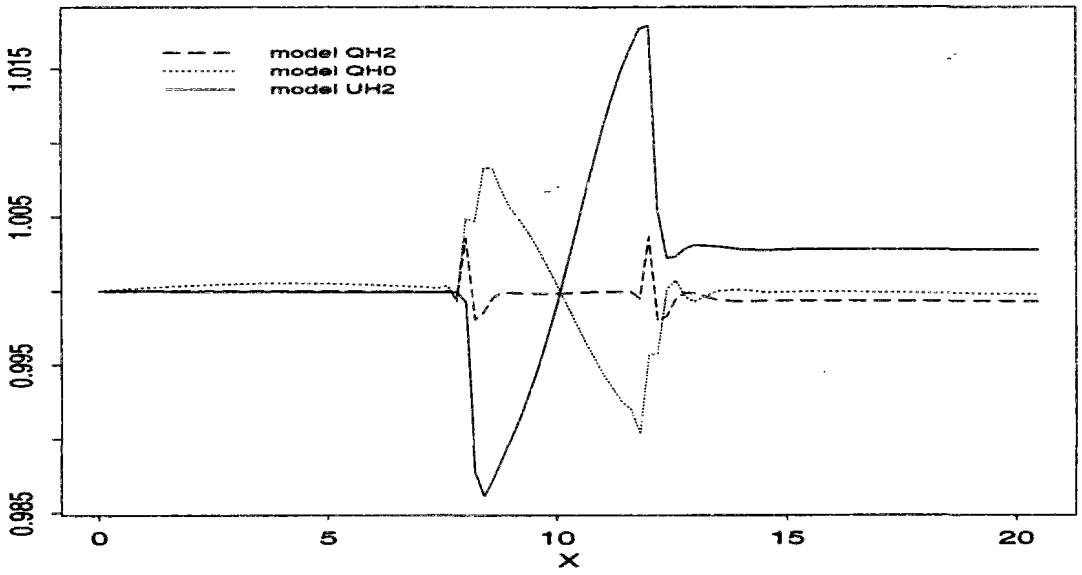


Figure 9.6: Ratios of computed by exact flow rate / sill test : $\Delta t = 0.10$ s

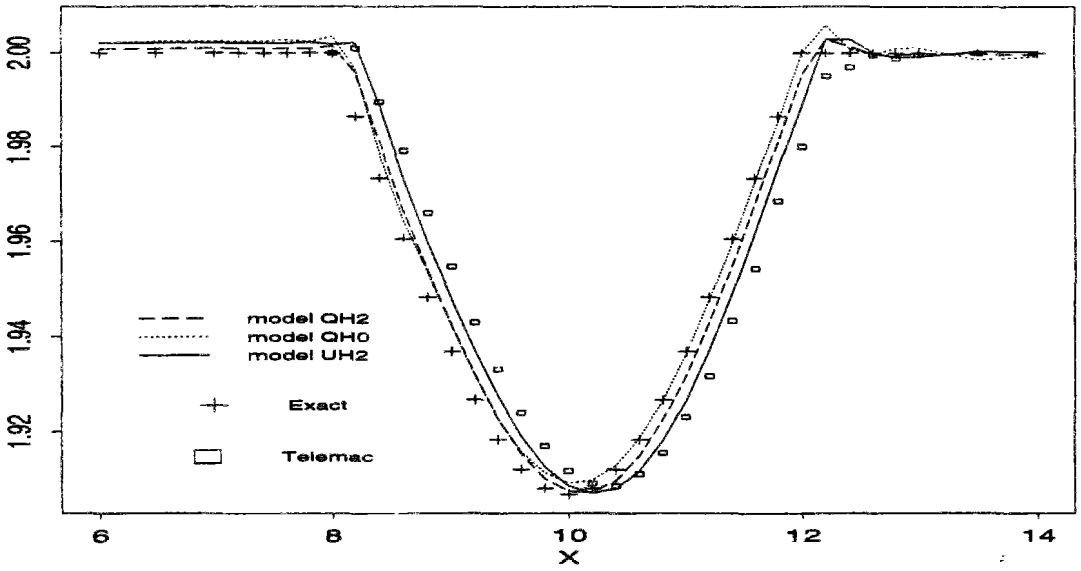


Figure 9.7: Computed free surface elevations / sill test : $\Delta t = 0.05$ s

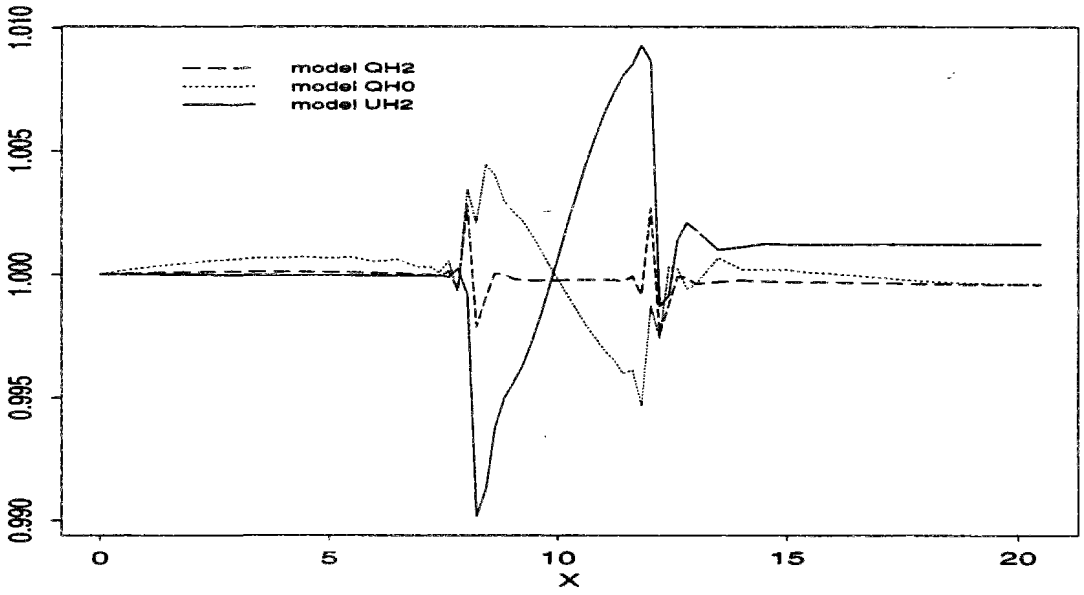


Figure 9.8: Ratios of computed by exact flow rate / sill test : $\Delta t = 0.05$ s

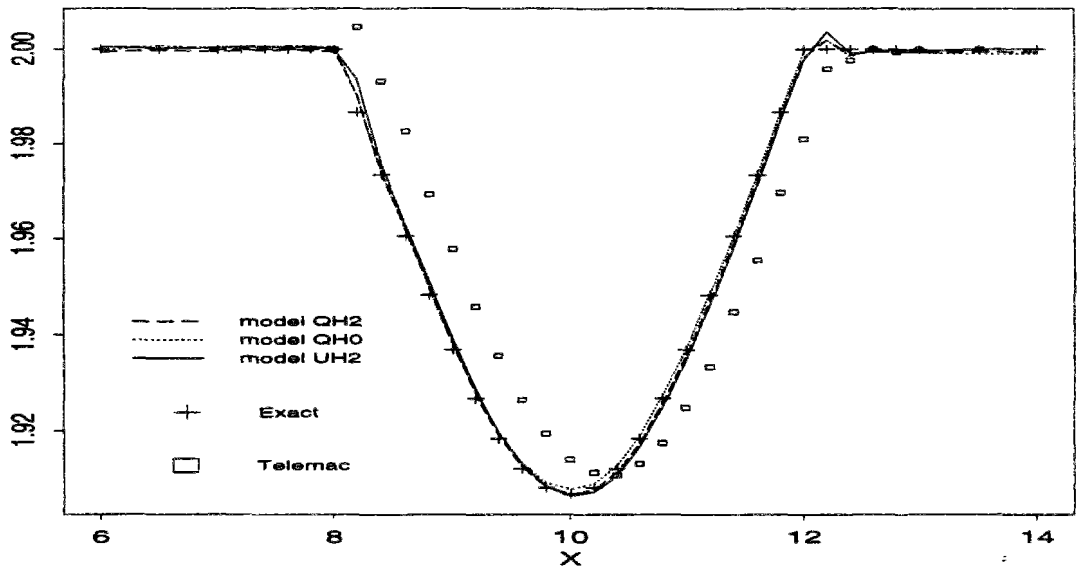


Figure 9.9: Computed free surface elevations / sill test : $\Delta t = 0.01$ s

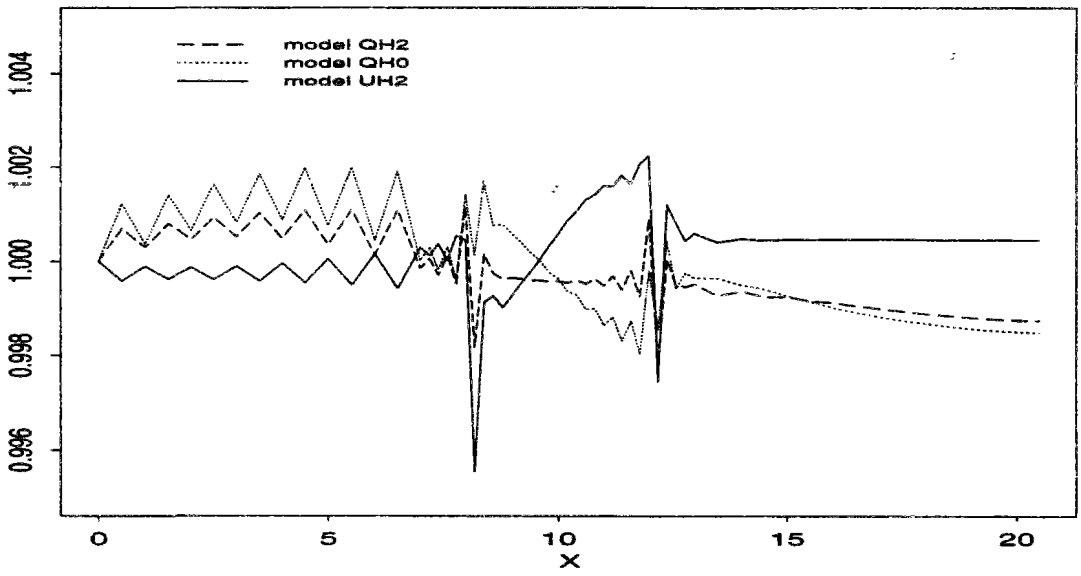


Figure 9.10: Ratios of computed by exact flow rate / sill test : $\Delta t = 0.01$ s

9.4 Conclusions

The two tests performed here allow the following conclusions :

1. It is essential to deal properly with the equations governing the propagation step. As minor as they may seem at first sight, simplifications as proposed in (Dan N'Guyen, 1988) induce systematically a significant loss of accuracy.

   These tests correspond to situations where either bottom friction is important or water-depths display strong gradients. It is possible we would observe less difference between simplified and complete models where friction is less significant and/or surface profiles smoother, for instance when dealing with marine applications. However, complete equations are not really more complex to write down and discretize, nor more costly to solve. Thus, there is no reason for not using them systematically.

2. On the whole, results yielded either by the "depth/discharge" model QH2 or by the "depth/velocity" model UH2 are satisfying, even if the excellent performance of QH2 casts some shadow on the lesser performance of UH2.

   Possible reasons for the discrepancies between models QH2 and UH2 are numerous. First we may notice that in both tests the boundary conditions concern the flow rates, so that choosing the velocity components as dependent variables necessitates some extra manipulations, which provide additional sources of errors. However, as proves the similarity between UH2 and Telemac results, the major source of inaccuracy lies probably in the splitting of the continuity equation applied in the "depth/velocity" formulation. This appears to have notably negative impacts on flow conservation.

   Yet we must not forget that, on the other hand, the "depth/discharge" model QH2 has a slight problem with the solution of the advection step. Indeed, as discussed in section 8.3.1.2, the advected discharges are estimated with a systematic error of approximately $(h^{n+1} - h^n) . \vec{U}$, besides all other errors which may stem from the backtracking, spatial interpolation, etc ... This has no consequence when dealing with steady-state flows but could be critical when simulating rapidly varying unsteady flows.

   Consequently, we decided to pursue our tests, not only with the most successful model version till now, namely QH2, but also with model UH2.

9.5 Résumé français : "Cas-tests sur écoulements permanents"

Comme dans le cas de l'étude des solutions de l'équation d'advection-diffusion, nous sommes confrontés au problème de trouver des cas-tests qui nous permettent d'évaluer objectivement les modèles et ne soient pas trop éloignés des conditions réelles d'application de ceux-ci. Malheureusement, les problèmes dont on peut apprécier a priori la solution de façon quantitative (existence d'une solution analytique) ou qualitative (grandes caractéristiques majeures de l'écoulement) avec un bon degré de fiabilité sont rares ! Divers cas-tests ont été introduits dans la littérature mais on ne note pas d'effort de synthèse comparable à celui qui a débouché sur le cahier de tests du Forum Convection-Diffusion. L'effort le plus poussé en ce domaine a été réalisé par Électricité de France et s'est concrétisé dans le "cahier de validation" du code bidimensionnel TELEMAC (Galland & Hervouet, 1988), document dans lequel nous avons largement puisé.

Les deux problèmes que nous étudions dans ce chapitre concernent les situations les plus proches de l'hydraulique fluviale que nous ayons pu trouver. Ils ne permettent bien sûr pas une analyse complète :

- Il s'agit d'écoulements **permanents** et **unidirectionnels**.
- Il s'agit de cas où l'on néglige la diffusion horizontale de quantité de mouvement. Heureusement, ceci est assez courant dans les problèmes fluviaux.

Cependant, ils ont des avantages indéniables :

- L'un comme l'autre sont **proches de problèmes typiques** (calcul de ligne d'eau sous influence aval, passage d'une singularité bathymétrique).
- Ils possèdent une **solution analytique**, ce qui permet d'évaluer en tout objectivité la précision des simulations. De plus, comme il s'agit de calculer un état stationnaire, on pourra aussi apprécier la vitesse de convergence des algorithmes.
- Enfin, il faut noter que l'approche que nous utilisons (pas fractionnaires et spécificité du traitement de l'étape de propagation), originellement suggérée par (Dan N'Guyen, 1988), a été jusqu'ici validée essentiellement pour des écoulements marins et côtiers, où le terme de propagation est trés largement dominant. Les deux tests sélectionnés concernent des situations plus équilibrées où propagation, frottement ou advection sont du même ordre de grandeur. Nous verrons que ceci nous permettra d' **observer des phénomènes numériques nouveaux**.

Pour des raisons de simplicité nous utiliserons des abbréviations pour désigner les différentes versions de modèles présentées au chapitre précédent :

- formulation "hauteur/vitesse". UH2 désigne le modèle complet, sans aucune simplification concernant l'équation en δz de l'étape de propagation (eq. 8.31). UH1 renvoie à la version obtenue en négligeant les variations spatiales du facteur de frottement f_b (cf eq. 8.32).
- formulation "hauteur/débit". Quand aucune simplification n'est appliquée à l'étape de propagation (utilisation eq. 8.36) les modèles sont désignés respectivement par les sigles QH2 et DAN2 selon

que l'étape d'advection est résolue suivant l'approche proposée par (Benque *et al.*, 1982) (modèle CYTHERE) ou (Dan N'Guyen, 1988) (modèle ECOMOD).
Quand on applique un niveau de simplification intermédiaire (hypothèse sur f_b, eq. 8.37) on obtient les versions QH1 et UH1.
Quand on simplifie encore plus (cf eq. 8.38, (Dan N'Guyen, 1988)), on a finalement les modèles QH0 et UH0.

Calcul d'une courbe de remous (section 9.2)

Le premier test concerne le calcul classique d'une ligne d'eau, le débit amont et la cote avale étant imposés (cette dernière impliquant une profondeur supérieure à la profondeur normale).

- On considère un canal rectiligne uniforme de longueur 5000 m, largeur 200 m et pente $9\,10^{-4}$ m/m ($z_f(0) = 4.5$m, $z_f(5000) = 0$). La rugosité est paramétrée par un coefficient de Strickler K_s d'une valeur de 40.
- Le débit amont Q_x par unité de largeur est de 1.2 m^2/s, ce qui correspond à une profondeur normale de 1 m. La cote aval est égale à 2 m. Si l'on néglige la diffusion de quantité de mouvement, ce type d'écoulement uniforme admet une solution calculable analytiquement (en résolvant eq. 9.1).
- Les conditions initiales de la simulation correspondent à une méconnaissance totale de la ligne d'eau : celle-ci est tout simplement interpolée linéairement de l'amont à l'aval. Par suite, les erreurs relatives sur la hauteur d'eau et la vitesse atteignent des maxima respectifs de +60% et −37.5%.
- Le pas d'espace Δx dans le sens de l'écoulement est de 20 m. La simulation démarre à $t = 0$ s. Le pas de temps est tout d'abord fixé à 10 s, jusqu'à $t = 3000$, puis passe à 5 s. Les nombres de Courant de propagation et d'advection varient, pour $\Delta t = 10s$, respectivement entre 1.6 et 2.2, et entre 0.3 et 0.6. La simulation est poursuivie jusqu'à convergence vers une ligne d'eau stable.

Compte tenu des conditions ci-dessus, le nombre de Froude de l'écoulement varie le long du canal entre 0.13 et 0.40. On en déduit que le facteur adimensionnel p_1 qui traduit le poids des termes de propagation par rapport aux termes inertiels est compris entre 6 et 60. En considérant que l'échelle de longueur typique du problème est la longueur du canal on obtient de plus que p_3, quantifiant le poids des termes de frottement par rapport aux termes inertiels, est dans la gamme 12 à 30. Par conséquent, on se trouve dans un cas où **les forces qui gouvernent en premier lieu l'hydraulique du canal sont la propagation et le frottement**, qui excèdent en gros d'un facteur 10 les termes inertiels.

Nous discutons en premier lieu les performances des différentes versions en **formulation "hauteur-débit"**. Celles-ci sont évaluées en étudiant les erreurs absolues et relatives entre hauteurs d'eau et débits calculés et solution analytique :

1. **Les meilleurs résultats sont produits par la version QH2, c'est à dire telle que l'étape d'advection est traitée comme dans (Benque *et al.*, 1982) et qu'aucune simplification n'affecte l'étape de propagation.** L'erreur relative sur le débit est en effet au maximum de 0.13%, en moyenne de 0.04%. Quant à l'erreur relative sur la hauteur elle atteint 0.05% au maximum, 0.01% en moyenne. La convergence est achevée à $t = 4500$ s.
2. Les tableaux 9.10 et 9.11 résument les performances des autres versions par rapport à QH2 (voir aussi table 9.2). On y a porté le ratio des erreurs relatives de chaque version par rapport à celles engendrées par QH2. (nb : dans les cases, le chiffre en haut à gauche correspond au ratio des erreurs moyennes, celui qui est en bas à droite au ratio des erreurs maximales)

Table 9.10: Ratio des erreurs relatives moyennes et maximales sur Q (canal)

	Versions 2 Pas de simplification	**Versions 1** Simplification frottement f_b	**Versions 0** Simplification hauteur
Versions QH Advection CYTHERE	1 / 1	9.5 / 13.7	11.5 / 15.1
Versions DAN Advection ECOMOD	2.3 / 1.9	10.3 / 13.8	12.8 / 15.5

Table 9.11: Ratio des erreurs relatives moyennes et maximales sur h (canal)

	Versions 2 Pas de simplification	**Versions 1** Simplification frottement f_b	**Versions 0** Simplification hauteur
Versions QH Advection CYTHERE	1 / 1	12 / 9	15 / 11
Versions DAN Advection ECOMOD	60 / 42	71 / 50	75 / 53

De ces tables on déduit immédiatement que :

- **Le traitement des termes advectifs tel qu'il est proposé par (Dan N'Guyen,**

1988) est peu recommandable.

- **Simplifier l'équation gouvernant les incréments δz dans l'étape de propagation entraîne une nette dégradation des résultats. C'est l'hypothèse portée sur le frottement qui est particulièrement préjudiciable.** (En effet, l'accroissement des erreurs est trés important entre les versions 2 et 1, moindre des versions 1 à 0).

Les versions 1 et 0 convergent également moins rapidement ($\simeq$ 1000 s plus tard) que les versions 2.

3. On note également que les performances de QH2 sont quasiment identiques que l'on résolve l'étape d'advection avec une méthode aux caractéristiques bicubique, bilinéaire ou par la méthode UPWIND (cf table 9.2). Elles sont également insensibles au choix du pas de temps (cf simulations réalisées avec un Δt constant de 10 ou 15 s).

Fort de ces résultats sur les formulation "hauteur-débit", **nous n'avons évalué que la version la plus complète des formulations "hauteur-vitesse", à savoir UH2. Ses résultats sont équivalents à ceux de QH2 mais la convergence est plus lente** (achevée vers $t = 6000$ s). UH2 est également plus sensible à l'algorithme de résolution de l'advection (cf tab. 9.2) (erreurs multipliées en gros par 4 quand on passe de l'interpolateur bicubique à l'interpolateur bilinéaire ou la méthode UPWIND). UH2 est de même plus affecté par l'augmentation du pas de temps (cf tab. 9.3) (erreurs presque doublées pour Δt constant à 10 s, multipliées par 2.5 pour Δt à 15 s).

Passage d'une singularité (seuil) (section 9.3)

Ce second test permet d'évaluer l'aptitude des modèles à s'accommoder d'une singularité bathymétrique.

- On considère un canal rectiligne de longueur et largeur respectivement égales à 20.5 et 2 m (cf table 9.4). Ses fonds sont plats ($z_f = 0$, pente nulle) à l'exception d'un seuil parabolique situé entre les abscisses $x = 8$ et $x = 12$ m, de hauteur maximale 20 cm. On suppose le frottement nul donc **les seules forces en présence seront les termes inertiels et de propagation.**
- Là encore, les conditions limite sont un débit amont par unité de largeur imposé et une cote aval fixée ($Q_x = \sqrt{2g} = 4.4294$ m^2/s et $\zeta = 2$ m). ζ dans le canal est alors régie par une équation qui peut être résolue analytiquement, ceci produisant une solution de référence (cf section 9.3.1).
- La ligne d'eau initiale est plate, les vitesses en sont déduites pour assurer l'uniformité du débit. La ligne d'eau va trés rapidement se déformer pour accuser un creux au niveau du seuil.
- Le maillage adopté est variable. Il est raffiné au passage du seuil ($\Delta x = 0.2$ m pour $7 \leq x \leq 13$, $\Delta x = 0.5$ m sinon).

  On étudie les solutions numériques obtenues pour 4 valeurs différentes du pas de temps, à savoir

Δt = 0.2, 0.1, 0.05 et 0.01 s. Pour $\Delta t = 0.2$, le nombre de Courant de propagation C_p varie entre 1.8 et 4.4, le nombre de Courant d'advection C_r entre 0.9 et 2.6 (cf table 9.5).
Les simulations sont pousuivies jusqu'à stabilisation de la ligne d'eau.

On remarque tout d'abord que **le traitement de l'advection comme suggéré par (Dan N'Guyen, 1988) est inapplicable.** En effet, les conditions initiales satisfont les équations hydrauliques, le terme d'advection étant simplifié. Par suite, l'état de la ligne d'eau ne s'éloigne pas de sa version initiale, erronnée. Par ailleurs, compte tenu de l'absence de frottement, les versions QH2 et QH1 d'une part, UH2 et UH1 de l'autre, sont équivalentes. Le nombre de Froude est de l'ordre de 0.5 à 0.6, il n'y a donc pas prédominance marquée des termes de propagation par rapport aux termes inertiels ($2.8 \leq p_1 \leq 4$).

En résumé, les lignes d'eau calculées présentent les inexactitudes suivantes, d'autant plus prononcées que le pas de temps est grand : **déphasage** par rapport à la ligne d'eau exacte, **surélévation à l'amont** du seuil, présence d'**oscillations parasites à l'aval**. Le passage du seuil induit également quelques problèmes de conservation du débit.

1. **Les résultats les plus satisfaisants sont de nouveau obtenus avec la version "hauteur-débit" complète QH2.** Pour $\Delta t = 0.2$, les erreurs relatives maximum sont de 0.5 et 1.4 % sur débit et hauteur respectivement, les erreurs moyennes sont de 0.06 et 0.67 % respectivement. Avec le plus faible pas de temps $\Delta t = 0.01$, les erreurs relatives maxi et moyenne tombent à 0.05 et 0.22 % sur le débit, à 0.05 et 0.18 % sur la hauteur.

   Les mesures d'erreur (maximum et moyenne des erreurs relatives) peuvent être trompeuses. Elles indiqueraient en effet (cf tables 9.6 à 9.9) un comportement de QH0 systématiquement meilleur que celui de UH2 et globalement meilleur que celui de QH2 en ce qui concerne le calcul des hauteurs. Or, un coup d'oeil aux lignes d'eau (figures 9.3, 9.5, 9.7 et 9.9) montre que **les solutions de QH0 sont celles qui présentent le plus d'oscillations parasites**. La solution pour $\Delta t = 0.2$ en particulier est inacceptable.

2. **L'avantage de QH2 sur son second, la version "hauteur-vitesse" complète UH2, est net, notamment en ce qui concerne la conservation des débits.** La table 9.12 illustre cet avantage. Elle indique le ratio des erreurs relatives engendrées par UH2 et QH2, l'étape d'advection étant résolue dans chaque cas par méthode aux caractéristiques bicubique.

   En ce qui concerne le débit, on notera également que les problèmes de conservation sont, avec QH2, limités aux deux extrémités du seuil, alors qu'avec UH2 et QH0, la perturbation s'étend sur tout le seuil et perdure à son aval (figures 9.4, 9.6 et 9.8). Les résultats produits par QH2 et UH2 pour $\Delta t = 0.01$ sont globalement trés bons mais on décèle tout de même la manifestation de faibles oscillations de grille (fig 9.10).

Table 9.12: Ratio des erreurs relatives d'UH2 et QH2 - test du seuil

	Erreurs sur Q		Erreurs sur h	
	moyenne	maxi	moyenne	maxi
$\Delta t = 0.20$	20.8	7.7	1.90	1.90
$\Delta t = 0.10$	11.5	4.8	2.00	2.00
$\Delta t = 0.05$	6.3	3.5	2.00	1.75
$\Delta t = 0.01$	2.0	2.0	1.60	1.95

3. L'influence du mode de résolution de l'étape d'advection a été évaluée pour QH2 et UH2. Tant que Δt est supérieur à 0.1 seule une méthode aux caractéristiques peut être appliquée, compte tenu de la gamme des nombres de Courant. L'utilisation d'un interpolateur bilinéaire plutôt que bicubique ne dégrade que faiblement les résultats (cf tables 9.6 et 9.7). UH2 se montre plus sensible que QH2.
   La dégradation est déjà plus nette à $\Delta t = 0.05$ mais **le mode de résolution de l'advection ne joue vraiment que pour** $\Delta t = 0.01$. En effet, si l'on applique interpolation bilinéaire ou méthode UPWIND, on obtient des résultats plus médiocres qu'avec le pas de temps précédent, 5 fois plus élévé.

4. La vitesse de convergence de UH2 est plus rapide que celle de QH2, sauf pour $\Delta t = 0.01$ s. En effet, la ligne d'eau se stabilise entre 40 et 45 s pour $\Delta t = 0.2$ et 0.1, entre 45 et 50 s pour $\Delta t = 0.05$ alors qu'elle survient toujours, pour QH2, entre 55 et 60s.

5. On notera enfin que les résultats de UH2 sont toujours trés proches de ceux de TELEMAC, modèle d'EDF aux éléments finis, qui travaille également en variables "hauteur-vitesse" et résoud l'étape d'advection par une méthode aux caractéristiques. La seule divergence notable concerne les simulations avec $\Delta t = 0.01$. Là les résultats de TELEMAC sont éloignés de ceux de la meilleure version de UH2 (interpolation bicubique) ce qui nous fait soupçonner que pour ce cas on a utilisé dans TELEMAC une interpolation bilinéaire, et non pas comme d'habitude l'interpolation bicubique d'Holly-Preissmann.

<u>Coûts de calcul</u> **Le coût de calcul, par noeud de la grille et par pas de temps, dépend essentiellement du mode de traitement de l'advection**, la résolution de l'étape de propagation consommant le même CPU quelque soit la version UH, QH ou DAN considérée (le gain apporté par l'une ou l'autre simplification est indétectable).

- Quand on applique une méthode aux caractéristiques, le coût de calcul individuel est également fonction du pas de temps. Plus celui-ci est grand, plus les trajectoires des particules sont longues et la remontée des caractéristiques coûteuse.

- Pour le test du canal ($C_r \leq 1$), le temps CPU consommé par pas de temps pour 1000 noeuds est de 0.26 s avec les caractéristiques, 0.10 s avec la méthode UPWIND. Le coût de l'étape de propagation est un peu inférieur à 0.09 s.
- On a un coût CPU par pas de temps similaire (0.25 s pour 1000 noeuds) pour le test du seuil aux deux pas de temps tels que C_r est plus faible que 1. L'adoption de la méthode UPWIND permet un gain considérable (temps réduit à 0.082s/1000 noeuds). Par contre, pour $\Delta t = 0.10$ et 0.20 s, les coûts montent à 0.32 et 0.39 s respectivement.

- L'augmentation des coûts individuels est cependant contrebalancée par la réduction du nombre de pas de temps, et donc du nombre de calculs, qu'il est nécessaire de pratiquer avant l'obtention d'une solution stable. Par exemple, pour $\Delta t = 0.01$, on résoud de l'ordre de 20 fois plus de calculs qu'avec $\Delta t = 0.2$, tandis que le coût de chaque calcul n'est réduit que de 36 % (méthodes aux caractéristiques) ou d'un facteur 5 (méthode UPWIND).

Conclusion des tests en permanents (section 9.4)

Les tests pratiqués nous permettent d'avancer les conclusions suivantes :

1. **La simplification des termes d'advection (cf (Dan N'Guyen, 1988)) est à éviter.** Elle était peut-être admissible pour des écoulements marins ou côtiers où le terme de propagation est archi-dominant. Dans les deux cas fluviaux considérées ci-dessus elle est à rejeter sans ambiguité possible.
2. **Les simplifications de l'étape de propagation dégradent les résultats.**
3. **la formulation "hauteur-débit" s'avère légèrement supérieure à la formulation "hauteur-vitesse"**, notamment en ce qui concerne la préservation des débits. Cette conclusion est sans doute celle des trois qui a le caractère le plus provisoire. En effet, nous avons travaillé sur des cas permanents, où l'erreur systématique (cf section 8.3.1.2) qui affecte la résolution de l'étape d'advection en formulation QH va en s'atténuant. Cette erreur, proportionnelle à l'incrément de hauteur durant un pas de temps, pourrait être plus grave dans un cas instationnaire. C'est pourquoi nos avons décidé de poursuivre nos tests pas uniquement avec la meilleure version à l'issue de ce chapitre, à savoir QH2, mais également avec son second UH2.

Chapter 10

Validation on benchmark tests : unsteady flows

Tests presented in this chapter are behavioral ones, as they do not possess any analytical solutions. They can be considered as being more demanding than tests of the previous chapter. They are rather far apart from riverine applications but who can do the more can do the least !

1. **The first test, about wave propagation in a closed basin, is a good check of the performance of the factorization method.** Indeed it deals with a quite unsteady, truly twodimensional flow, the free surface displaying quickly evolving sharp gradients. Besides, we shall see it highlights the difference between the "depth/unit discharge" and "depth/velocity" formulations and helps understanding why the latter version, less conservative, can be preferred.
2. **The second test, about a tidal basin, allows to check how the algorithms cope with a salient angle. It is also the first one where diffusion of momentum is not neglected ... and turns out to influence strongly flow patterns.** We also found, when working on this test, that there was only one consistent way of dealing with the open boundary conditions (details not given here).
3. **The last test**, relative to an expanding flume, is a bit peculiar. It **consists of the interpretation of a laboratory experiment.** Physics tells us that a the depth-averaged model without turbulence representation is unable to deal properly with this case ... which does not prevent attempts (Stelling & Wang, 1984; Dan N'Guyen, 1993). We nevertheless presented this test because :
   - **It precisely demonstrates the limits of the St-Venant model.**
   - **It highlights the influence of** parameters we had not the opportunity to discuss previously, namely **solid boundary conditions and the representation of the diffusion of**

momentum.

- **It also stresses the difficulty of achieving proper experimentation.** Indeed, however carefully planned the experience was, it turns out that the time scale of the observations (every 10s) is too large to allow one to fully understand how the separating flow develops.

We think thus that there is a lot to learn when working on that problem, both from a numerical and a physical viewpoint. Consequently, we judged it was worth a little advertising !

10.1 Surface wave propagation in a closed homogeneous basin

10.1.1 Presentation

The preceding tests are basically one-dimensional ones. We now need to check how the model behaves when the flow features are undoubtedly two-dimensional. A test suited to that purpose has been discussed both in (Dan N'Guyen, 1988) and (Galland & Hervouet, 1988). It deals with surface-wave propagation in a closed, square and homogeneous basin. The basin length is 21 m, its bottom is flat ($z_b = -2.4$m) and frictionless. The initial water surface is given by a gaussian distribution (cf figure 10.1) :

$$\zeta(x,y,0) = 2.4 \exp\left(-\frac{x^2+y^2}{4}\right) \tag{10.1}$$

where x and y denote the coordinates in a cartesian system whose origin is located at the basin center and whose directions are parallel to the basin sides. The initial velocity field is null.

This test is solved in (Galland & Hervouet, 1988) by TELEMAC, a finite element model based on a "depth/velocity" formulation, and in (Dan N'Guyen, 1988; Dan N'Guyen, 1993) by a finite difference model based on a "depth/unit discharge" formulation, whose factorization approach to the solution of the propagation step we have been following. As pointed out in (Dan N'Guyen, 1993), this is a difficult case to apply space splitting because surface slopes and curvature in each direction are large : the initial water surface achieves a maximum slope of -1.029 for $x^2+y^2 = 2$ and a maximal curvature of -1.2 at the basin centre.

There is no analytical solution available for this test case. However, the qualitative evolution of the flow can be determined. Given the basin features and the initial conditions, the wave motion is isotropic : as the gaussian hill collapses, we should observe first the propagation of a circular wave. Then, as soon as reflection on the walls occurs, radial symmetry will disappear but velocities and depths should nevertheless keep on displaying some symmetry with respect to both coordinate axes and to both diagonals of the basin. These are the features (besides mass conservation) that our models should reproduce.

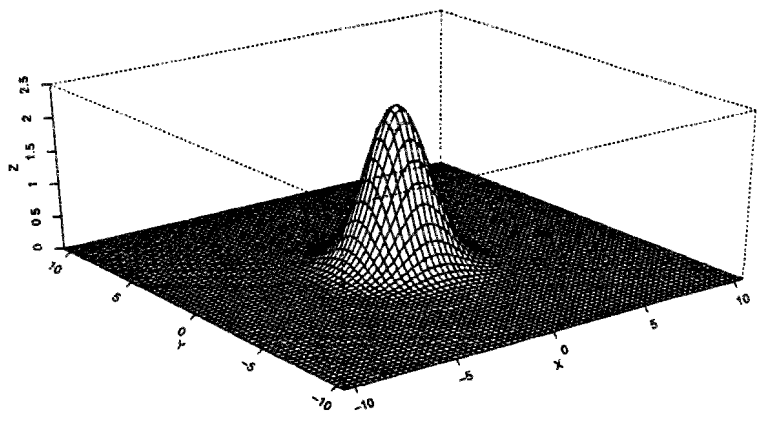


Figure 10.1: Initial free surface profile in closed homogeneous basin

The basin is discretized with space steps $\Delta x = \Delta y = 0.3$ m. Thus we have 5041 (71×71) computational grid nodes. TELEMAC (Galland & Hervouet, 1988) has been run on this test case with $\gamma = 0.6$, $\Delta t = 0.04$s and till $t = 4$s. Dan N'Guyen's model has been applied with implicitation parameter $\gamma = 0.5$ and different time steps, namely $\Delta t = 0.04$, 0.08 and 0.15 s : illustrations of its forecasts till $t = 1.2$s (which corresponds approximately at the time when reflection on the walls occurs) are supplied in (Dan N'Guyen, 1988; Dan N'Guyen, 1993). For $\Delta t = 0.04$s, the propagation Courant number at the beginning of the simulation ranges between 0.915 (near the walls) and 1.294 (at the basin centre).

In contrast to previous steady-state tests, the flow variability should induce a more important sensitivity of the model forecasts to the accuracy of the temporal discretization of the working equations : the influence of implicitation parameter γ and time step Δt on the simulation results should be more obvious. As for the previous tests, we let γ vary between 0.5 and 1. Models were run for time steps $\Delta t = 0.04$, 0.08 and 0.12 s.

Similarly, this test should also allow us to assess the relevance of iterative approaches to the solution of the propagation step, whether they include a correction to the factorization error or not. A look at the initial conditions may already help us to guess when accounting for the factorization error is advisable. This error is negligible when inequalities 8.42 are satisfied. At $t = 0$, the critical point is the basin centre where the curvature is maximum. There, 8.42 appears to be satisfied if and only if $(\gamma \Delta t)^2 \ll 0.04$. With $\Delta t = 0.04$, this is satisfied whatever value γ has. For $\Delta t = 0.08$, the magnitude of the factorization error exceeds 10 % of the other terms in the free surface equation if $\gamma > 0.8$. We observe the same problem for any $\gamma > 0.5$ with $\Delta t \geq 0.12$s.

On the basin walls, we applied the following treatment :

- The equation governing the free surface evolution is discretized applying a second order approximation to second-order derivatives, based on the nullity of free surface normal gradient (second option presented in 8.6.2).
- Normal velocities are set to zero.
- Tangential velocities are let free.

The results are surveyed each 0.4 s from $t = 0$ to 4 s. Besides, we have been examining in more detail some specific points of the basin, which are indicated on figure 10.2 : the centre P_0 of the basin and points located half-way between P_0 and the basin walls, on the coordinates axis and on the basin diagonals. There, flow variables have been stored at each time step.

The forecasts are to be appraised both from a physical and numerical point of view. The visualization of free surface profiles and velocities fields, as well as the intercomparison of flow

evolution at points $\{P_1, P_3, P_5, P_7\}$ on the one hand, points $\{P_2, P_4, P_6, P_8\}$ on the other hand, will enable us to check whether the applied models preserve the symmetries of the problem.

Besides, we have been using the following quantitative indicators :

- $\delta\mathcal{M}$ denotes the relative mass error (expressed in %). It is estimated at each time level.
- In section 8.5.3, we suggested a method allowing control of the behaviour of the propagation step algorithm. It consists of studying the difference between δz^{n+1}, the increment of free surface between time levels n and $n+1$ as estimated by the model, and $(\delta z)^*$ which, given the estimated $\vec{U}^{n+1}$ (or $\vec{Q}^{n+1}$), satisfies the continuity equation.
  Let ϕ denote this difference : $\phi = \mid (\delta z)^* - \delta z^{n+1} \mid$. We have been surveying the evolution of its maximum $\phi_{\max}$ and of its spatial average $\overline{\phi}$ throughout the simulations.

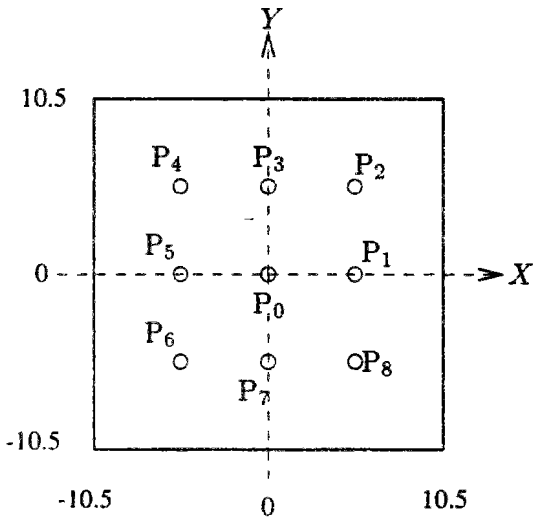


Figure 10.2: Control points in closed homogeneous basin

10.1.2 Discussion of results

10.1.2.1 Formulation "depth/velocity" (model UH)

We begin by examining the model behaviour for $\Delta t = 0.04$ s.

1. **Qualitative features of forecasted flows** Model UH appears to reproduce faithfully the expected symmetries of free surface profiles and velocity fields, whatever the chosen γ (cf figures G.1 to G.6, appendix G.3). It is perfectly devoid of spurious polarization often observed in classic ADI methods.

   Wall reflection appears to occur slightly before $t = 1.2$ s.

2. **Quality of the numerical resolution**

   A survey of $\overline{\phi}$ and $\phi_{\max}$ indicates that the propagation step is resolved fairly accurately.

   $\overline{\phi}$ remains respectively inferior to 0.4, 0.2, 0.1 mm for $\gamma = 0.5$, $\gamma = 0.6$, $\gamma = 0.8$ and 1. This is less than the precision threshold used in iterative versions.

   $\phi_{\max}$ stays always less than 6 mm for all γ values, which is negligible with respect to depths whose order of magnitude is one meter high. The largest values of $\phi_{\max}$ are systematically observed at the very beginning of the simulation, at the basin centre. Then they tend to stabilize at different levels according to γ (4, 1, 0.5 and 0.3 mm for $\gamma = 0.5$, 0.6, 0.8 and 1 respectively) except around $t = 2.4$s, for which $\phi_{\max}$ displays local maxima (of 6, 3, 1.2 and 0.6 mm for $\gamma = 0.5$, 0.6, 0.8, 1). In the second-half of the simulation, $\phi_{\max}$ is systematically reached on the basin walls, which indicates that our treatment of closed boundary conditions is not perfect.

3. **Mass conservation** is good : final $\delta\mathcal{M}$ equates respectively to -0.69, -0.39, -0.22, and -0.16 % for $\gamma = 0.5, 0.6, 0.8, 1$.

4. **Comparison of plain and iterative versions** Considering the good numerical behaviour of the "plain" model version, it is no wonder the iterative solution of the propagation step brings few improvements. It helps in reducing $\phi_{\max}$ in the first time steps (cf figure 10.3) but then the levels of mean and maximum errors raised by plain or iterative versions are strictly equivalent.

   The impact on mass conservation is negligible (an improvement of 0.01 % whatever γ).

   Sub-iteration slightly modifies the forecasted surface profiles and velocity fields but these modifications are minor (an absolute difference of approximately 0.5 mm as regards depths, a relative variation of about 0.5 % as regards velocities).

   Outcomes from iterative versions (1) and (2) (cf section 8.5.1) are the same but the first version converges faster than the latter. Indeed, version (1) converges after 2 iterations except for $t \leq 0.4$ s, when it requires 3 loops, while with version (2) 3 loops are always performed (and even 4 loops for $t \leq 0.24$). This is undoubtedly due to the fact that the linearization applied to initialize the sub-iterative process is cruder in version (2) than in version (1).

   Computational costs are of course more important for iterative versions. While the "plain" version consumes 0.30 s per 1000 points and per time step, iterative version (1) costs 0.38 s per 1000 points and per time step : this amounts nearly to a 30 % increase.

   In conclusion, there is little point in applying a sub-iterative resolution of the propagation step in this specific test case.

5. **Influence of the choice of γ**

   The sensitivity of model forecasts to γ is obvious (cf figure 10.4 and also figures G.7 and

G.8, app. G.3). Surface profiles and velocity fields at a given time display similar patterns whatever γ, but, as γ is closest to 1, the forecast variations of water levels and velocities are smoother. The straightforward explanation is that, as γ is farther apart from the optimal value of 0.5, the error term linked to the temporal discretization of the flow equations grows bigger : this results in larger errors concerning both celerity and amplitude of wave propagation, which are somewhat dampened.

(nb : The smoother the free surface and velocity profiles are, the smaller any error linked to the finite difference approximation of the related flow equations is. This probably explains why numerical indicators of the resolution quality are better for larger γ values (cf point 2 above).)

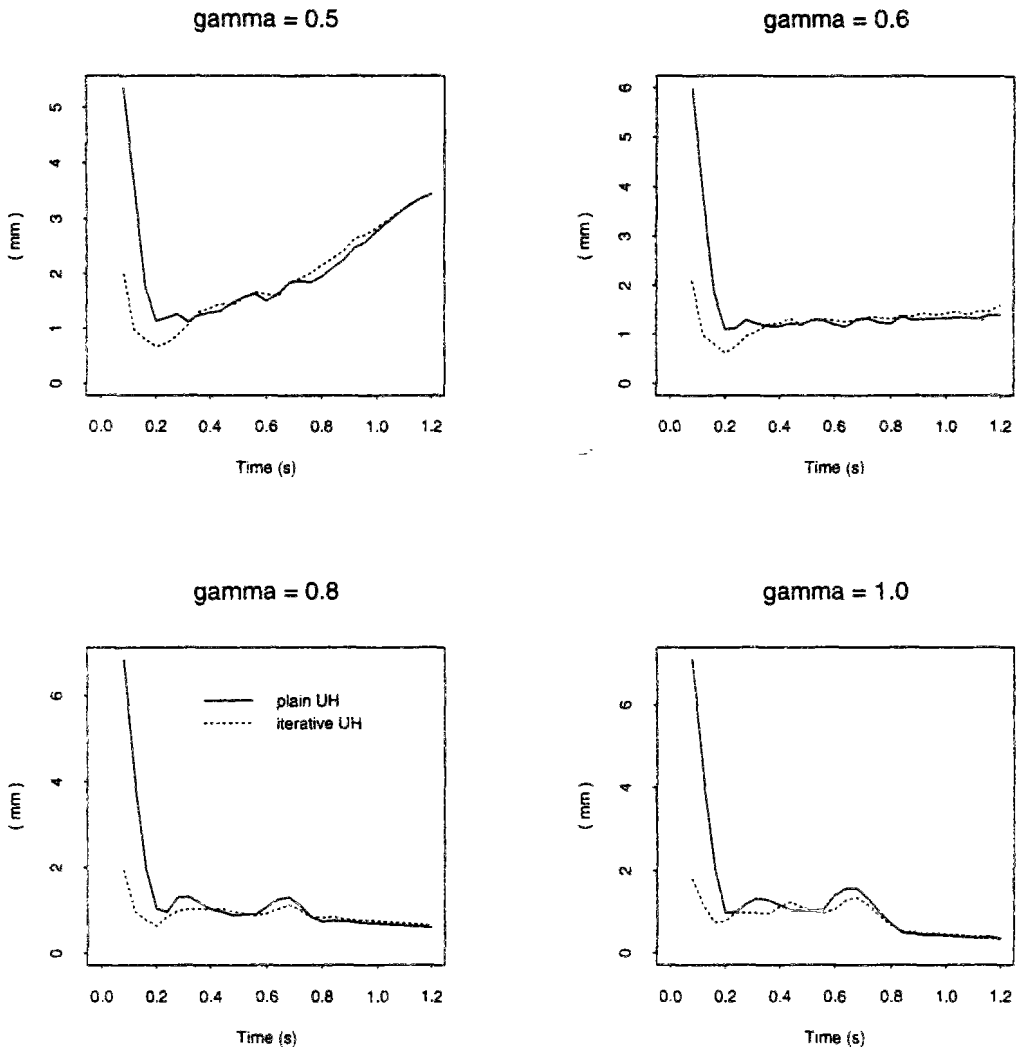


Figure 10.3: Comparison of $\phi_{\max}$ ($\Delta t = 0.04$s) for iterative and plain versions of model UH

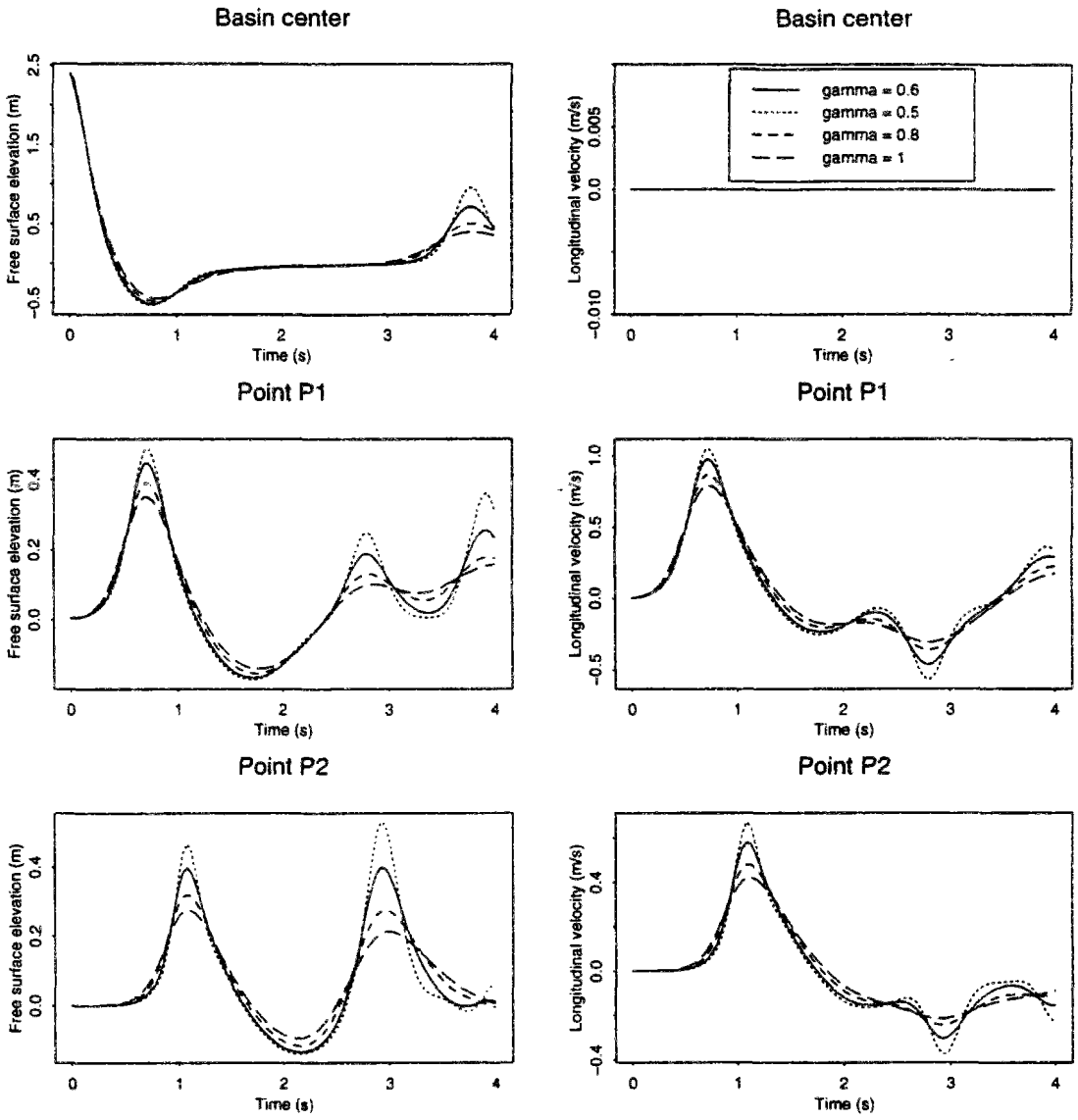


Figure 10.4: Influence of the choice of γ on forecasts at control points ($\Delta t = 0.04$)

10.1.2.2 Formulation "depth/discharge" (model QH)

Once again, we discuss the results obtained with $\Delta t = 0.04$ s.

The forecasts of model QH are consistent with the physical nature of the studied problem, all properties of symmetry being well preserved.
Mass conservation is excellent, with a final loss inferior to 0.001 %, whatever the value of γ !
The indicators of good numerical behavior ($\overline{\phi}$ and $\phi_{\max}$) are of the same order of magnitude (slightly lower) than with model UH. Similarly, iterating the resolution of the propagation step brings minor improvement of the resolution and negligible changes in the forecasts.
Increasing the value of implicitation parameter γ has the same effect as observed with UH.

We mentioned earlier (cf section 8.3.1.2) that the method we apply to compute advected discharges introduces a systematic error which amounts to $\left(h^{n+1} - h^n\right).\vec{U}^n$. As the phenomenon we are studying is quite unsteady (local variation of water levels within a time step can reach several tens of centimeters) it seems advisable to control the magnitude of this error and its relative importance with respect to the estimates of advected discharges. It turns out that, according to time and to the value of γ, the error represents between 1 and 2 % of the estimated advected discharges on average. However, it may locally (particularly at the beginning of the simulations) reach a tenth of the estimates. This led us to conclude that we could not discard a priori the hypothesis that the advection step error has some influence on the model forecasts. On the other hand, as this error appeared to be generally moderate, it seemed feasible to try and correct it rather straightforwardly, in an explicit manner which induces but very slight modifications of our algorithms. The new version of model QH, which allows for an iterative correction of the error in the computation of advected discharges, is presented next page.

The trials performed demonstrate that indeed the solution of the advection step influences strongly the model results. While "individual" errors may seem small, even negligible, their addition, both in time and in space, leads to important differences. In fact, it appears that there exist major discrepancies between forecasts of the "plain" and "corrected" QH model, the latter yielding results fairly close to those produced by the UH model (while ensuring a better mass conservation). This is illustrated first on figure 10.5, which displays the computed water levels and velocities at three control points (see also figures G.9 to G.11, appendix G.3.) Secondly, we have been plotting on figure 10.6, for different γ tunings, the evolution of indicators *mean(|DH|/H)* and *mean(|DU|/U)*. These denote respectively, at a given time, the spatial average - based on all computational grid nodes - of the relative differences between water profiles and velocity fields (in the first case, the absolute difference is normalized by the local water-depth; in the second case, it is normalized by the local velocity modulus).

"Plain" model QH (without sub-iteration) and "corrected" iterative version consume respectively 0.3 s and 0.41 s per 1000 nodes per time step.

Iterative "depth/unit discharge" formulation - Version 2

0 Initialisation :

$\delta z^{[0]} = 0$, $h^{[0]} = h^n$, $\zeta^{[0]} = \zeta^n$, $\vec{Q}^{[0]} = \vec{Q}^n$, $f_b^{[0]} = f_b^n$, $\vec{Q}_{adv}^{[0]} = \vec{Q}^{n+2/3}$

where $\vec{Q}^{n+2/3}$ denotes the outcome of the advection step,

$\vec{Q}_{adv}^{[q]}$ the advected discharges estimated after iteration $[q]$.

1 Equation satisfied by the free surface increment :

$$\delta z^{[q+1]} - g\,(\gamma\,\Delta t)^2\ \mathrm{div}\left(\frac{h^{[q]}}{1+\gamma\,f_b^{[q]}\,\Delta t}\,\nabla\,\delta z^{[q+1]}\right) = -\,\Delta t\ \mathrm{div}\,\frac{\vec{B}^{[q+1]}}{1+\gamma\,f_b^{[q]}\,\Delta t} + \mathcal{G}^{[q+1]}$$

with

$$\vec{B}^{[q+1]} = (1-\gamma)\,\vec{Q}^n + \gamma\,\vec{Q}_{adv}^{[q]} + \gamma\,(1-\gamma)\,\Delta t\,\left(f_b^{[q]} - f_b^n\right)\,\vec{Q}^n - g\,(\gamma\Delta t)\,\left(h^n + \gamma\delta z^{[q]}\right)\,\nabla\zeta^n \tag{10.2}$$

and $\mathcal{G}^{[q+1]}$ factorization error corrective term, as indicated in 8.5.2.

2 Updating of the advected discharge

Once $\delta z^{[q+1]}$ has been computed :

$$\vec{Q}_{adv}^{[q+1]} = \vec{Q}^{n+2/3} + \delta z^{[q+1]}.\vec{U}^n \tag{10.3}$$

3 Updating of the discharge

$$\begin{aligned}\left(1 + \gamma\,f_b^{[q]}\,\Delta t\right)\,\vec{Q}^{[q+1]} &= \vec{Q}_{adv}^{[q+1]} - (1-\gamma)\,\Delta t\,\left(f_b\,\vec{Q}\right)^n \\ &- g\,(1-\gamma)\,\Delta t\,(h\nabla\zeta)^n - g\,\gamma\,\Delta t\,(h\nabla\zeta)^{[q+1]}\end{aligned} \tag{10.4}$$

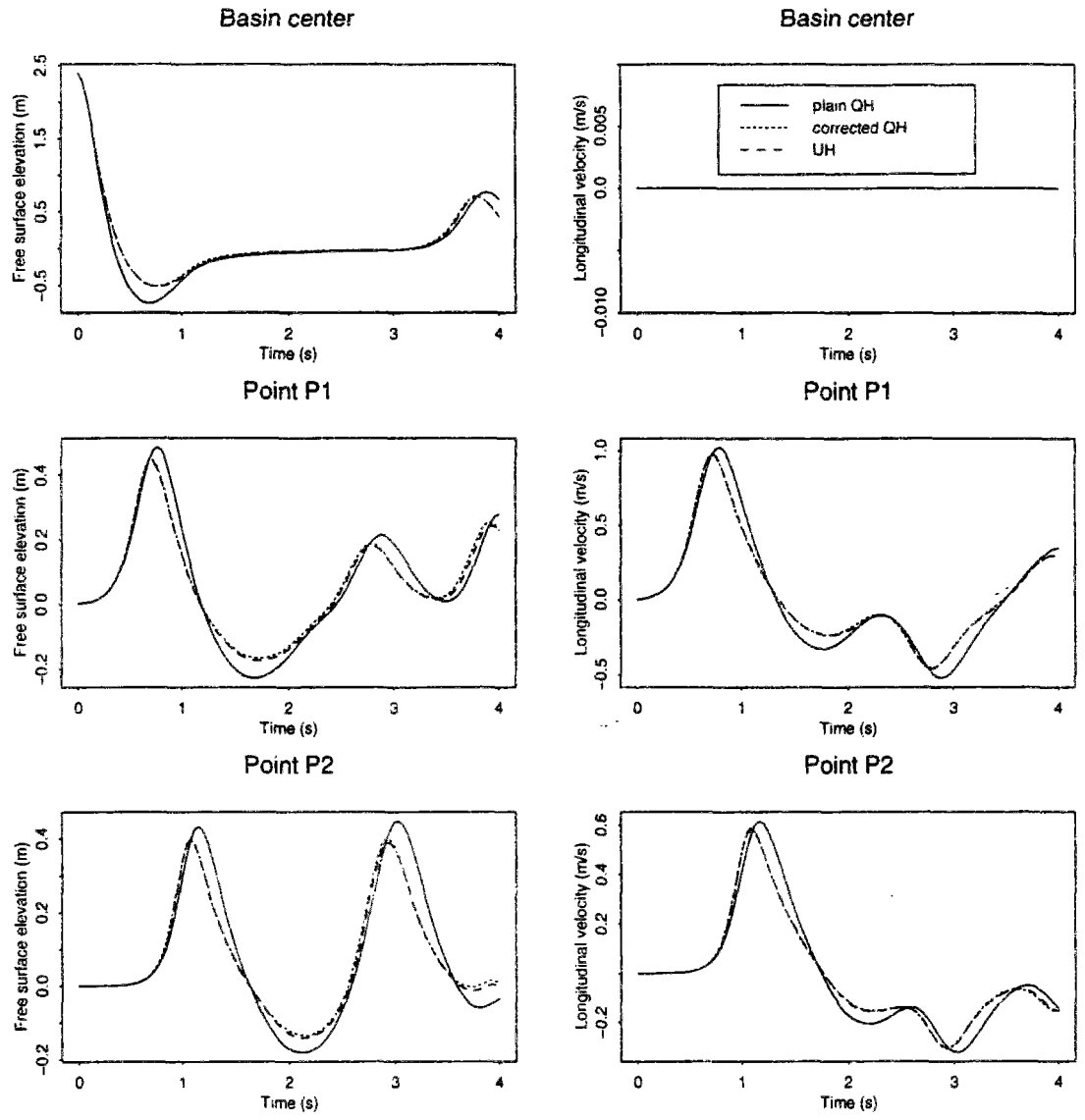


Figure 10.5: Forecasts according to "plain", "corrected" QH and UH models ($\Delta t = 0.04$, $\gamma = 0.6$)

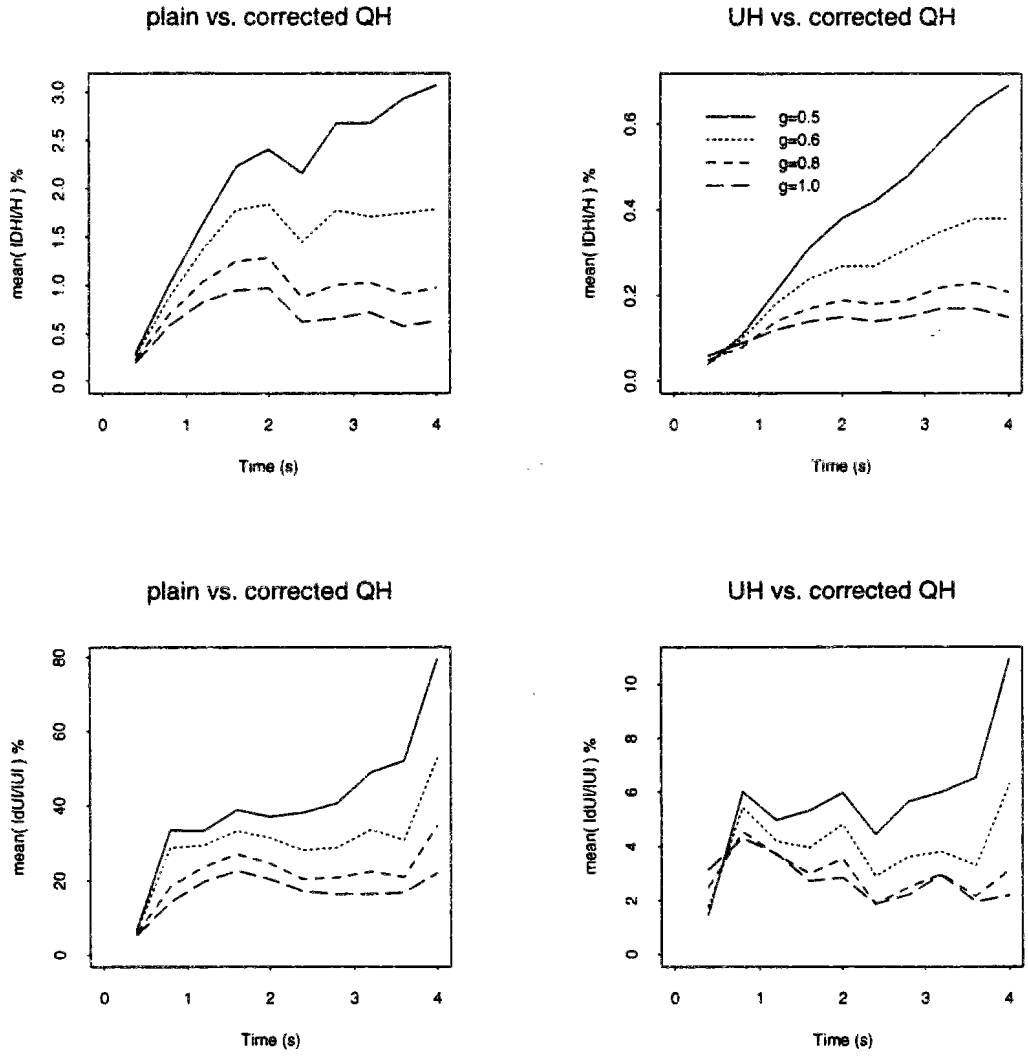


Figure 10.6: Relative differences between water profiles and velocity fields computed according to various models ($\Delta t = 0.04$)

10.1.2.3 Influence of time step variations

This has been investigated for model UH only.

Both the plain version of model UH and its iterative version (1) have been applied with $\Delta t = 0.08$ s, their respective CPU costs being 0.35 s and 0.46 s per 1000 nodes and per time step.

The improvement brought by the iteration of the propagation step resolution is this time more significant, as illustrated on figure 10.7. In order to achieve this improvement when $\gamma = 1$, it is necessary to include the correction of the factorization error term; this is not the case for lower γ values.

Similarly, the differences between forecasts yielded by plain and iterative versions are larger than with $\Delta t = 0.04$ s, notably as regards velocity fields. While the average difference *mean(|DU|/U)* ranges between 1 and 3 %, local relative differences may reach ten percent, even more for $\gamma = 0.5$. The absolute difference between computed surface profiles remains moderate : its spatial average ranges between 1 and 3 mm, its maxima never exceed 3 cm for $\gamma = 0.5$, 1 cm for larger γ values.

With the plain version, final $\delta\mathcal{M}$ are respectively -0.65, -0.26, -0.12, -0.09 % for $\gamma = 0.5, 0.6, 0.8, 1$. Iteration reduces the mass loss (of 0.02 %).

In fact, the main advantage of the iterative version is that its forecasts appear to be slightly less sensitive to the increase of the time step than when no iteration is applied. However, the dependency with respect to Δt does exist !

- Surface profiles forecast with $\Delta t = 0.08$s bear undoubtedly similarity with those forecast with $\Delta t = 0.04$s. They notably preserve too the symmetries of the studied phenomenon. However, they display generally a smoother shape, as was observed when increasing γ. This is already noticeable for the most favorable γ value (i.e. which displays the lowest sensitivity to Δt increase), namely $\gamma = 0.6$ (cf figures 10.8 and 10.9).
- The temporal dependency is drastically more important when $\gamma = 0.5$, which leads to suspect that, for this γ value, when Δt is set to 0.08 s, the model is on the verge of instability (cf figure 10.10).
- The estimation of velocities appear to be more influenced by the time step increase than is the computation of water level variation (cf figure 10.11).

All these trends are observed to a greater extent when Δt is further increased to 0.12 s. While the model keeps on preserving the flow symmetries, its forecasts are still farther apart from

results yielded with $\Delta t = 0.04$ s, as can be observed in figure 10.12 and fig. G.12 to G.14 (appendix G.3).

As expected, an increase in Δt induces worse (i.e. larger) values of quality indicators $\overline{\phi}$ and $\phi_{\max}$. However, the error ratio is less than the time steps ratio.
Similarly, mass conservation *per time step* is worst : for $\gamma = 0.6$ for instance, the relative mass loss per time step is respectively 0.0038 %, 0.0048 % and 0.0054 % for $\Delta t = 0.04$, 0.08 and 0.12 s.

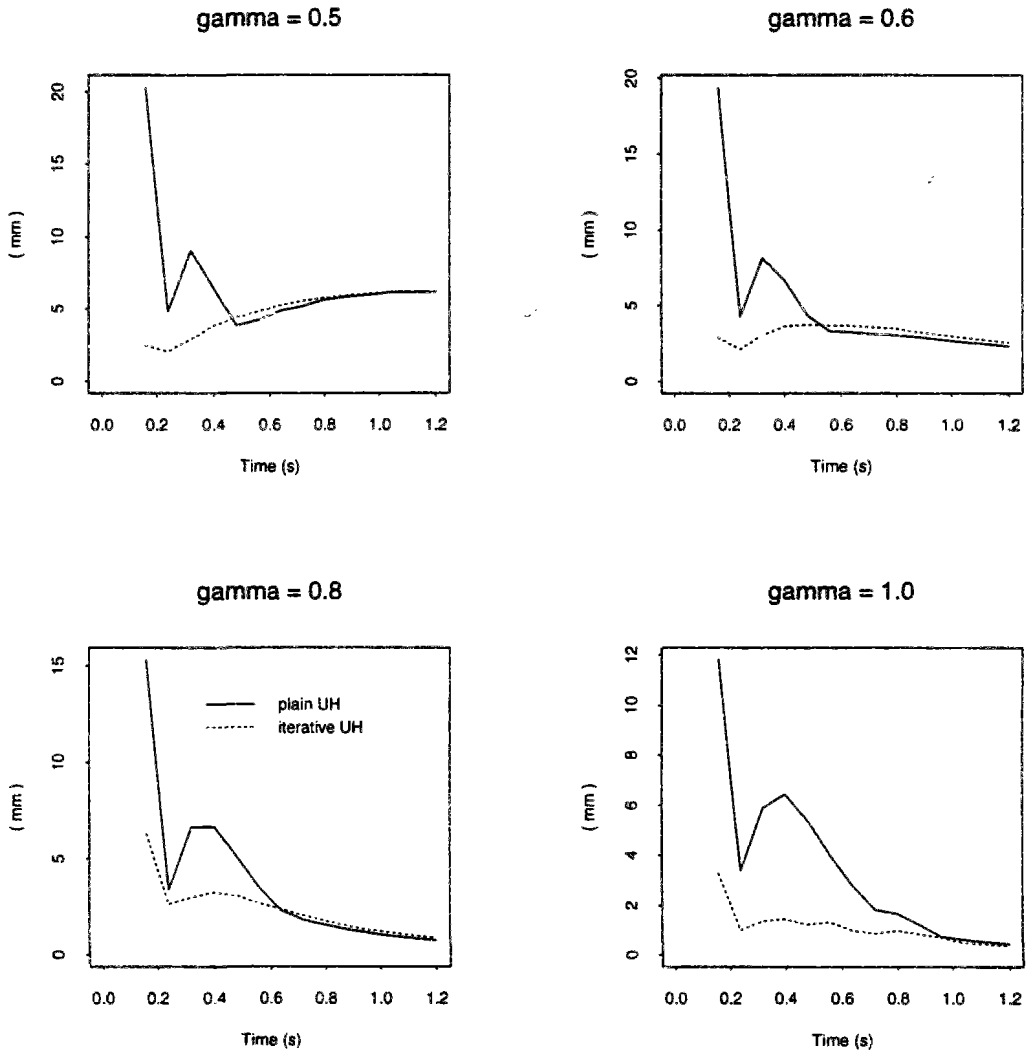


Figure 10.7: Comparison of $\phi_{\max}$ ($\Delta t = 0.08$s) for iterative and plain versions of model UH

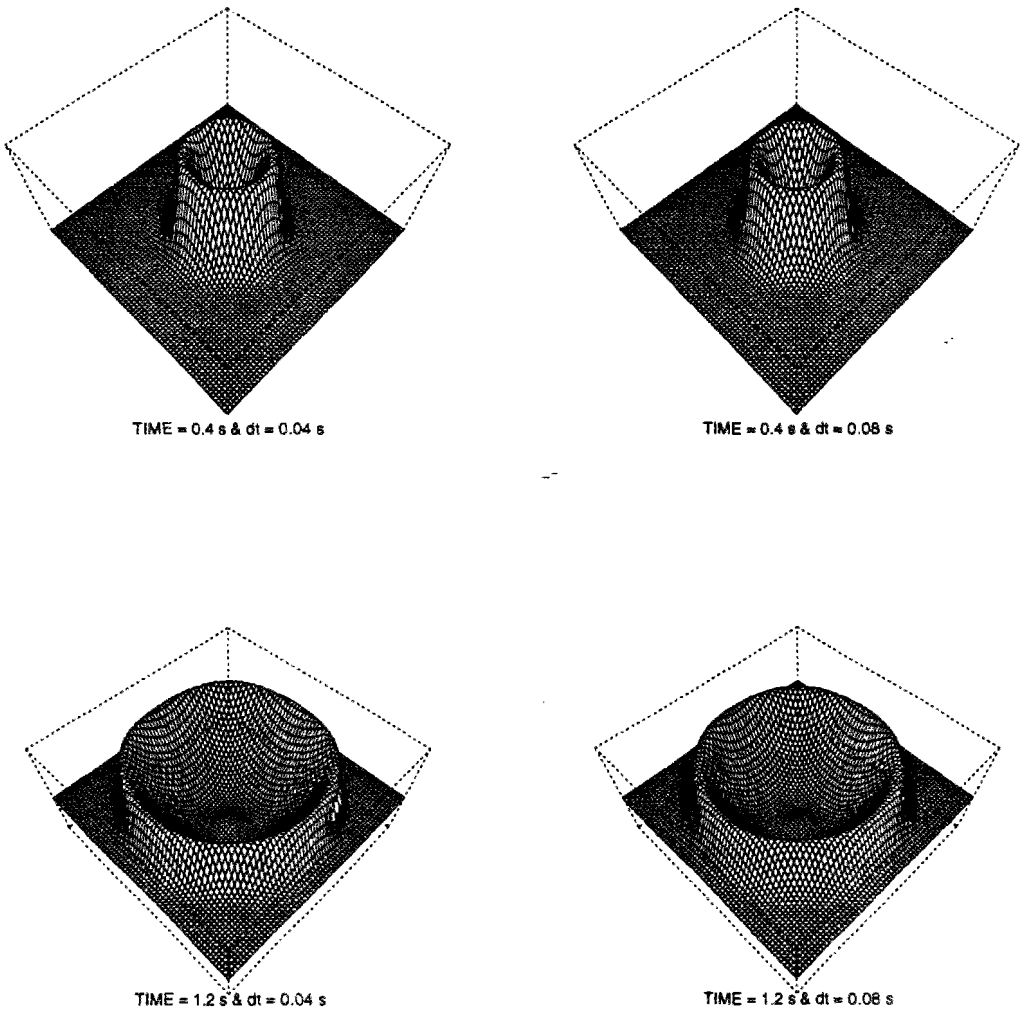


Figure 10.8: Influence of Δt increase on forecast water surface ($\gamma = 0.6$) : $t = 0.4$ & 1.2s

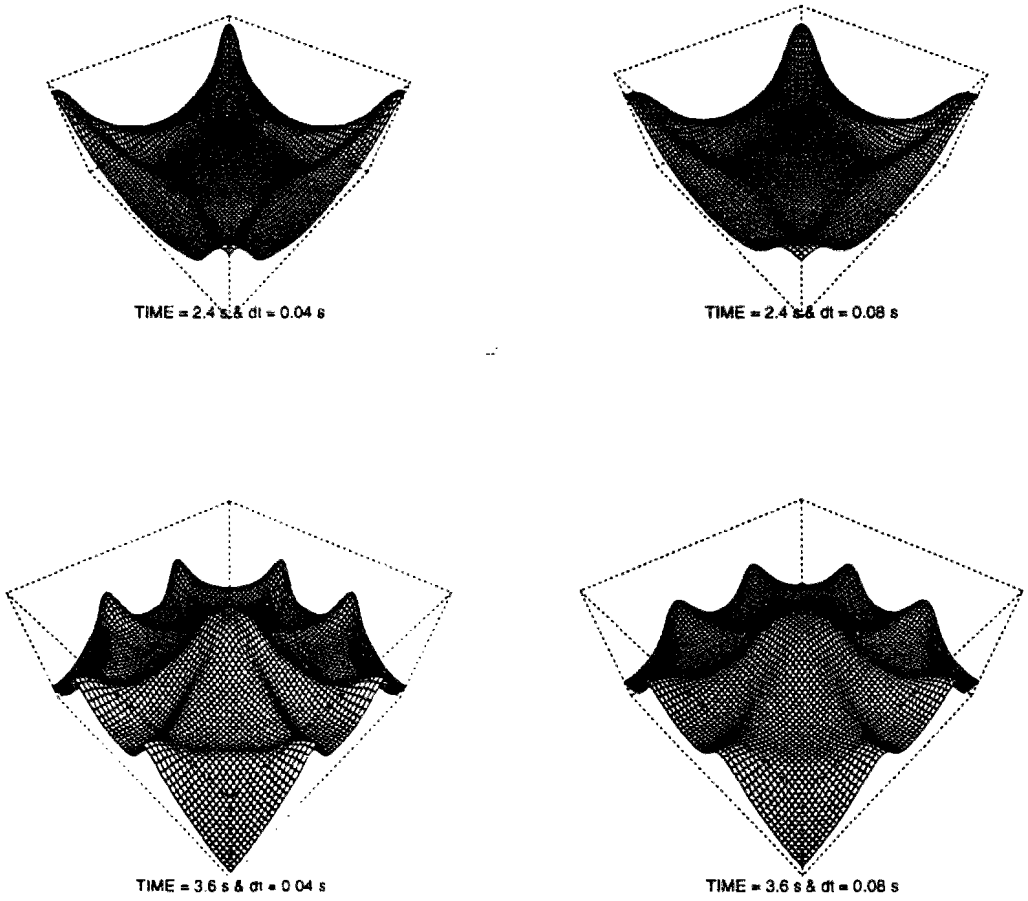


Figure 10.9: Influence of Δt increase on forecast water surface ($\gamma = 0.6$) : $t = 2.4$ & 3.6s

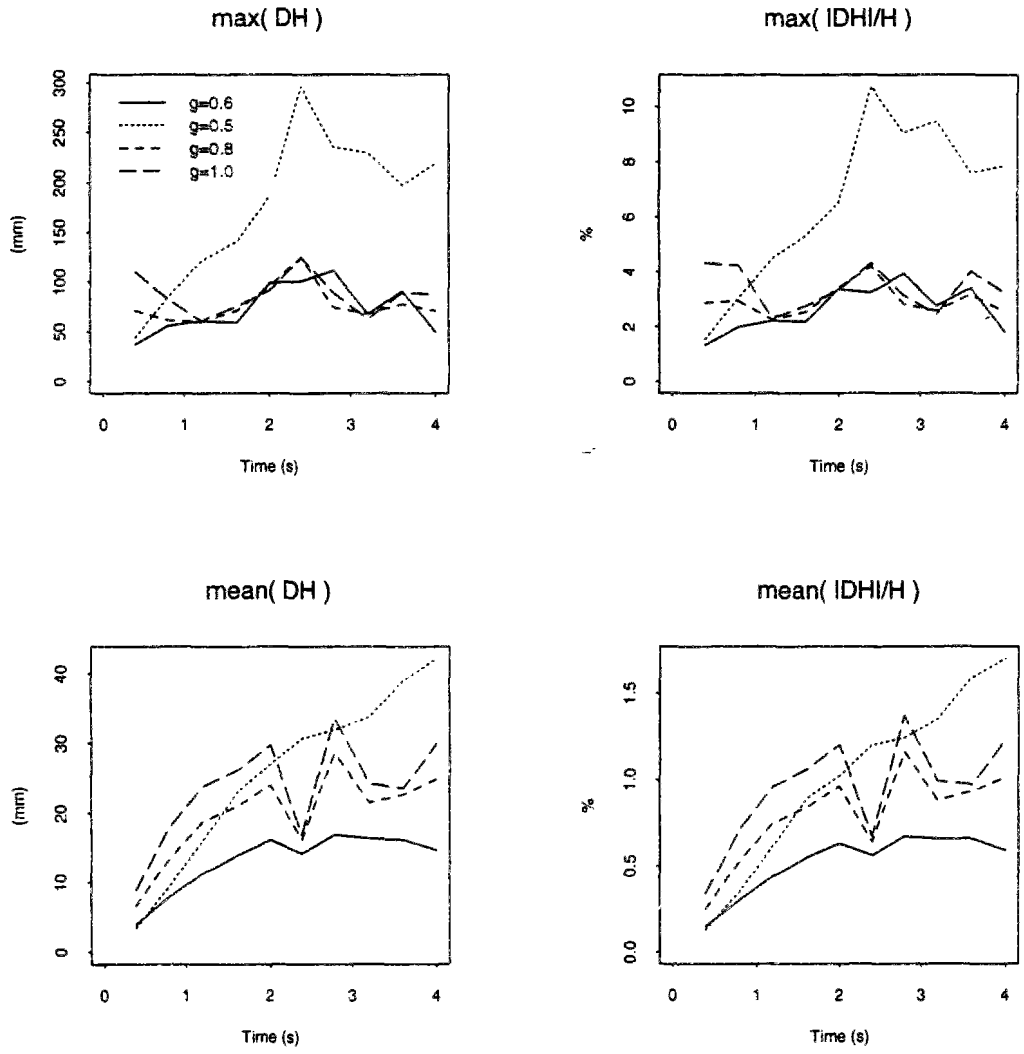


Figure 10.10: Sensitivity of surface profile estimates to Δt increase

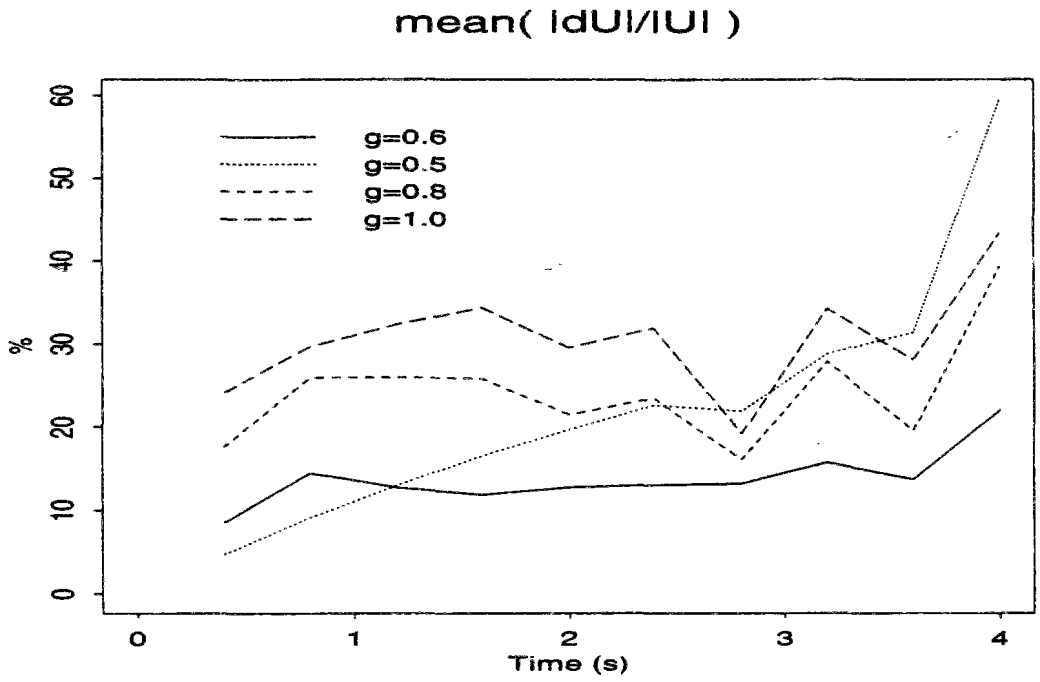


Figure 10.11: Sensitivity of velocity estimates to Δt increase

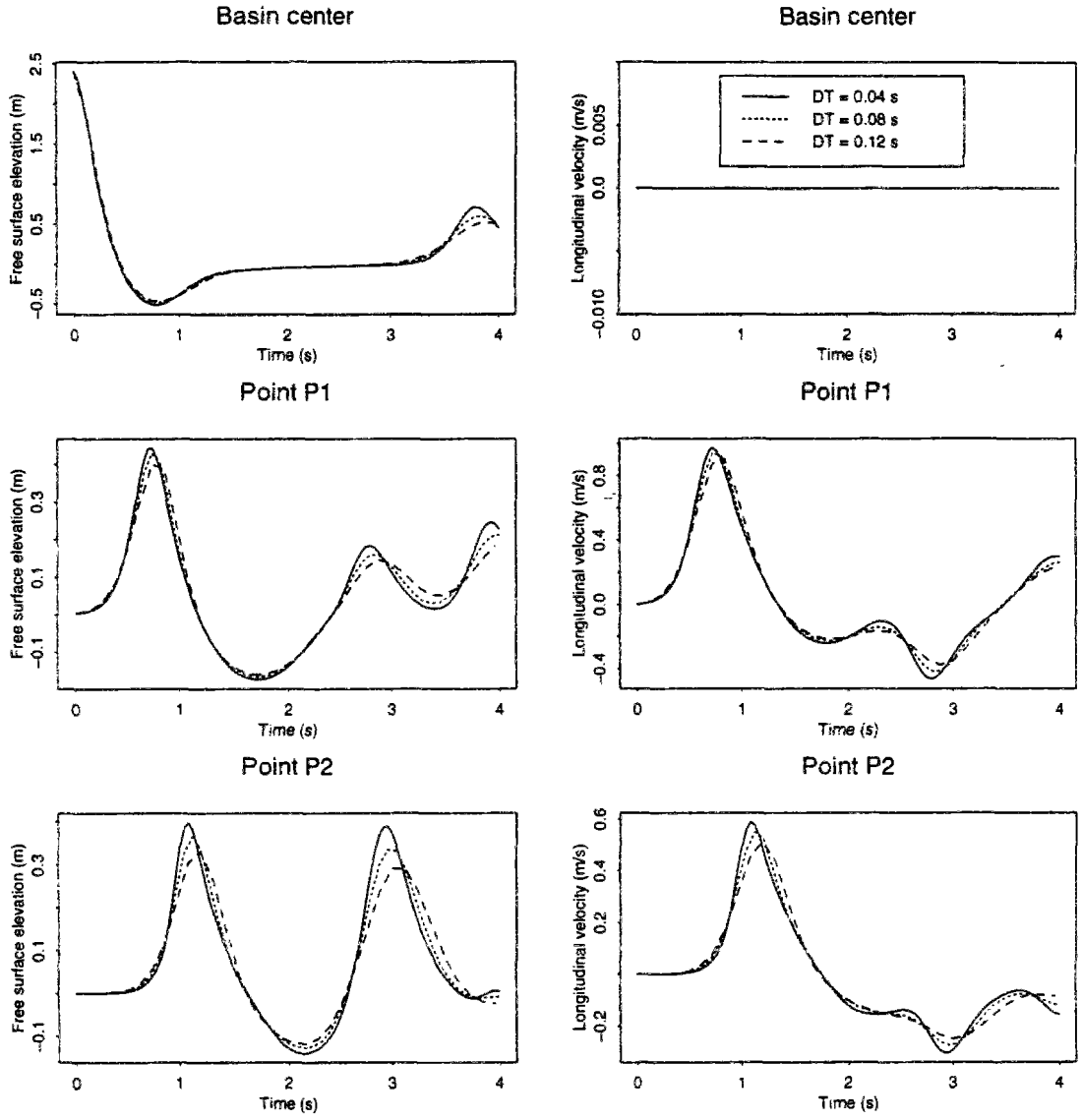


Figure 10.12: Forecasts at control points for different Δt

10.1.3 Conclusions

1. This test demonstrates the good performance of the factorization method applied in the propagation step to solve the two-dimensional equation governing the evolution of water levels. In spite of the strong irregularities and rapid variation of the surface profiles, the models produce results which remain consistent from a physical point of view, for all tested time steps and implicitation parameters.

2. The test also illustrates, in a very unsteady situation, the sensitivity of "depth/unit discharge" models to the correction of the systematic error embedded in the resolution of the advection step. The version accounting for the correction of this error and the plain, uncorrected, version raise respectively forecasts very close to and far apart from those yielded by the model based on a "depth/velocity" formulation.

3. Finally, since forecasts appear to be quite dependent on the temporal discretization and implicitation, this test underlines the need for "tuning" models with care whenever the studied flow is rapidly varying. We consider that this recommendation applies in fact to any model relying on a first-order (in time) approximation to the flow equations.

10.2 Tide in a rectangular harbour

10.2.1 Presentation

This test, introduced in (Galland & Hervouet, 1988), is once again a behavioural test. It deals with the study of a flat, rectangular harbour which is submitted to the influence of a tidal, sinusoidal, flow through a side feeder canal. What can be checked is mass conservation, and the general look of forecast flow patterns, notably whether they are disturbed at the vicinity of the junction of the feeder canal and the harbour. Indeed, salient angles are always delicate to deal with in numerical codes.

The open boundary conditions concern the flow velocities (cf table 10.1). There, as in the tests of sections 9.2 and 9.3, an open boundary condition for the free surface elevation is obtained by developing and discretizing semi-implicitly the momentum conservation equation : the resulting condition concerns the free surface gradients at the open boundary. (nb : Trials with an other kind of condition frequently used in tidal applications, namely the radiation condition we describe in next section, resulted in more stringent stability conditions - $\Delta t \leq 2$ s - and the development of undesirable velocity oscillations in the feeder canal)
At closed boundary nodes, the zero normal gradient condition on the free surface elevation is substituted for the general propagation step equation governing the free surface evolution : the

gradient is approximated with a first-order finite difference involving thus only the boundary node and its closest neighbour in the direction normal to the boundary. Other methods for dealing numerically with this zero normal gradient condition were found to induce spurious oscillations of the water surface in the narrow (5 mesh sizes) feeder canal.

Table 10.1: Conditions of harbour test

Harbour dimensions	864 m × 480 m
Feeder canal dimensions	160 m × 96 m
Mesh size	$\Delta x = \Delta y = 32$m
Initial conditions	$u = v = 0$, $h = 10$ m
Open boundary conditions	$v = -0.2 \sin \frac{2\pi t}{T}$, $u = 0$ $T = 12$ h 24 min $= 44640$ s
Closed boundary conditions	zero normal velocity perfect slip of tangential velocity
Diffusion of momentum	constant, isotropic diffusivity $D_{xx} = D_{yy} = 0.1$ m^2.s^{-1}
Bottom friction	Strickler coefficient $K_s = 60$
Initial simulation time Final simulation time	$t = 0$ s $t = 44640$ s (a full tide)

In (Galland & Hervouet, 1988), the model TELEMAC is applied with time step $\Delta t = 30$ s. However, for this time step, both "depth/discharge" (QH) and "depth/velocity" (UH) models turn out to be unstable. Consequently, we halved the time step to $\Delta t = 15$ s. With time steps of this order of magnitude, it is possible to solve explicitly the diffusion step, with a priori minor errors with respect to a fully implicit treatment of this operator. The advection step was solved with the backward characteristic method coupled with a bilinear interpolator. For the propagation step we fixed first the implicitation parameter as suggested in (Galland & Hervouet, 1988), namely $\gamma = 0.6$.

Flow forecasts on the whole computational domain were stored every hour. In addition, depth and velocity were recorded every time step at a point located in the middle of the feeder canal. Finally, mass evolution is surveyed every ten time steps.

10.2.2 Results

1. **Comparison of different model versions** (Δt = 15 s, γ = 0.6)
   Model UH was applied first. A look at the velocity fields confirm that they are quite consistent which what can be expected from a physical point of view (cf figures 10.13 to 10.15, and G.15 to G.17 in appendix G.4). A tide-induced circulation does appear in the basin. The salient angle has minor influences on the velocity field. It only induces a slight perturbation along the wall of the feeder canal as the basin is emptying (fig. G.17 and 10.15). The mass conservation after a full tide is good but not perfect : a mass loss of 0.82 % is observed. The tidal range is important (21.1 m). Consequently, propagation Courant numbers achieve large values during the simulation (a maximum of 9.6 at the tide reversal).

   Solving iteratively the propagation step modifies but slightly the model outcomes. Velocity forecasts throughout the simulation differ from less than 0.3 % on average, less than 3% at maximum discrepancies points. Water depth forecasts differ by less than 0.07 % at the utmost. These differences are indistinguishable on visualization of velocity fields, limnigraphs or hydrographs.
   Convergence is fast (obtained after two sub-iterations). The removal of the factorization error term brings a small improvement as regards mass preservation : final mass loss reduces to 0.74 % .

   When model QH is applied without correction of the advection step error, it quickly diverges. On-line survey reveals indeed that this advection step error becomes quickly overwhelming as the tide gains in strength. When this is corrected iteratively, as explained in the preceding section, model QH becomes stable, and gives forecasts nearly indistinguishable from those of model UH. A further improvement of mass preservation is observed (final mass loss of 0.69 %).

2. **Influence of modifications in numerical parameters of the simulation**
   (nb : this was tested mainly with model UH / iterative version)

   - The choice of implicitation parameter γ has a moderate influence over the forecasts. First, the indicators relative to the quality of the numerical resolution slightly worsen. Hydrographs and limmnigraphs computed with γ = 1 are slightly shifted with respect to those obtained with γ = 0.6. Besides, the predicted velocities are slightly lower (between 2 and 4 % on average) and the amplitude of the tide is slightly reduced (by 2 cm). The phase shift results in a worse mass conservation : with γ = 1, the basin has not finished emptying at the end of the tidal period; the excess mass amounts to 1.65 % (1.30 % with the QH model).
   - Treating the diffusion operator explicitly or dealing with it implicitly, with optimal

implicitation parameter 0.5, has no influence on model results.

- The maximum allowable time step which ensures the model stability appears to be $\Delta t = 20$ s. However, for this time step, the model accuracy is worst (continuity errors approximately double). As the tide rises, the velocities forecast with $\Delta t = 20$ s are somewhat larger, with the consequence that the tide maximum is achieved faster, and is higher (by 40 cm). Symmetrically, as the tide reverses, the velocities within the basin are decreasing faster than with the smaller time step, so that the basin fails to empty at the end of the tide (excess mass of about 2.53 % for $\gamma = 0.6$).
- Finally, while the advective Courant numbers are small (0.1 at the utmost), the method chosen to solve the advection step has some impact on the forecasts. When advection is solved with the upwind method, forecast velocities and depths behave as observed when increasing the time step. The tidal range is higher (by 70 cm) and final mass excess ranges between 2.6 and 4 % according to γ value.

3. **Computational costs**

   For $\Delta t = 15$s, the computational cost of model UH is 0.242 s per 1000 nodes and per time step. Approximately 64 % are devoted to the solution of the advection step, 30 % to the solution of the propagation step.

   When sub-iteration is applied, the cost rises by approximately 40 % (0.340 s per 1000 pts, and time step for both iterative versions, of UH and QH). Since the differences between the forecasts of the plain and iterative versions are small, it seems it is not computationally efficient to apply sub-iteration in this specific test case.

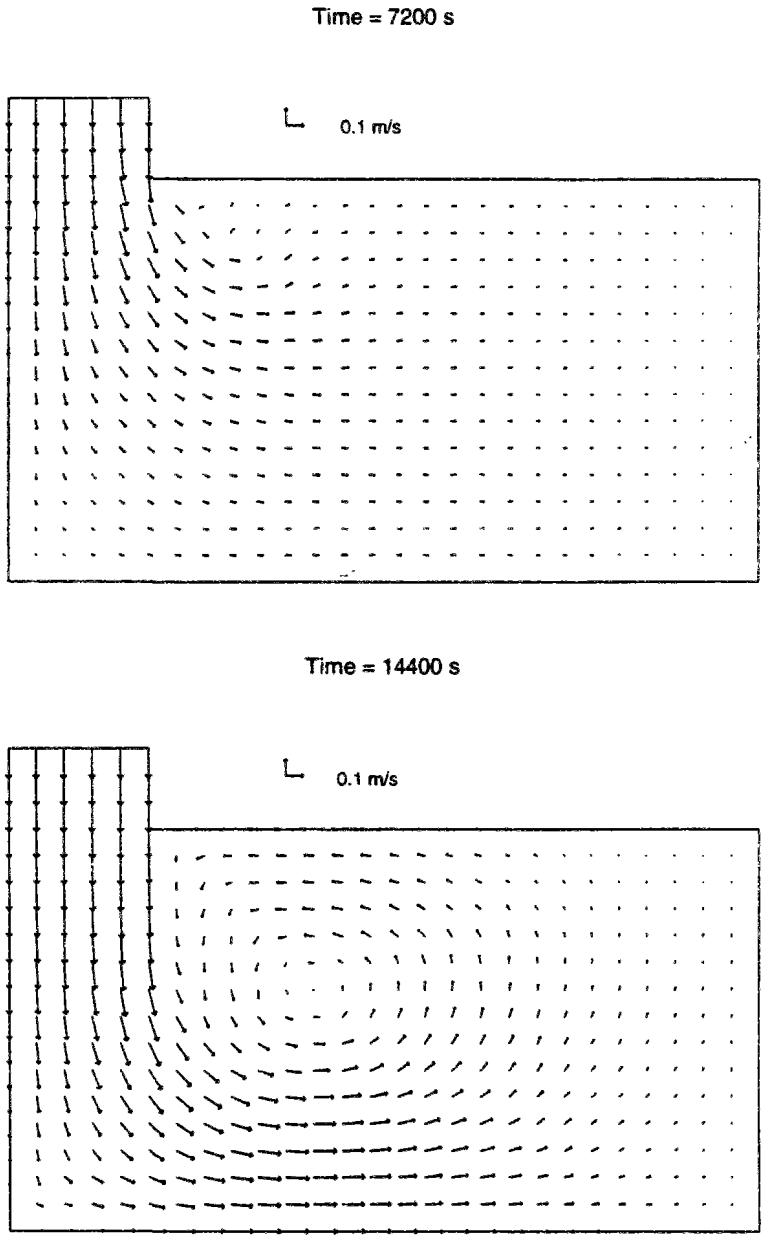


Figure 10.13: Velocity field in the harbour after $T = 2$ and 4 h

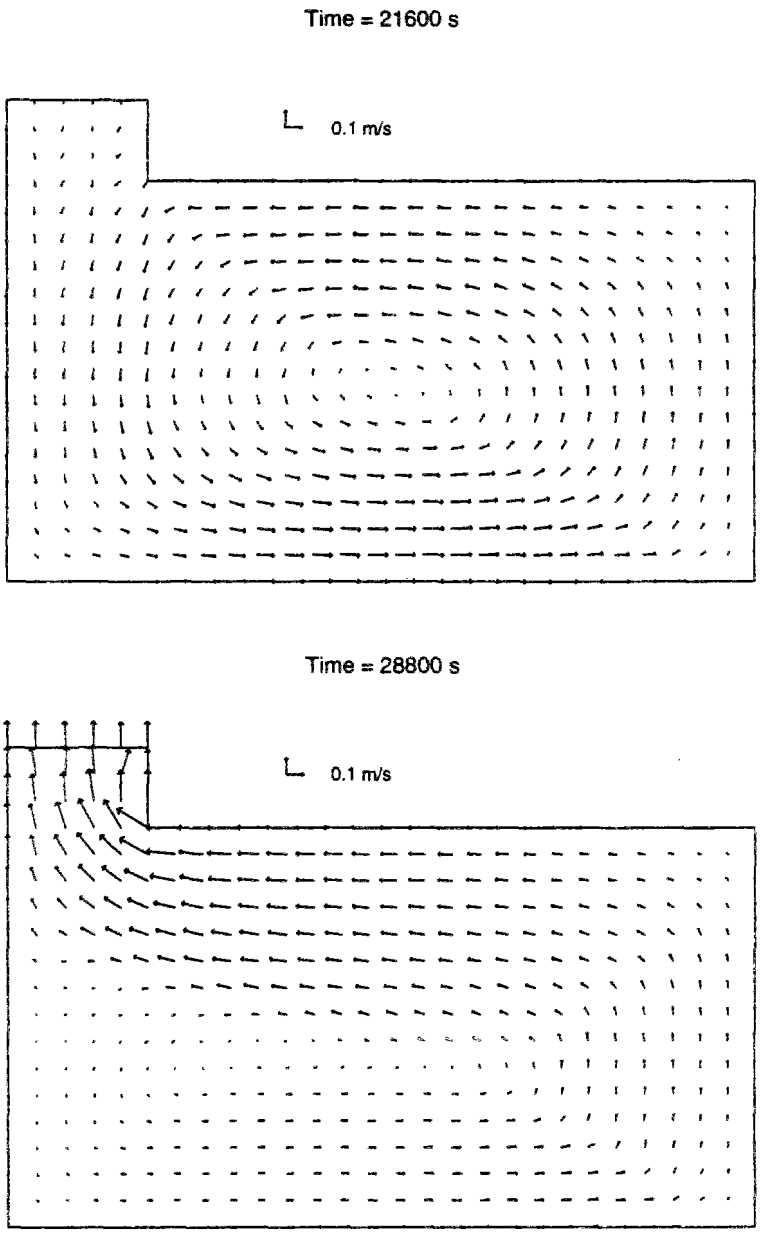


Figure 10.14: Velocity field in the harbour after $T = 6$ and 8 h

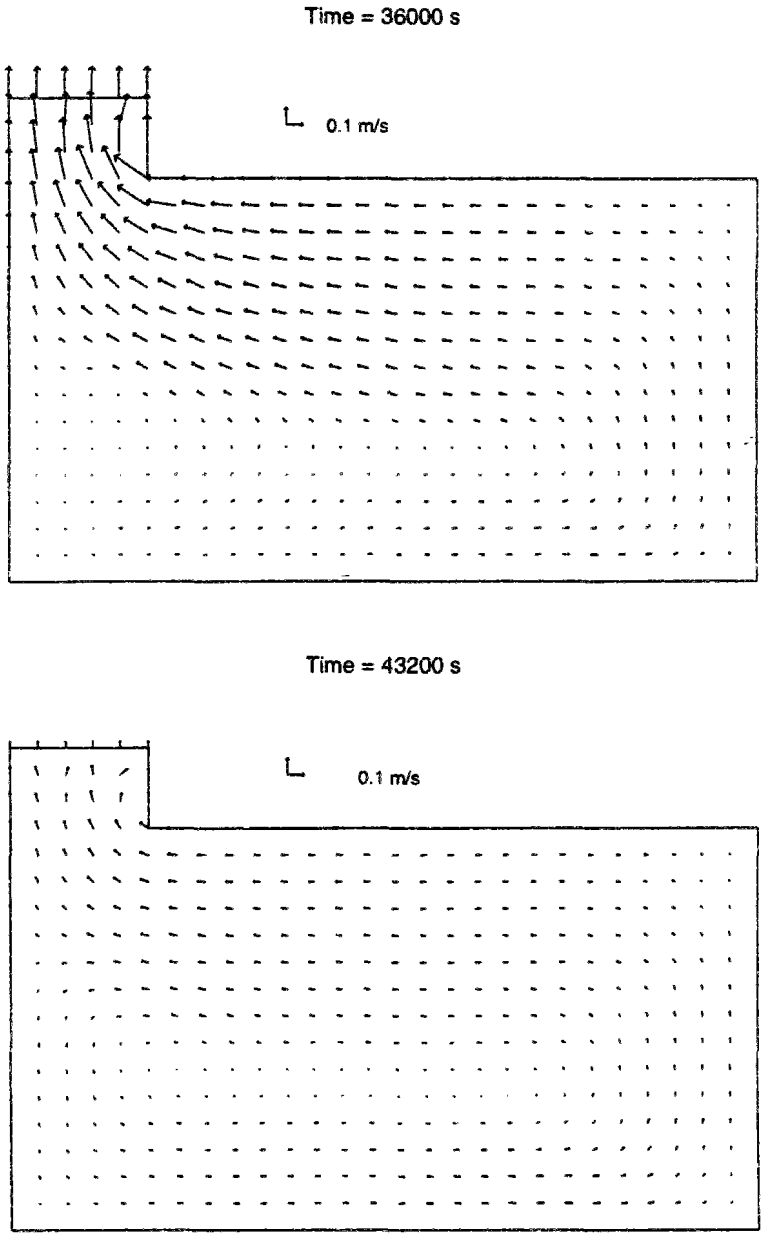


Figure 10.15: Velocity field in the harbour after $T = 10$ and 12 h

10.3 Separating flow in an expanding flume

The following application deals with a computational domain whose geometry is somewhat similar to the harbour previously studied, as it displays too a salient angle and a sudden widening. This test is based on experiments conducted in 1981 and 1982 by Koppel & Wang in the Laboratory of Fluid Mechanics, Departement of Civil Engineering, Delft University of Technology. It consists of simulating an unsteady separating flow in an expanding flume. Experimental results are reported in (Stelling & Wang, 1984), which includes also, as (Dan N'Guyen, 1993), a numerical investigation about the phenomenon.

The availability of measurements allows us to appraise the overall model behaviour, not only from a qualitative point of view but also quantitatively. It provides an opportunity to investigate the influence of *physical* parameters, namely wall friction and eddy diffusivity, and to underline some limitations of depth-averaged modelling approaches.

10.3.1 Experimental conditions and observations

The figure 10.16 presents the geometry of the experimental flume. Its bottom is flat : its elevation in the chosen coordinate system is $z_b = -0.1$ m. The magnitude of bottom friction is characterized by a Chezy coefficient of 62.64 $m^{1/2}.s^{-1}$.

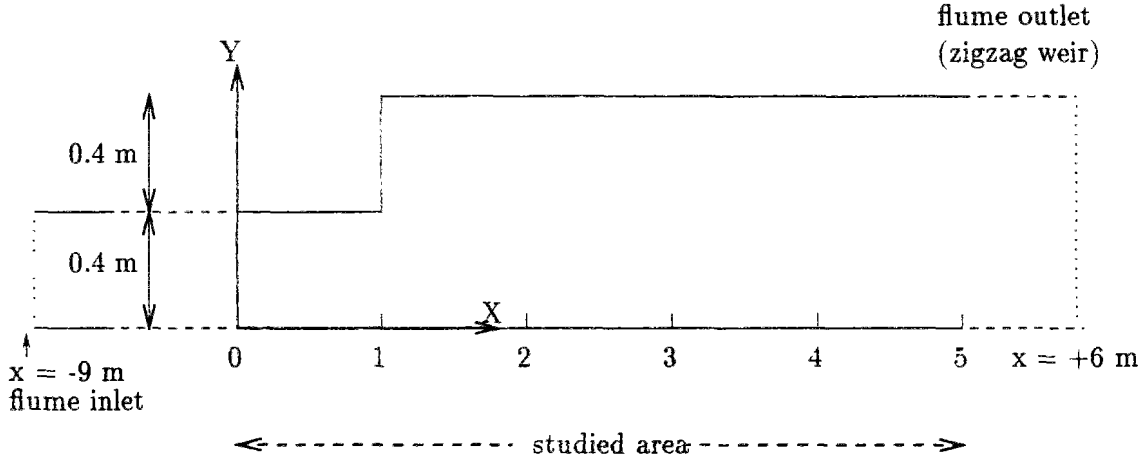


Figure 10.16: Experimental flume

At the beginning of the experiment, the flume was at rest ($u = v = 0$) and we had $\zeta \simeq 0$ (more precisely $h(x, y, 0) = 0.096$ m).

Open boundary conditions A sinusoidal discharge, with amplitude 16 l/s and period 150 s, has been imposed at the flume inlet (section $x = -9$ m). However, owing to the deformation during the propagation within the upstream, narrow, part of the flume, neither velocity nor wave height is a simple sine function at the expansion. According to the measured data, these variables can be expressed in the form of Fourier series :

- at section $x = 0$, for $0 \leq t \leq 75$s,

$$v(x,y,t) = 0 \quad \text{and} \quad u(x,y,t) = \sum_{j=1}^{3} u_j \sin \omega_j t$$

with $\omega_j = \frac{2\pi}{150} j$ $(j = 1,2,3)$, $u_1 = 0.375$ m.s^{-1}, $u_2 = 0.05$ m.s^{-1} and $u_3 = 0.01$ m.s^{-1} .

- at section $x = 5$ m,

$$\zeta(x,y,t) = 0 \text{ for } 0 \leq t \leq 5\text{s}$$
$$\zeta(x,y,t) = \sum_{j=1}^{3} h_j \sin \omega_j (t-5) \text{ for } 5 \leq t \leq 75$$

where $h_1 = 0.021$ m, $h_2 = 0.001$ m and $h_3 = 0.0005$ m.

The above expressions were used as open boundary conditions in subsequent numerical modelling.

Observed flow patterns During the experiment, the two velocity components in the $x-$ and $y-$ directions were monitored continuously at 145 verticals in the separation region, while the water-depth was surveyed across two sections of the flume. The location of these measuring points is detailed on figure G.18 (appendix G.5.1). From fig. G.18, it is obvious that the experimental scheme allows one to describe fairly well what occurs in the one-meter long area immediately downstream of the expansion, but far less precisely what happens further downstream.

Based on the continuous measurements, depth-averaged velocities were estimated every 10 s, starting from $t = 5$ s : these data are given in (Stelling & Wang, 1984). The corresponding flow patterns are displayed on figures 10.19 and 10.20 :

- A separating eddy occurs immediately behind the protruding corner at the very beginning of the flow. It creates a recirculation region which develops with time : from 15 to 35 s (which corresponds to the accelerating phase of the flow), the eddy size increases as the eddy centre is moving downstream.

- A very slight recirculation develops in the concave corner. It is barely noticeable, except for $t = 25$ and 35 s.

- In the decelerating phase, there appear to be two eddies. (Stelling & Wang, 1984) consider that the secondary eddy is perceptible only for $t \geq 45$ s. However, on close observation of experimental results, it appears that the secondary eddy is beginning to develop as soon as $t = 35$ s (cf area around point $(x = 1.48, y = 0.48)$ on third drawing of figure 10.19).
- Between $t = 45$ s and $t = 65$ s, the main eddy seems to keep approximately the same length but it tends to widen. In the meantime, the secondary eddy lengthens steadily. This leads to further increase of the size of the recirculation region, but this increase proceeds at a much slower pace than in the accelerating phase of the flow.
- During acceleration, the main stream is driving the flow in the separation region, while a slight meandering of the flow is observed during deceleration. Notably, it seems that another, weaker, eddy is developing in the downstream part of the flume ($x \geq 3.5 - 4$ m), along the opposite wall ($y = 0$).

The main features of the eddies, as estimated by (Stelling & Wang, 1984), are given in table G.6 : they consist of the location of eddy centres (x_c, y_c) (indicated with respect to point P $(x, y) = (0,1)$) and the length L_s of the recirculation region downstream of section $x = 1$ m . Due to the location of monitoring points, the exact size of the recirculation area and the exact place of the main eddy centre cannot be estimated very accurately as soon as $t \geq 45$ s. The weakness of velocities in the secondary eddy also makes the estimation of this eddy centre somewhat difficult.

Eddy diffusivity Direct measurements of turbulence quantities were not carried out. However, in order to obtain some information for their numerical applications, (Stelling & Wang, 1984) evaluated eddy viscosities ε from the velocity measurements, with the help of approximate formulae. These estimations are reported in table G.5 (appendix G.5.1).
Eddy viscosities are found to be space and time varying. Their instantaneous values belong to the range 10^{-4} to 10^{-3} m^2.s^{-1}, the strongest values being observed when $40 \leq t \leq 60$ s . Comparing the values in table G.5 and the observed eddy locations, it appears that eddy viscosities, not surprisingly, reach their maximum in the recirculation area corresponding to the main eddy, more exactly at its periphery, where velocities are largest. Viscosities are much smaller outside this zone (approximately 3 to 7 times smaller, when $t \geq 30$ s).
In (Stelling & Wang, 1984) as in (Dan N'Guyen, 1993), for the purpose of numerical calculations, the eddy diffusivity is set to a constant, uniform and isotropic value. In the first case, ε is set to $2.3\,10^{-4}$, then to 10^{-3} in order to begin assessing its influence on the forecasts. In the second case, only one simulation with $\varepsilon = 2.3\,10^{-4}$ is performed.
We have looked for an explanation of the use of value $\varepsilon = 2.3\,10^{-4}$ m^2.s^{-1}. It appears that this corresponds approximately to the geometric mean of the viscosities estimated throughout the experiment.

Supplementary boundary conditions Quoting (Stelling & Wang, 1984), "it is believed that the characteristics of the unsteady separating flow (the development of eddies, the splitting of the main eddy, the meandering of the main stream, etc ...) *depend on boundary conditions* as well as flow conditions, such as the magnitude of the velocity, acceleration and deceleration, the period, ..." That's why we devote some space hereafter to describe the applied boundary conditions.

(a) closed boundaries (Stelling & Wang, 1984) apply a model based on an ADI/predictor-corrector scheme, where the different flow variables are defined on staggered grids (cf figure 10.17). At a wall boundary, the normal velocity is set to zero. The tangential velocity is assumed to obey relation 8.72 (introduced in section 8.6.2), namely

$$(1-\alpha)\,U_t + \alpha\,\delta l\,\frac{\partial U_t}{\partial n} = 0$$

where U_t denotes the tangential velocity, n the direction normal to the boundary, δl a length related to the distance between the closed boundary and the nearest computational point, α (in $[0,1]$) a "slip parameter". Let superscripts "w" and "next" refer respectively to the value of velocity at the wall and at the closest interior node. If $\frac{\partial U_t}{\partial n}$ is first-order approximated by $\left(U_t^w - U_t^{\text{next}}\right)/\delta l$, the boundary condition reduces simply to : $U_t^w = \alpha\,U_t^{\text{next}}$.

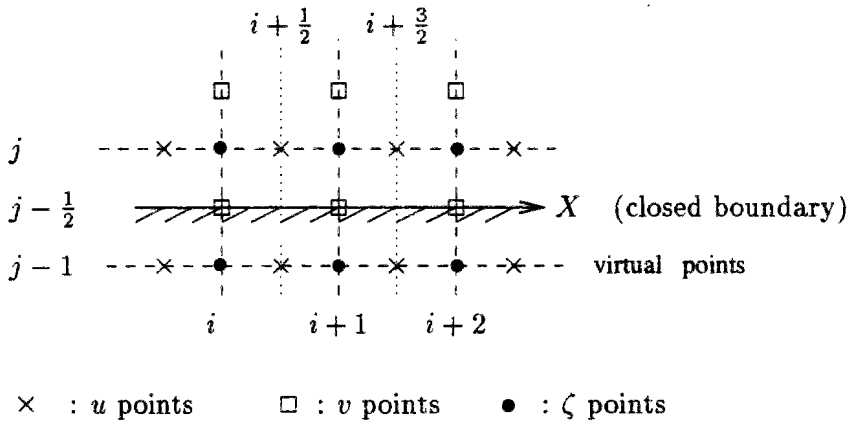


Figure 10.17: Staggered grid used in Stelling & Wang model

α theoretically depends on wall roughness and local flow conditions and should thus be space and time varying. However, as no deterministic formula allowing computation of α was available to Stelling & Wang (by the way, the extension to small-scale experimental devices of empirical formulations developed by (Mary, 1982) in the case of fluvial flows in large water bodies appears to be dubious), they applied the following strategy. For one simulation, α is set to a constant and uniform value; yet several simulations were

performed with different α values, allowing to apply alternatively no-slip, perfect-slip and partial-slip conditions, and examine their impact on the forecast flow patterns.

The boundary condition is introduced in the equations as follows. Let us consider a wall boundary parallel to Ox. The staggered grid is such that u-nodes (i.e. computational nodes where the velocity along the x-direction is computed) are not located on the wall but half a mesh size apart from it (cf fig 10.17). By using the slip condition it is possible to estimate u on the wall and then, assuming that u-gradient across the wall remains constant, to define u at "virtual" nodes, which are mirror images of the boundary nodes with respect to the wall (cf fig 10.18). In Stelling & Wang model, all derivatives are estimated by centred differences, except for some of the advective terms (see more details in (Stelling & Wang, 1984)). The use of the virtual points allows to compute these derivatives close to the boundary as for fully interior nodes.

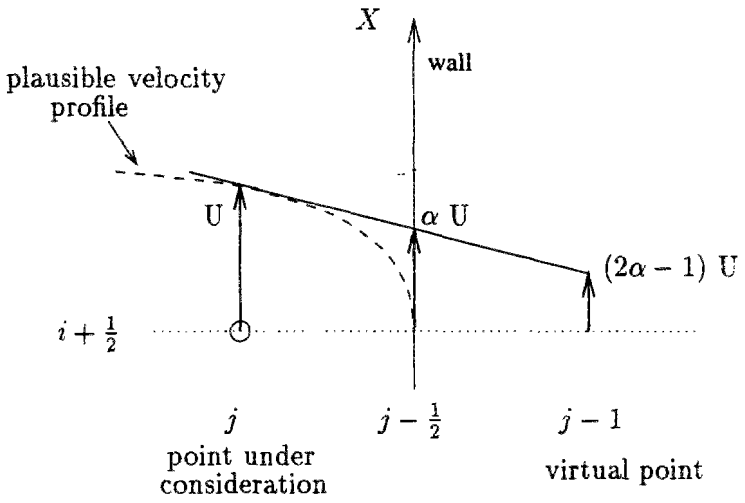


Figure 10.18: Treatment of slip condition on velocity in Stelling & Wang model

Dan N'Guyen (Dan N'Guyen, 1993) applies the finite difference method whose principles have been described throughout chapter 8 and which makes use of a non-staggered grid. He imposes a no-slip (zero velocity) condition on walls.

We have been applying the same condition as (Stelling & Wang, 1984). However, it is worth noting that, due to the differences between the applied discretizations (staggered versus non-staggered grid, different mesh sizes), the use of the same α value for the present model and for the Stelling & Wang model does not necessarily imply that the same amount of wall friction is prescribed and consequently that the resulting velocity profiles at the wall will be similar.

(b) Open boundary conditions (Stelling & Wang, 1984) give no indications as regards the treatment of boundary conditions. In order to run N'Guyen's model, as ours, it is necessary

to supply an additional condition on the free surface elevation at the flume inlet.

In (Dan N'Guyen, 1993), the author uses a radiation condition (Orlanski, 1976), as frequently applied in tidal and marine applications involving periodic flows :

$$\frac{\partial \zeta}{\partial t} + c\frac{\partial \zeta}{\partial n} = -\left(\frac{\zeta - \zeta_k}{T_f}\right) \tag{10.5}$$

where c is the phase speed of the wave ($c = \sqrt{gh}$). The right-hand term of the equation implies that ζ is forced to some equilibrium value ζ_k with a time scale of the order of T_f . ζ_k can be itself time and space varying : it can describe for instance periodic "background" variations of the mean elevation due to the large-scale circulation around the studied area, as derived from data or from the application of global tidal models.
When T_f is small with respect to the time scale of the simulated phenomena, ζ cannot depart much from ζ_k : this corresponds to the use of "clamped" boundary conditions (open boundary elevation a priori prescribed). On the contrary, when T_f is large, 10.5 tends to represent a pure radiation condition, that is to say the open boundary is transparent to waves travelling at the phase speed c.
(Dan N'Guyen, 1993) sets ζ_k to 0, T_f to 10800 s, which amounts to imposing pure radiation, and discretizes relation 10.5 as suggested in (Blumberg & Kantha, 1985).

According to the analysis of flow equations performed in section 8.2, it appears that the pure radiation condition is adequate when the propagation terms are overwhelmingly dominant within the flow equations so that both ζ and $\vec{U}$ are found to obey indeed a wave equation with wave celerity $\sqrt{gh}$.
We applied once more the dimensional analysis introduced in 8.2. Considering that the typical depth-scale is 0.11 m (according to observations depths varied between 0.1 and 0.12 m during the experiment) and that velocities at the flume inlet can reach about 0.38 m/s, it appears that advective terms are not completely negligible : the Froude number can reach 0.36 so that propagation terms are only 8 times larger than advective terms. This led us to discard the use of the radiation condition. We applied instead the same approach as at the basin inlet in previous section 10.2 : we recall it relies on a close approximation of the momentum conservation equation in order to derive the value of the free surface gradient at the open boundary.

Spatial and temporal discretization Finally, for the numerical experiments, the flume was discretized in (Stelling & Wang, 1984) with mesh sizes $\Delta x = \Delta y = 0.025$m . Dan N'Guyen (Dan N'Guyen, 1993) used a coarser grid, with $\Delta x = \Delta y = 0.05$m . Both chose time step $\Delta t = 0.125$ s. In the first case, the average propagation Courant number C_p (with depth of 0.1 m) is approximately 7, in the second case 3.5 . The maximum advective Courant numbers C_r reached during the simulation are respectively 1.88 and 0.94 . We chose the same spatial

discretization as (Dan N'Guyen, 1993).

Table 10.2: Conditions of past numerical experiments

	Stelling & Wang (84)	Dan N'Guyen (93)
Grid size	$\Delta x = \Delta y = 0.025$ m	$\Delta x = \Delta y = 0.050$ m
Time step	$\Delta t = 0.125$ s	
Flume inlet : velocity (m/s)	$v = 0, \quad u = \sum_{j=1}^{3} u_j \sin \omega_j t$ $\omega_j = \frac{2\pi}{150} j$ and $u_1 = 0.375, u_2 = 0.05, u_3 = 0.01$	
elevation	? ? ?	radiation condition $\frac{\partial \zeta}{\partial t} + c \frac{\partial \zeta}{\partial n} = -\left(\frac{\zeta - \zeta_k}{T_f}\right)$ $T_f = 10800$ s, $\zeta_k = 0$
Downstream boundary : elevation (m)	$t \leq 5 \Longrightarrow \zeta = 0$ $t \geq 5 \Longrightarrow \zeta = \sum_{j=1}^{3} h_j \sin \omega_j (t-5)$ $h_1 = 0.021, h_2 = 0.001, h_3 = 0.0005$	
Closed boundaries : normal velocity	$U_n = 0$	
tangential velocity	$(1-\alpha) U_t + \alpha \delta l \frac{\partial U_t}{\partial n} = 0$	
	no-slip : $\alpha = 0$ perfect-slip : $\alpha = 1$ intermediate : $\alpha = 0.75$ and 0.9	no-slip : $\alpha = 0$
Initial conditions	$u = v = 0$ and $\zeta = 0$	
Bottom friction	Chezy coefficient : $C_h = 62.64$	
Eddy diffusivity (m^2 s)	$2.3\,10^{-4}$ then 10^{-3}	$2.3\,10^{-4}$

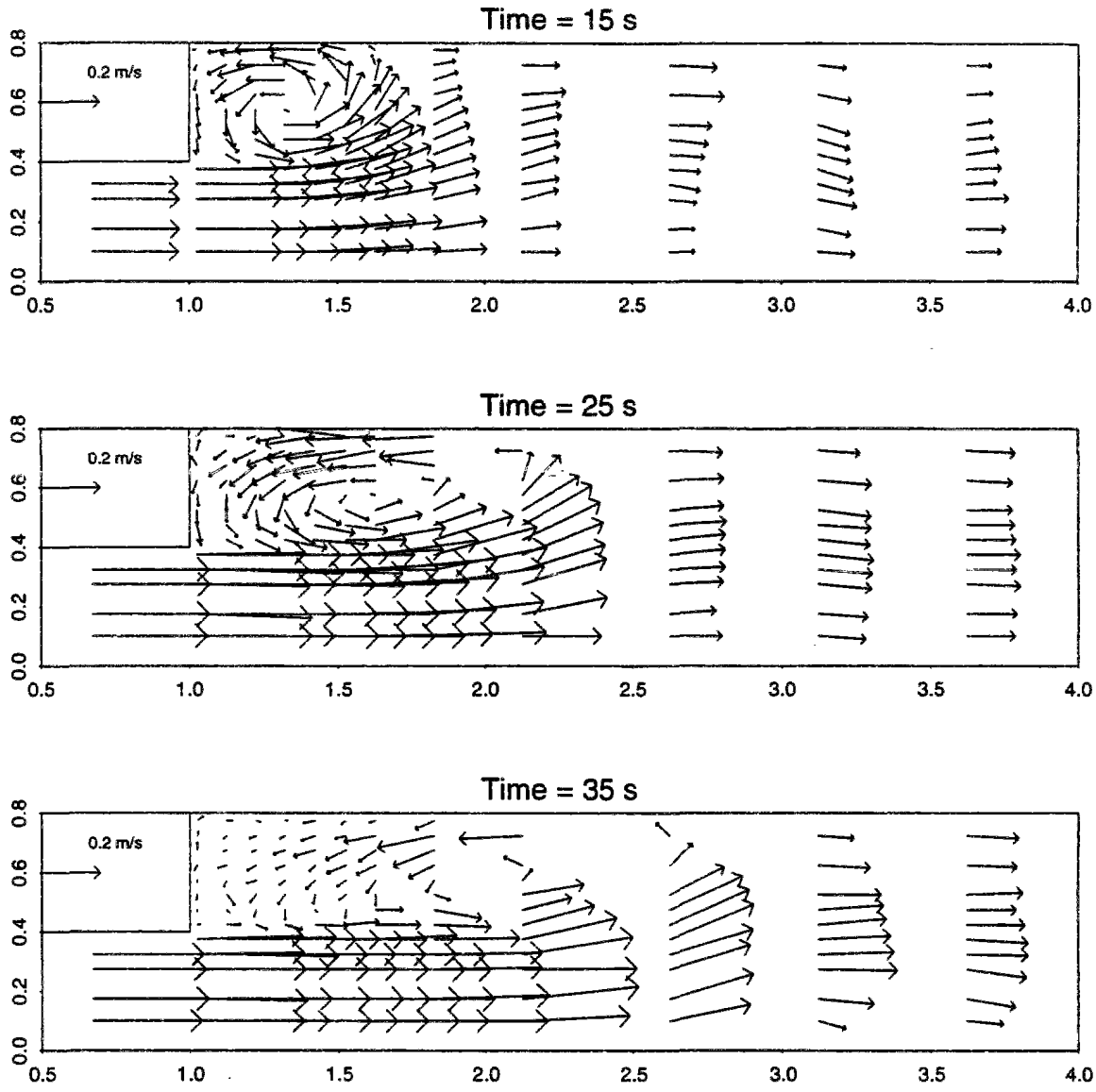


Figure 10.19: Velocity field estimated from experiment at $t = 15$, 25 and 35 s

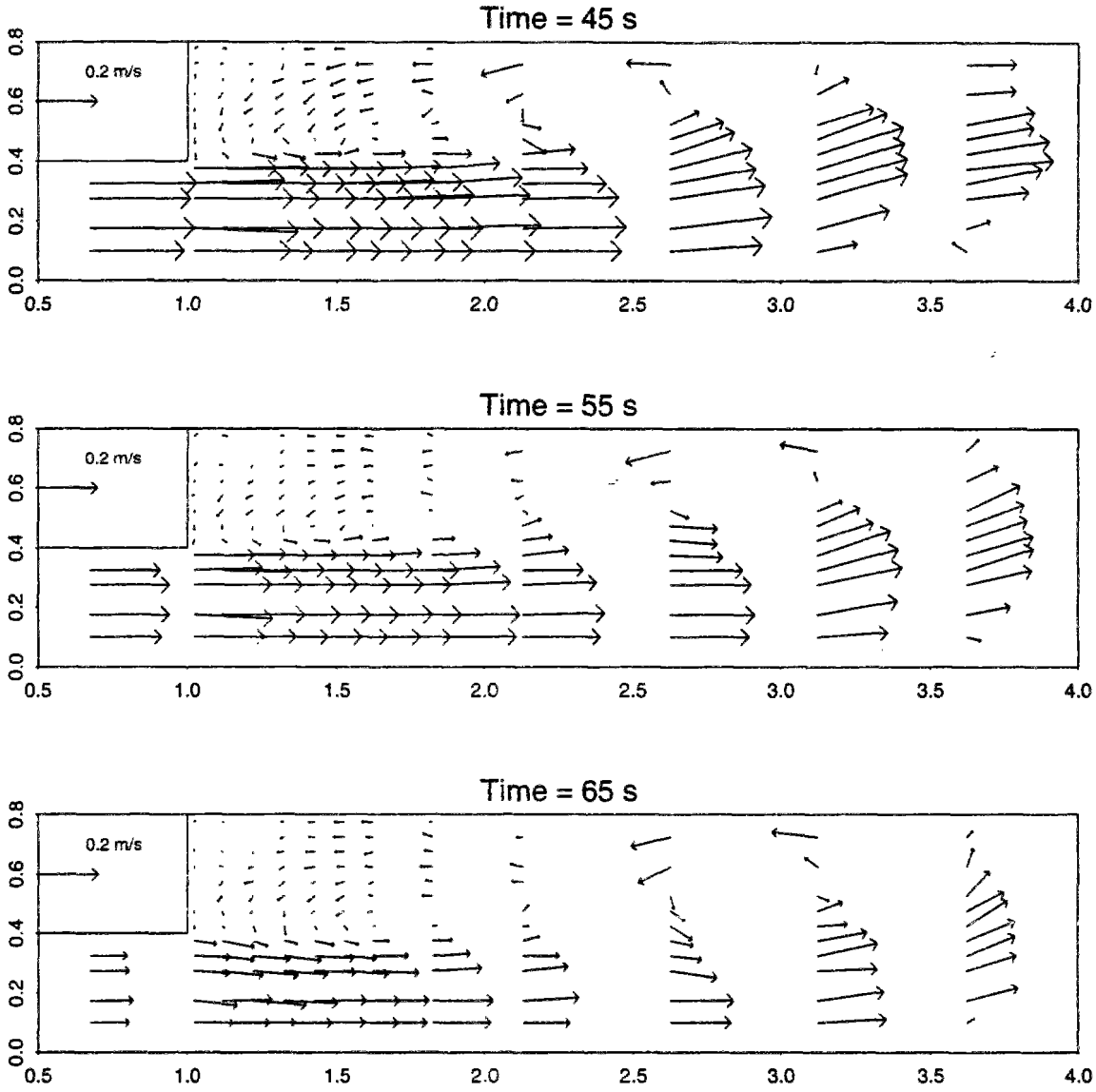


Figure 10.20: Velocity field estimated from experiment at $t = 45$, 55 and 65 s

10.3.2 Numerical experiments

(Stelling & Wang, 1984) observe that computational results are considerably influenced by the conditions specified at closed boundaries and by the parametrization of eddy diffusivity. When these parameters are correctly "tuned", their model succeeds in reproducing qualitatively the main features of the flow, namely the development, growth and downstream displacement of one, then two eddies.
With no slip boundary condition ($\alpha = 0$), several secondary eddies appear, their development is too fast when compared to measurements. When the amount of slip applied along the side walls is increased ($\alpha = 0.75$ or 0.9), the emergence of secondary eddies is delayed, the recirculation length becomes more, the eddies shape is more rectangular. Eventually, with perfect slip boundary condition ($\alpha = 1$), only one eddy develops. Similarly, too much viscosity ($\varepsilon = 10^{-3}$) suppresses secondary eddies.
The origin of the secondary eddy is not fully elucidated. When $\alpha = 0.9$, the secondary eddy appears to be shedding from the salient corner. When $\alpha = 0.75$, Stelling & Wang claim that the second eddy observed at $t = 45$ corresponds to the upstream half of the main eddy, which has split. However, the evolution of this eddy is somewhat strange : it appears to have shrunk between $t = 45$ and 55 !
Quantitatively, the agreement between "best" forecasts and exprimental results remains rather poor, as can be checked in table G.6 (appendix G.5.1). Stelling & Wang state that this is probably due to the the crude modelling both of closed boundary conditions and eddy viscosity and suggest that a full turbulence model is needed.

As we shall see hereafter, our numerical experiments confirm Stelling & Wang's conclusions. Before discussing them, the following comments about numerical resolution can be made :

- Forecasts yielded by "depth/velocity" and "depth/discharge" models are not significantly different.
- In order to reduce as much as possible the temporal discretization error yielded by the semi-implicit development of the flow equations, we always set the implicitation parameters θ (diffusion step) and γ (propagation step) to 0.5 and 0.6 respectively. With these parameters, and the grid sizes and time step mentioned above, numerical behaviour appeared to be fairly satisfying. The average continuity error was observed to remain always inferior to 0.06 mm (worst case, corresponding to perfect slip conditions of simulations). Errors kept close to their mean except at the protruding corner, where they can be as large as 1 or 2 mm.
- The treatment of advective terms influences markedly the computational results. This is illustrated in appendix G.5.2. In summary, when the characteristic method is applied in combination with the bilinear interpolator instead of the bicubic one, velocities in the

recirculation region are dampened : they turn out to be notably inferior to velocities observed during the experiments, and computed using the bicubic interpolator (which are, on the contrary, in the same range as those measured). Eddies grow more slowly and the extent of the recirculation area is underestimated. The sensitivity to boundary conditions is less. This behaviour led us to suspect that the resolution of advection by bilinear interpolation yielded significant numerical diffusion, which is the reason why we discarded the use of this interpolator for that test case.

- Due to the scarcity of available measurements, it was impossible to check thoroughly the model accuracy as regards water-depth forecasts. We noted the following :
  - Computed water-depths at the two measuring cross-sections always are in the good range. However, they seem to display some phase shift. Indeed, according to measures, the maximum wave height is reached at $t = 35$. In forecasts, it occurs at $t \simeq 41 - 42$ and water depths at $t = 45$ are still larger than at $t = 35$.
  - Observed water depths immediatly downstream of the flume widening (22.5 cm downstream) display a transverse gradient : depths in the flume expansion are lower than in the main flow for $15 \leq t \leq 55$. The difference amounts approximately to 1 mm for $t = 25$, 1.5 mm for $t = 35$. At $t = 15, 45, 55$, the surface profile across the flume section is nearly linear. On the contrary, for $t = 25$ and 35, the free surface is nearly flat in the main flow as in the flume expansion but there is a sharp decrease at the widening ($y \simeq 0.4$). The model fails to reproduce this. It does forecast a transverse variation of depth but this is much too smooth and too slight (never more than 0.3 mm).

    The transverse slope is less important further downstream. In the second control cross-section (60 cm downstream of the first), transverse gradients are noticeable only for $t = 25$ and 35, are much weaker, and better approximated by the model.

10.3.2.1 Influence of slip condition at side walls

(nb : All results discussed here have been obtained setting eddy diffusivity to $\varepsilon = 2.3\,10^{-4}\ m^2/s$. The role of eddy diffusivity will be studied in next section)

Figures 10.21 and 10.22 display the velocity fields computed under no-slip condition. Differences with the observations are obvious. The model tends to overestimate the slight circulation in the concave corner : it concerns a much larger area than observed (compare lowest parts of fig. 10.21 and fig. 10.19). On the other hand, it underestimates the development of the main eddy in the accelerating phase of the flow : the eddy length at $t = 25$ and 35 s is respectively 1.3 and 1.5 m, with respect to observed sizes of 1.4 and 1.9 m . The computed eddy is also somewhat thinner. In the decelerating phase, the emergence of a second eddy is indeed forecast. This second

eddy is more elongated than measured, while the main eddy is much smaller : its size remains nearly constant, as in the experiment, but reaches only 1.2 m (versus 1.7 m approximately). The extent of the nearly stagnant area keeps on being oversized, which finally induces an overestimation of the total recirculation region. A good point is that the computed velocities have a good order of magnitude (cf section G.5.2). The model also predicts the occurrence of other secondary circulations along the wall opposite to the expansion, which are either not observed, or should be much feebler and limited. Basically, the same behaviour (too fast emergence of secondary circulations, notably in the concave corner and along the opposite wall) is reported in (Stelling & Wang, 1984).

Mechanism of eddy development Prior to studying the influence of closed boundary condition on flow patterns we tried to achieve a deeper understanding of the model forecasts. For that purpose, we stored and surveyed model results every second. This tells us the following story :

- A single, round-shaped, eddy develops between 5 and 15 s.
- Between 15 and 35 s, the eddy grows and becomes more elongated. It begins to drift away from the salient corner at $t \simeq 20$.
- We observe four cycles of secondary circulation development :

  1. In the area downstream of the wall and upstream of the main eddy, next to the protruding corner, a very very slight circulation begins to develop. It spectacularly gains strength at $t = 29 - 30$ s but then immediatly merges with the first eddy. This process conveys to the single eddy observed at $t = 35$ its slightly asymmetric shape.

  2. The story repeats itself : birth, at the salient corner, $(t = 35 - 36)$ and reinforcement $(t = 36 - 38)$ of a secondary eddy, followed by its fusion with the main eddy $(t = 39 - 41)$.
  3. The third cycle starts at $t = 41 - 42$. Its duration is larger than before, as the flow decelerates (cf figures 10.23 and 10.24).
     This time the secondary eddy becomes more important : its size is closest to the main eddy size but the velocities within are definitely much weaker. It simultaneously develops and parts from the wall between $t = 42$ and 47 s.
     Merging occurs between $t = 48$ and 50. At the same time, a new secondary circulation develops upstream of the main eddy. This time it is unclear whether it sheds from the protruding corner or from the splitting of the secondary eddy undergoing the fusion.

  4. The last secondary eddy has an even larger life time, as it still exists at $t = 65$. It lengthens between $t = 51$ and 53 (and is still bigger than the preceding) then seems

to stabilize ($54 \leq t \leq 58$), its only modification being a slight widening.
This second eddy separates from the wall only at $t = 59-60$. As it drifts downstream, the flow between it and the corner remains first nearly stagnant. Then, a very weak recirculation once again appears to develop, at $t \simeq 65$.

As no more detailed measurements are available, it is impossible to check whether the model describes faithfully the flow dynamics. The numerical experiment demonstrates that the good time scale for observations should be no greater than 1 or 2 s. Single, isolated, observations may lead to erroneous conclusions. For instance, let us consider the flow pattern at $t = 49$ (figure 10.24). Without knowing the past and following history of the eddies, we could falsely conclude it represents the main eddy in the middle of splitting when, on the contrary, it corresponds to an intermediate stage in the merging of first and secondary eddies.

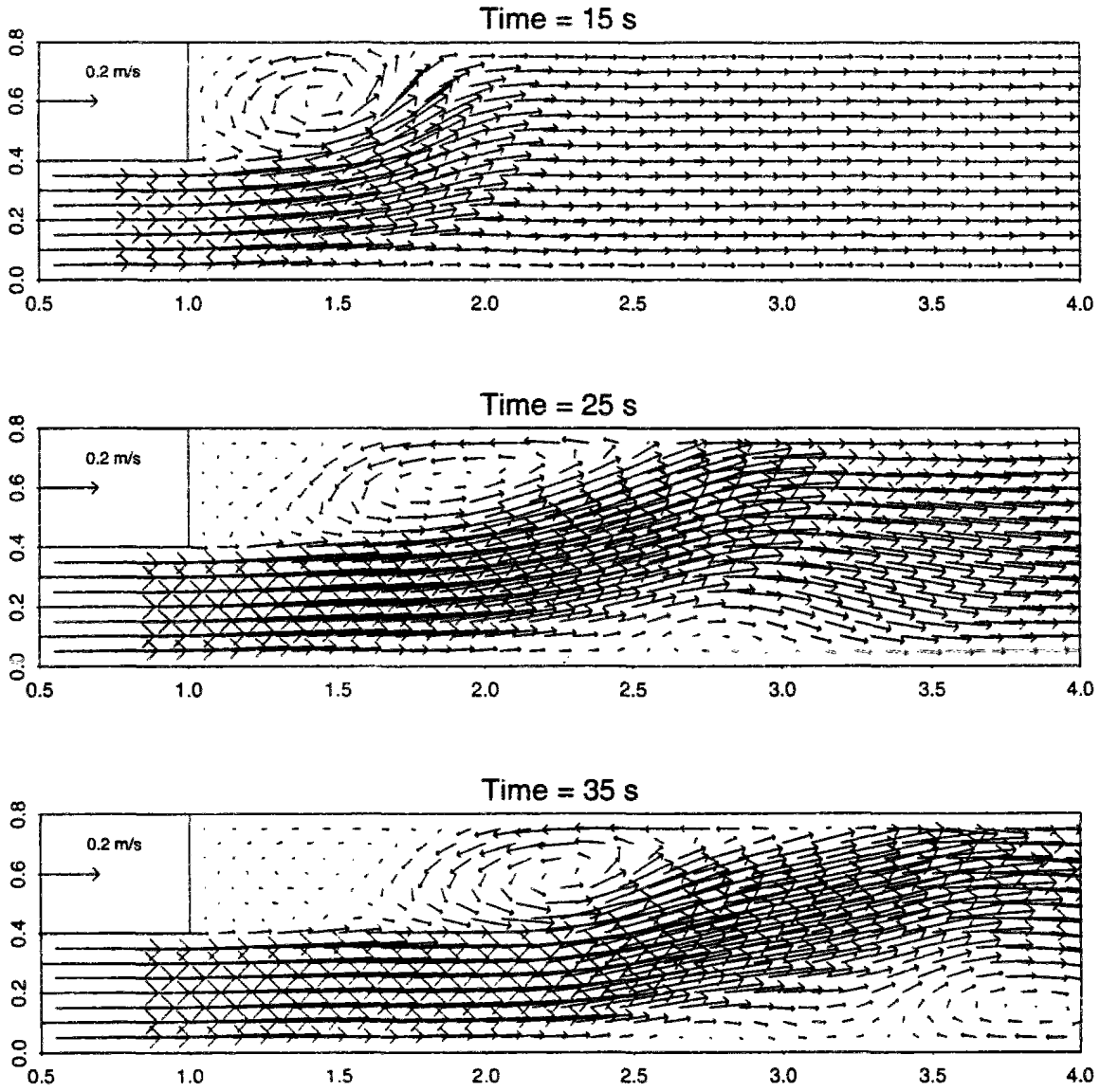


Figure 10.21: No-slip / $\varepsilon = 2.3\,10^{-4}$ / Velocities at $t = 15, 25, 35$ s

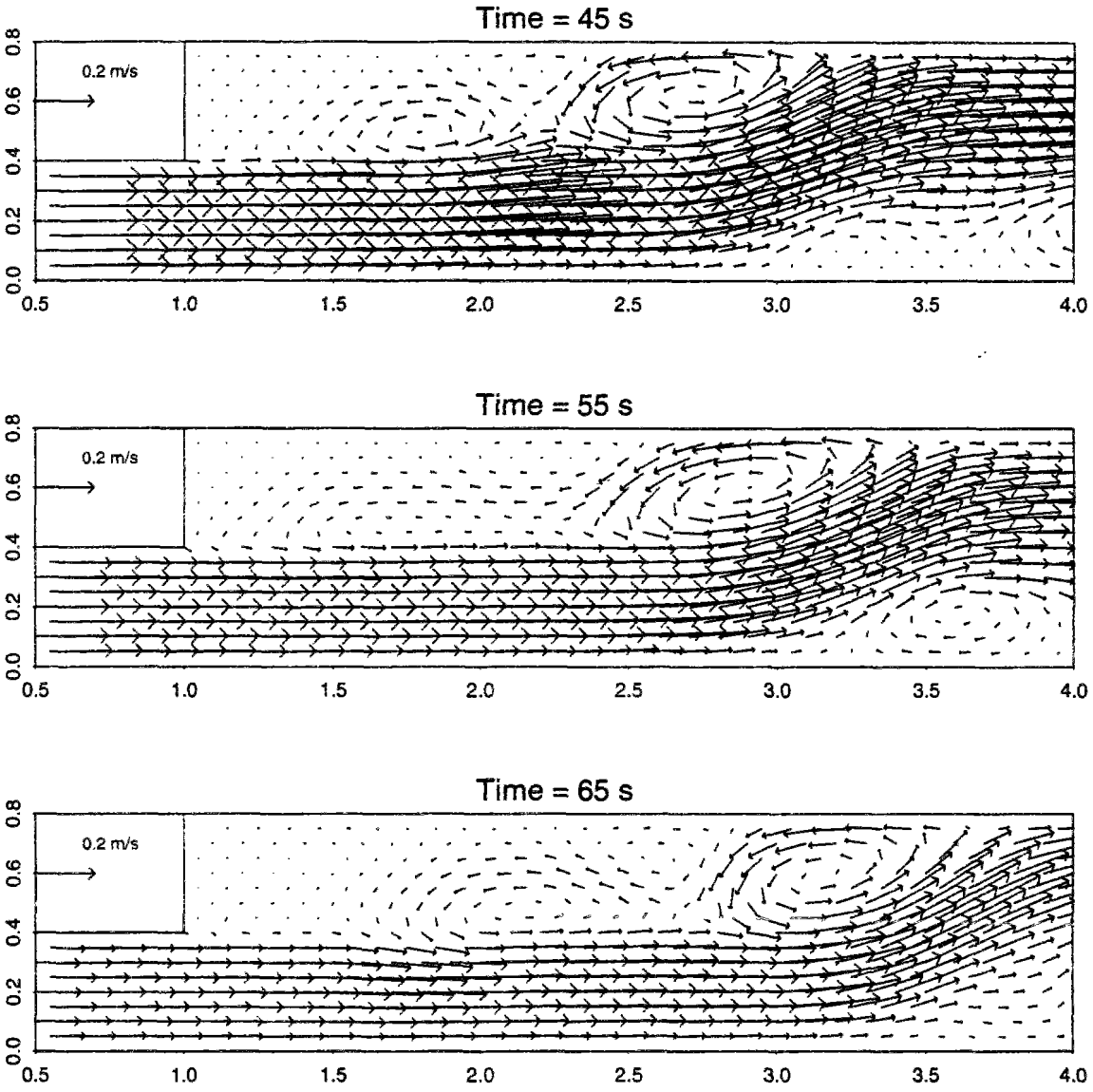


Figure 10.22: No-slip / $\varepsilon = 2.3\,10^{-4}$ / Velocities at $t = 45, 55, 65$ s

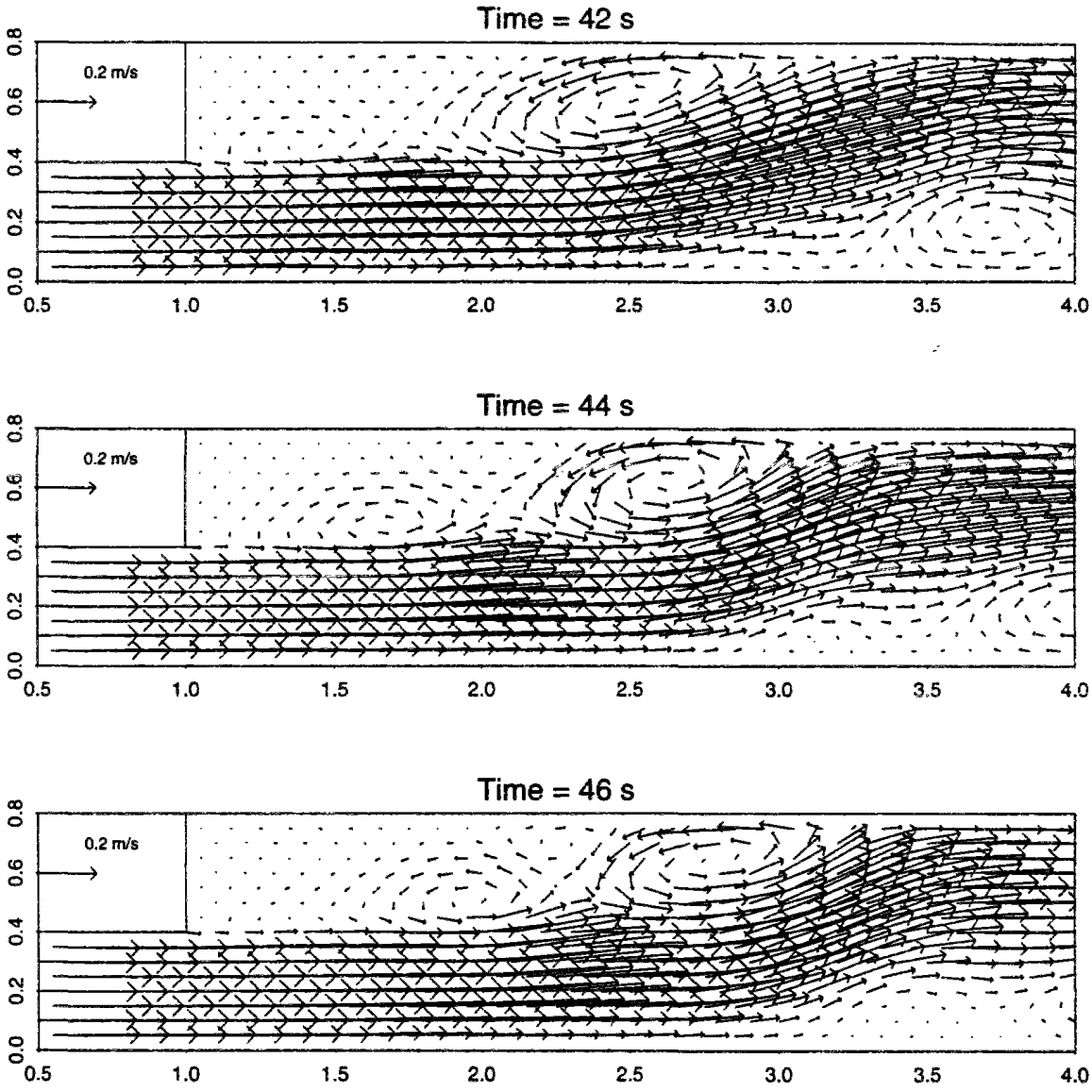


Figure 10.23: Secondary eddy life cycle ($\alpha = 0$, $\varepsilon = 2.3\,10^{-4}$) : (a) Birth and growth

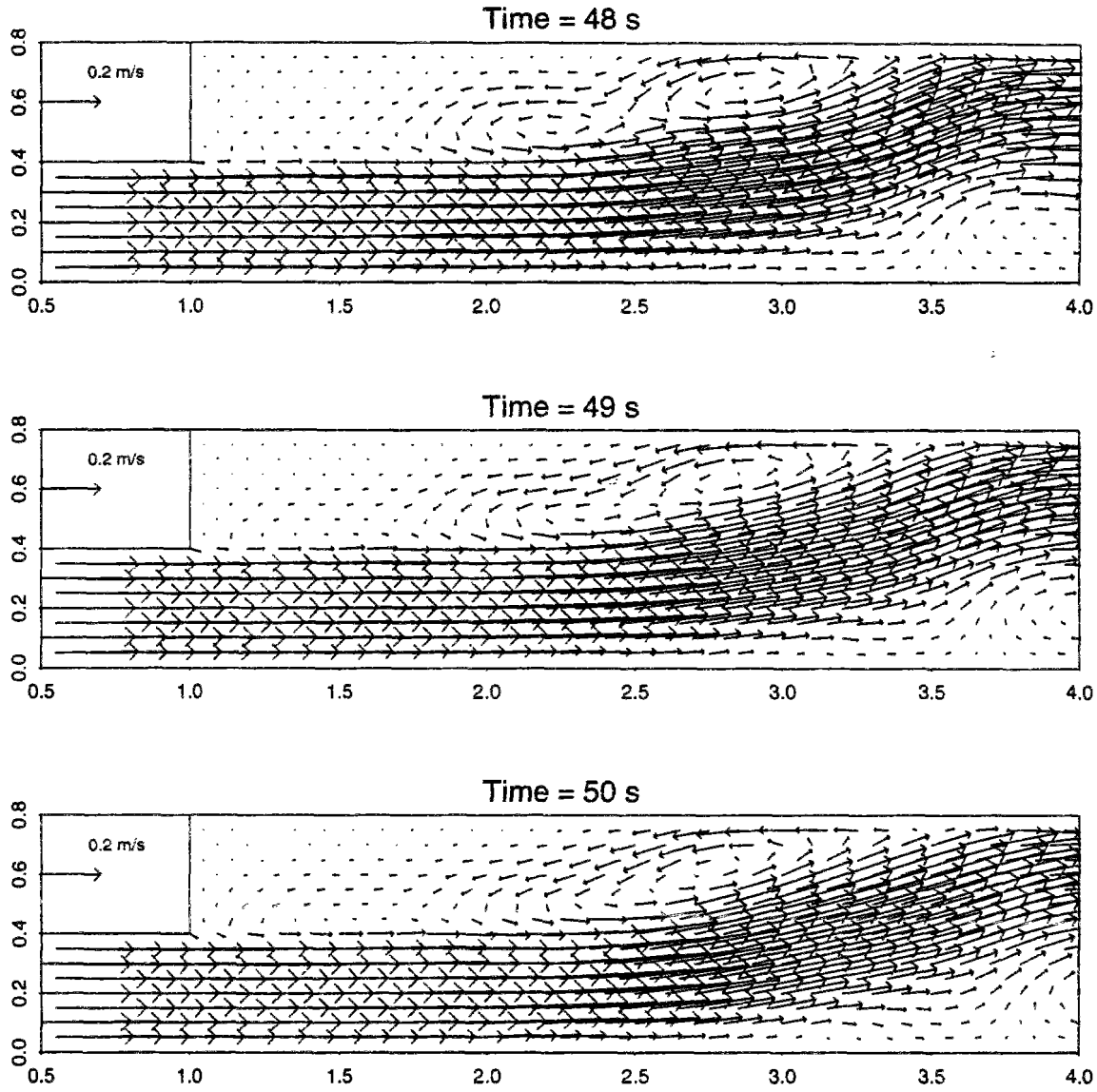


Figure 10.24: Secondary eddy life cycle ($\alpha = 0$, $\varepsilon = 2.3\,10^{-4}$) : (b) merging with first eddy

Secondary circulations according to different slip conditions

$\alpha = 1.00$ **(perfect slip)** As observed by Stelling & Wang, when perfect slip along the side walls is assumed, the development of a single, rectangular-shaped, eddy is forecast and the recirculation length is overestimated (cf figures G.22 and G.23, appendix G.5.3).

$\alpha = 0.75$ **(partial slip)** In the accelerating phase of the flow, a single eddy develops. With respect to observations, it appears to be somewhat too thin and too rectangular (cf figure G.24). At $t = 35$, the extent and strength of the slight recirculation within the concave corner of the flume is underestimated. Yet, these forecasts are in closer agreement with the experiment than when perfect, and above all no-slip, is imposed.

At the beginning of the decelerating phase, the eddy stays next to the salient corner. Its shape is changing. From $t = 44$ to $t = 47$, it seems to be on the verge of splitting, as the circulation within seems to be organizing itself around two distinct centres (cf figure 10.25), one located approximately at $(x, y) = (1.8, 0.6)$, the other at $(x, y) = (2.8, 0.6)$.

Between $t = 47$ and 49, the eddy drifts away from the corner as a very slight secondary eddy is developing. In the meantime, the circulation within the main eddy (which is about 2 meters long) becomes more symmetric. It proceeds around a point located at $x \simeq 2.4/2.5$. The secondary eddy is quickly absorbed by the main one (cf figure G.25).

For the second time, at $t \simeq 53 - 54$, a secondary eddy sheds from the protruding corner (cf figure G.26). In the meantime, the main eddy appears to be once again somewhat distorted, with nearly two centres ($53 \leq t \leq 55$). The secondary circulation is stronger than the previous one. Fusion with the main eddy occurs at $t = 58 - 59$.

A third cycle begins at $t = 61$. This secondary eddy achieves its maximum strength at $t = 64$ (cf figure G.27).

$\alpha = 0.50$ **(partial slip)** We observe this time 5 cycles of secondary circulation development. The first begins at $t = 35 - 36$. The tiny eddy is almost immediately absorbed by the main eddy ($t = 38 - 39$). In the following cycles, the secondary circulation succeeds in gaining strength (cf figure 10.26).

The second and third cycle have the same scenario, and overlap : birth ($t = 41 - 42$ then $t = 48 - 49$), growth (maximum at $t = 44$ and $t = 52$ respectively), fusion with the first eddy ($t = 48 - 49$ and $t = 54 - 55$). The main eddy becomes significantly wider at each cycle while the total recirculation length is increasing but very slowly.

The fourth cycle unfolds differently. The secondary eddy born at $t = 54 - 55$ strengthens till $t = 58$ then weakens but never completely vanishes. In fact, it combines with another eddy (fifth cycle) emerging at $t = 61 - 62$ (cf figure 10.27).

In summary, **the introduction of slip along the side walls has the following consequences :**

- **It delays the emergence of secondary circulations**, and eventually prevents them from developing (perfect slip case) : the first cycle of secondary circulation development starts respectively at $t \simeq 29$, 35 and 47 s for $\alpha = 0$, 0.5 and 0.75 . Slip also affects the life time of secondary eddies.

- **It modifies the balance between main and secondary eddies.** As the slip is increased, the main eddy becomes longer and wider (cf table G.10) while the secondary eddies are smaller and weaker (cf table G.11).

- **It tends to increase the final recirculation length.** While in the accelerating phase ($t \leq 35$), the recirculation length is fairly independent of the applied slip condition, significant discrepancies occur in the decelerating phase (cf table G.9).

- **Velocities in the recirculation area corresponding to the main eddy become larger** (cf table G.8). This is quite logical as increased resistance tends of course to slow down the flow near the side walls.

(nb : for more detailed results, the reader may refer to tables G.8 to G.11 in appendix G.5.3.)
At first glance, the slip condition which leads to the closest agreement between computations and experiment is not far from $\alpha \simeq 0.5$.

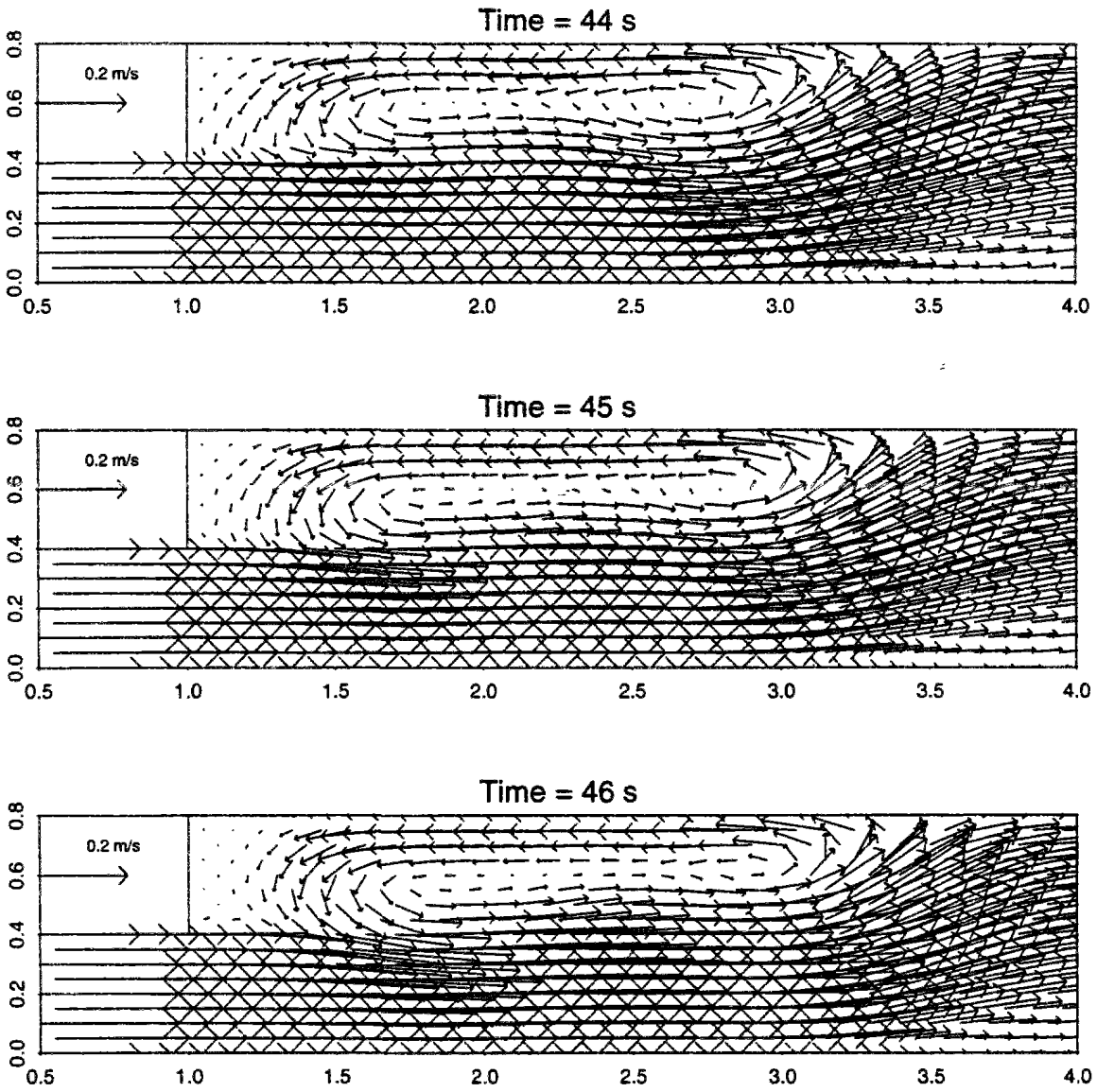


Figure 10.25: Partial slip : $\alpha = 0.75$ / $\varepsilon = 2.3\,10^{-4}$ / Distorsion of main eddy for $44 \leq t \leq 46$ s

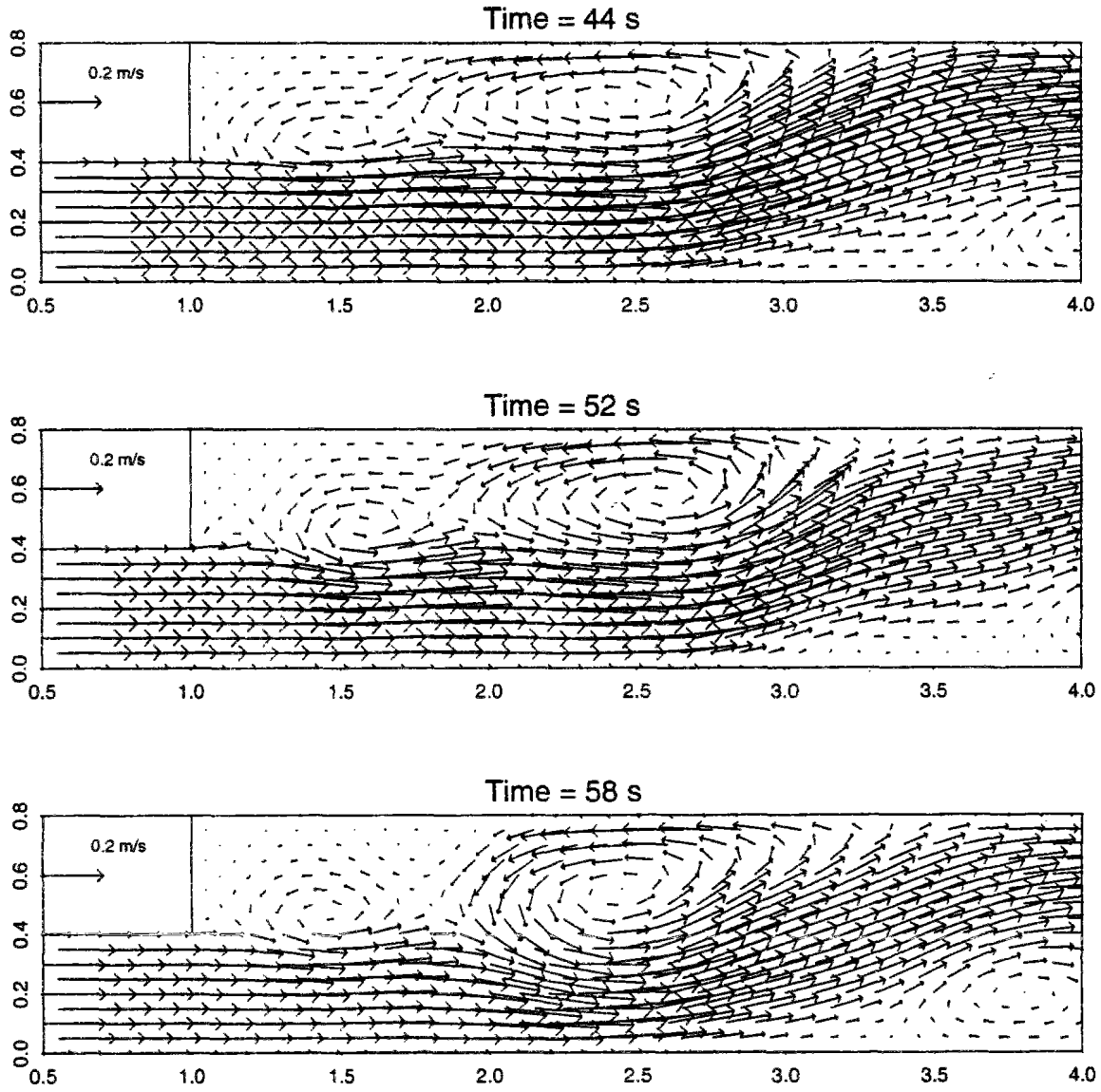


Figure 10.26: Partial slip : $\alpha = 0.50$ / $\varepsilon = 2.3\,10^{-4}$ / Secondary eddies at their stage of maximum strength

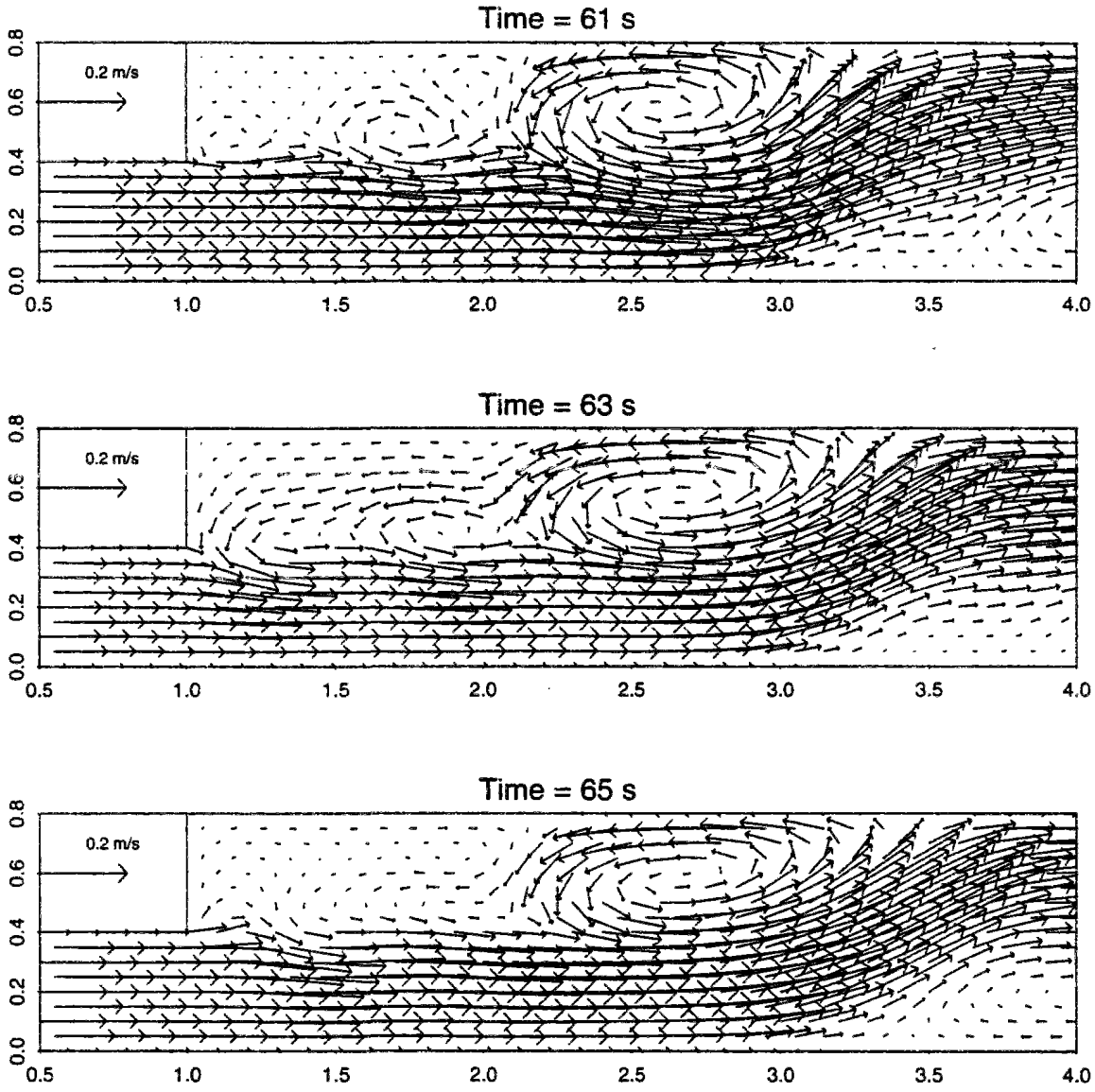


Figure 10.27: Partial slip : $\alpha = 0.50$ / $\varepsilon = 2.3\,10^{-4}$ / Evolution of the last secondary eddies

10.3.2.2 Influence of eddy diffusivity

Our investigation of the part played by eddy diffusivity is of limited extent. Assuming a partial slip condition along the side walls (with $\alpha = 0.5$), we set the eddy diffusivity to a constant and uniform value successively equal to 10^{-4} and 10^{-3} $m^2.s^{-1}$ (which are respectively the lower and upper limit observed by Stelling & Wang), then $4.6\,10^{-3}$ $m^2.s^{-1}$ (namely twice the value recommended in (Stelling & Wang, 1984)). This is insufficient to make firm conclusions about eddy diffusivity influence. However, we may already make the following comments :

- **Whatever the applied diffusivity, several cycles of secondary circulation development are observed.** We observe 5 cycles for $\varepsilon = 2.3$ and $4.6\,10^{-4}$, starting approximately at the same time, namely $t \simeq 35$ s, 4 cycles when $\varepsilon = 10^{-4}$, 2 cycles for $\varepsilon = 10^{-3}$ (starting in the last two cases at $t \simeq 41$).

- **The eddy diffusivity has a marked influence on eddy characteristics.**
  - As the eddy diffusivity is increased, the main eddy becomes larger and its centre is located further downstream (cf table G.12, appendix G.5.4).
  - Increased eddy diffusivity results in a longer recirculation area.
  - As slip, eddy diffusivity affects the magnitude of velocities within the recirculation region : the larger the applied diffusivity, the slower the flow.

- **The eddy diffusivity not only influences the number and duration of the cycles of secondary circulation development. Overall, it controls the mechanism of secondary eddy formation.**
  - For the two smallest values of ε, all cycles of secondary circulation development, except the last, unfold as follows : the secondary eddy appears to shed from the protruding corner, it grows (keeping smaller than the main eddy), drifts away from the corner and merges into the main eddy while a new circulation develops immediately downstream of the corner. The last cycle is different. Indeed, the secondary eddy observed at $t = 65$ comes from the merging of two secondary circulations, born respectively around $t = 55$ and $t = 61$.
  - For the largest ε value, the forecast flow dynamics are quite different. The first secondary eddy comes from the splitting of the main eddy ($42 \leq t \leq 44$) (figure G.30). It does not really deserve to be called "secondary". Indeed, it lenghtens steadily (figure G.31) while the other eddy (corresponding to the downstream part of the main eddy) shortens, weakens and finally vanishes (for $t \simeq 50$). The second cycle unfolds similarly : the existing eddy splits ($51 \leq t \leq 53$), its upstream half grows while its downstream half disappears. This time, there is no further emergence of secondary circulation and only one eddy is observed at $t = 65$. Final recirculation velocities are much too weak (about five times too small).

- The dynamics are even more complex for the intermediate $\varepsilon = 4.6\,10^{-4}$. The first two secondary eddies are due to the splitting of the main eddy (occurring respectively from $t = 36$ to $t = 38$, then between $t = 41$ and 43). However, contrary to what we observe with the greater eddy diffusivity, they remain smaller than the eddy downstream part, do not annex it but are eventually merged into it. The following cycles unfold as for the smallest diffusivity values. Recirculation velocities during the decelerating phase are also too weak (by 20 to 30 %).

- The modifications induced by different choices of eddy diffusivity are particularly obvious in the decelerating phase of the flow. In the accelerating phase, discrepancies do exist but they are definitely more limited (consider figures 10.28 to 10.30 on one hand, figures 10.31 to 10.33 the other hand).

  Considering the variation of eddy features (including recirculation velocity) according to diffusivity, it appears that the suggested value $\varepsilon = 2.3\,10^{-4}$ $m^2.s^{-1}$ is at the same time too large for the flow beginning (since results yielded by $\varepsilon = 10^{-4}$ are more satisfying till $t = 35$) and then too moderate (insufficient recirculation length and main eddy size).

In conclusion, the forecasts appear to be quite sensitive to the eddy diffusivity value. As underlined by (Stelling & Wang, 1984), the assumption of a constant and uniform eddy diffusivity is much too crude. A correct approximation of this complex flow requires most probably the use of a proper turbulence model able to prescribe some dependence of the eddy diffusivity on the local flow conditions.

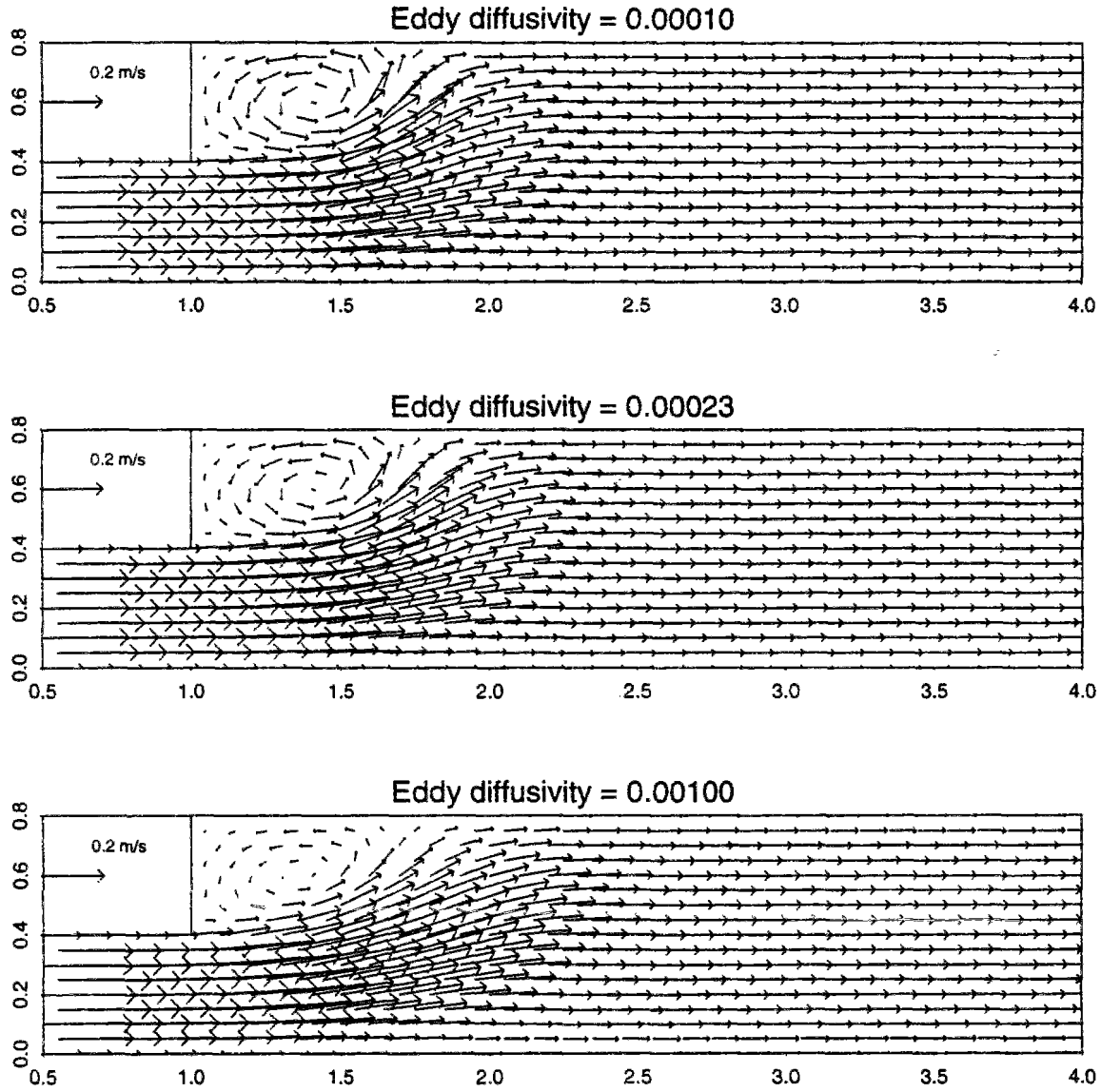


Figure 10.28: Influence of eddy diffusivity on flow patterns computed at $t = 15$ s

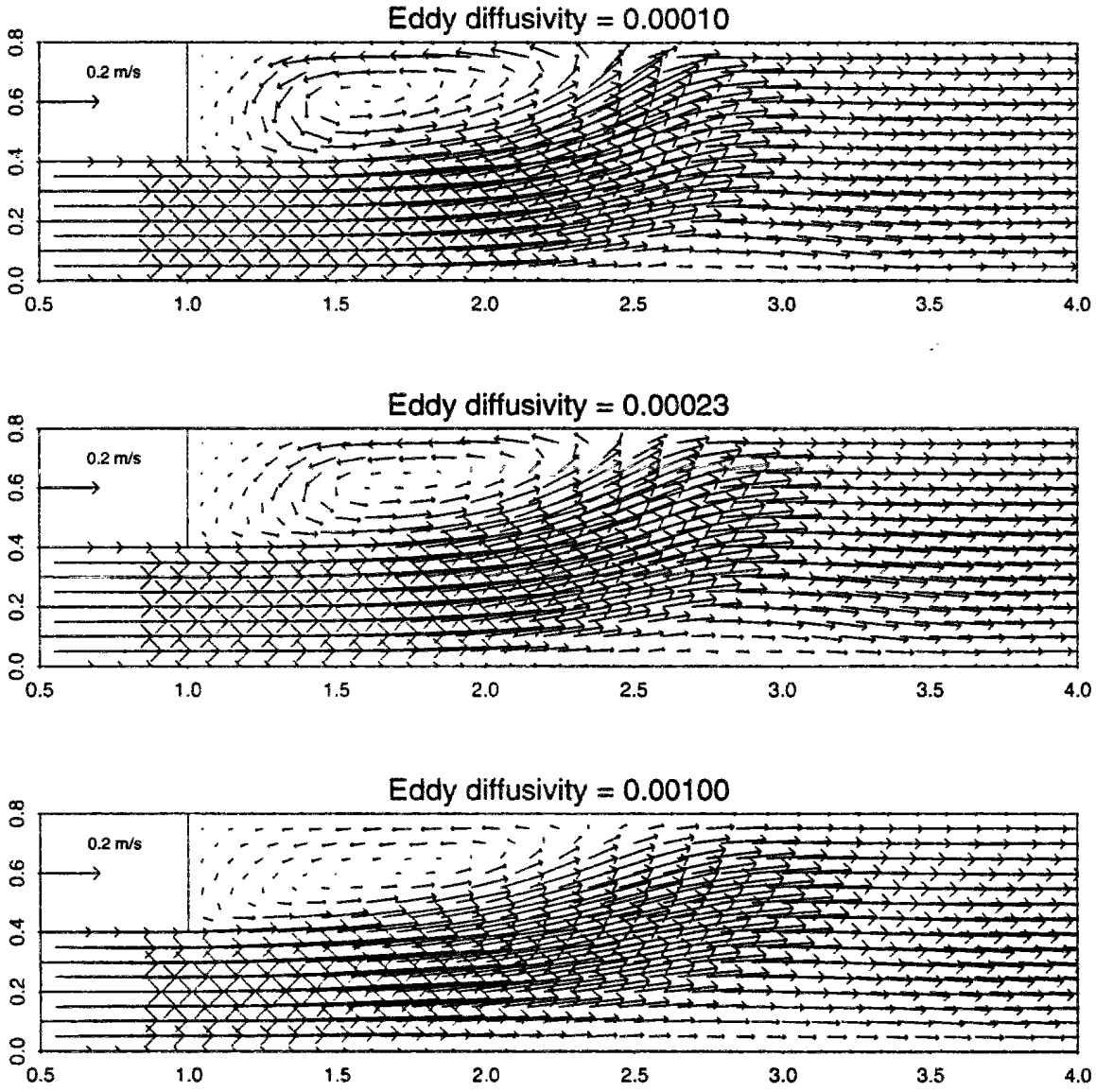


Figure 10.29: Influence of eddy diffusivity on flow patterns computed at $t = 25$ s

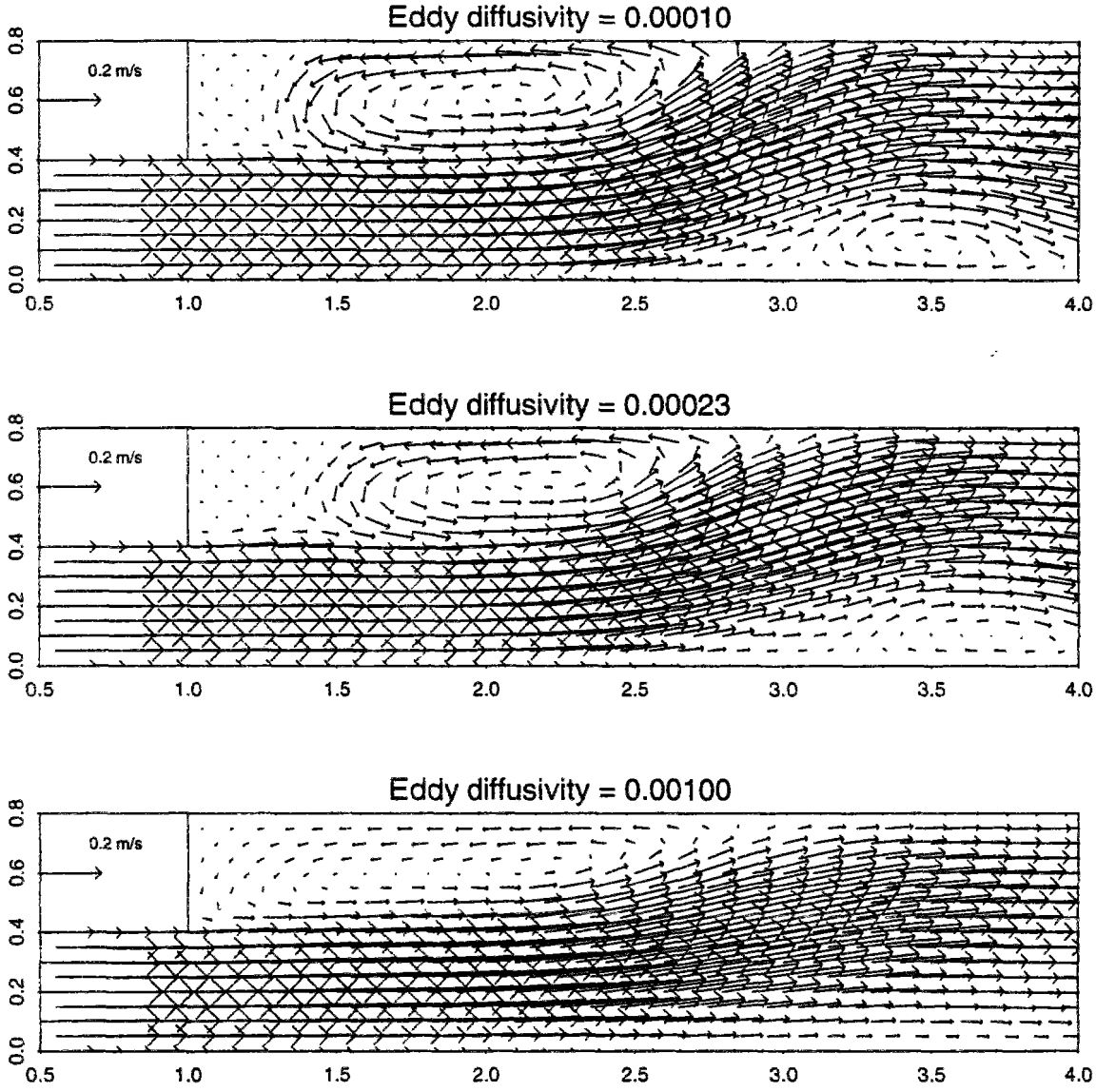


Figure 10.30: Influence of eddy diffusivity on flow patterns computed at $t = 35$ s

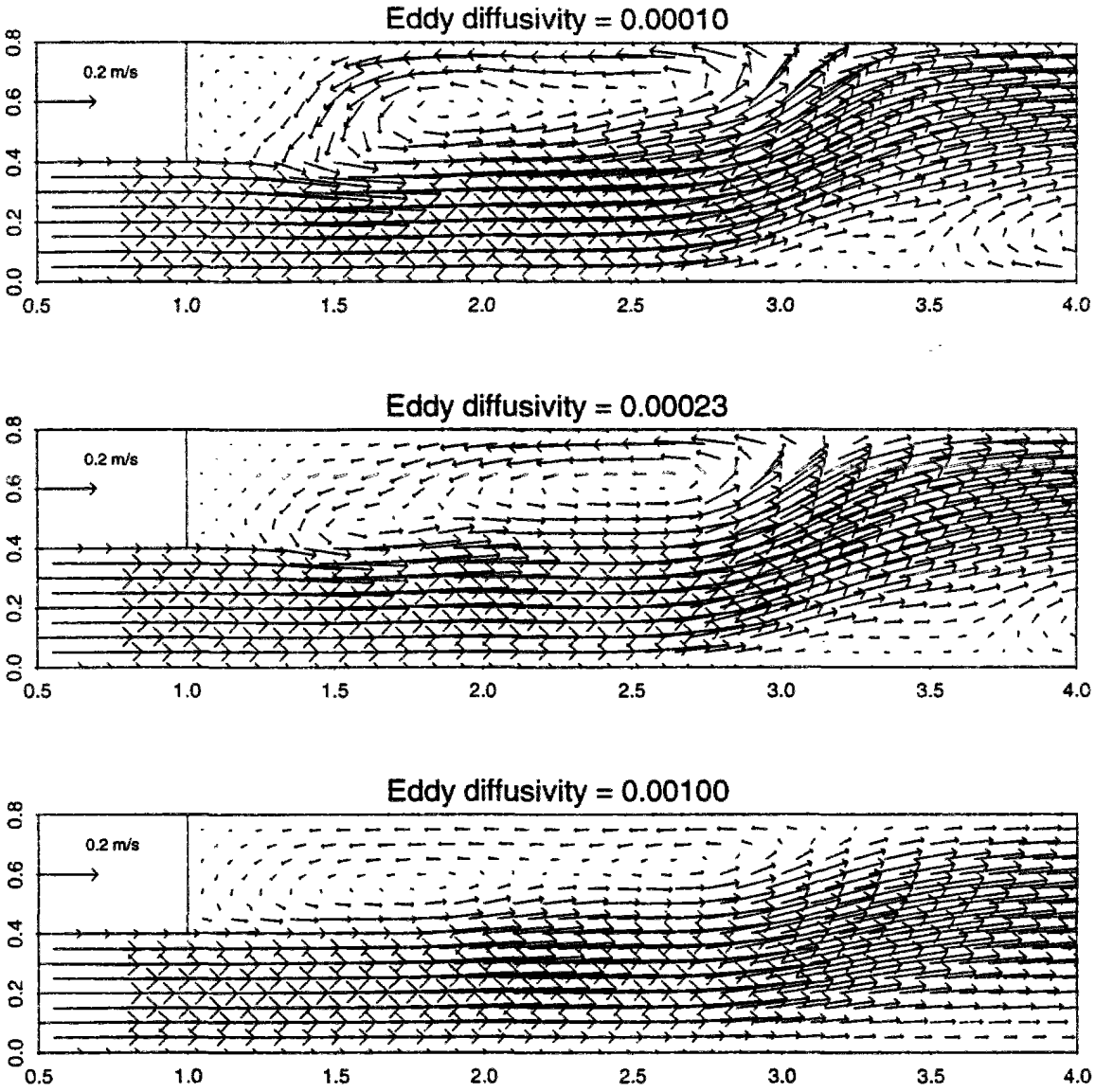


Figure 10.31: Influence of eddy diffusivity on flow patterns computed at $t = 45$ s

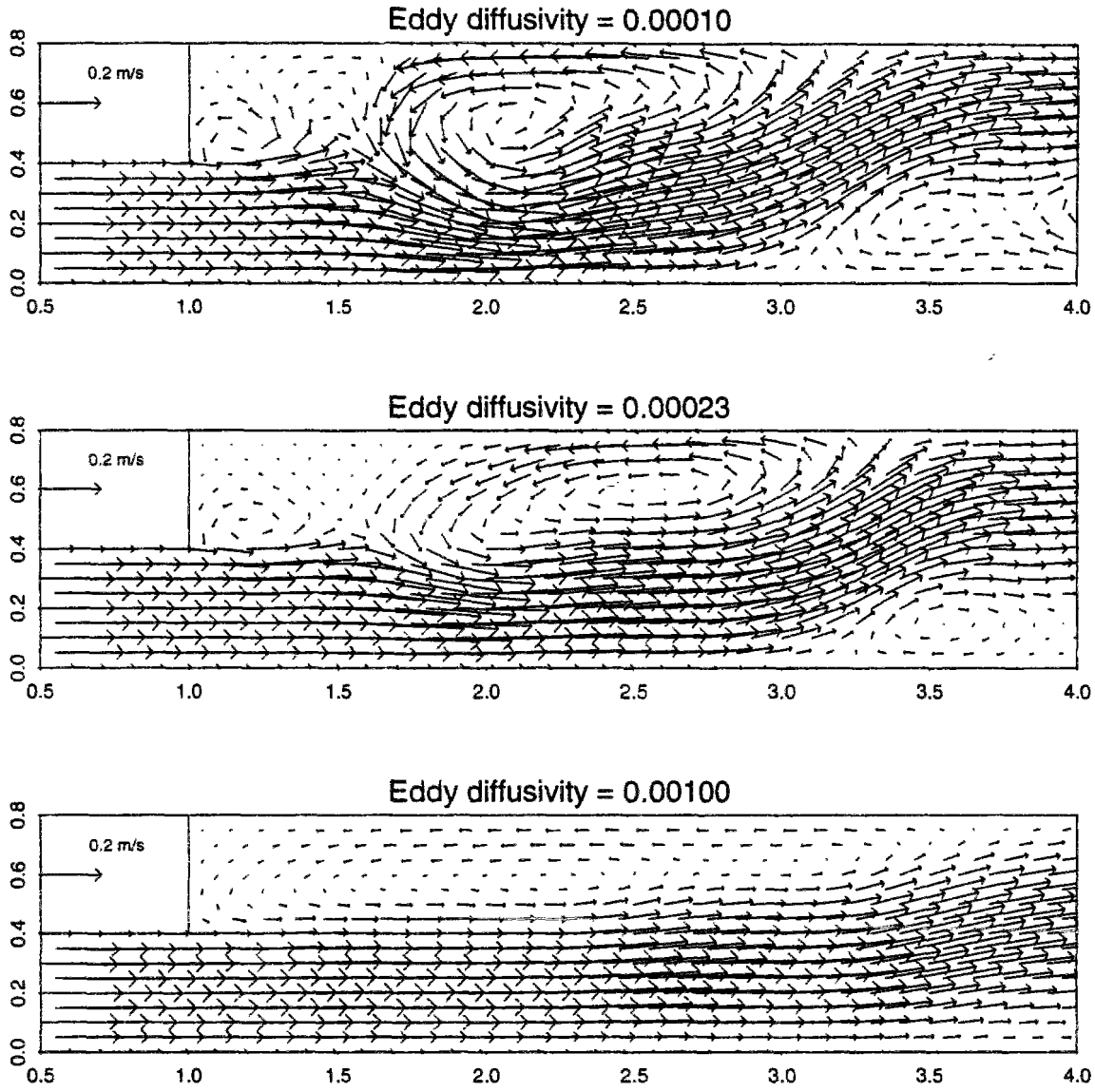


Figure 10.32: Influence of eddy diffusivity on flow patterns computed at $t = 55$ s

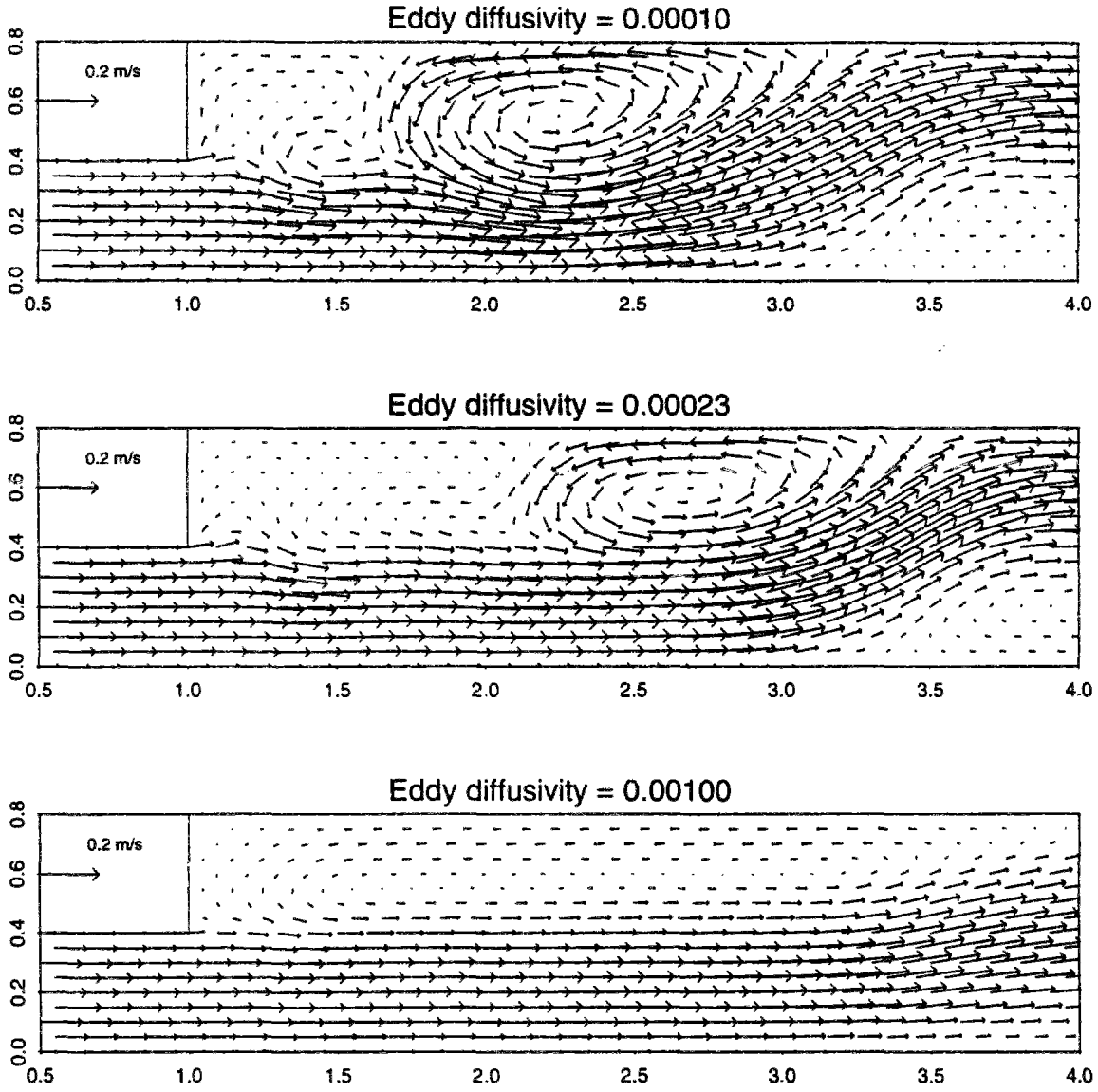


Figure 10.33: Influence of eddy diffusivity on flow patterns computed at $t = 65$ s

10.3.2.3 Conclusion

The model performance for this test case can be deemed satisfying ... but falls far from being excellent !

Numerical resolution appears to proceed fairly smoothly (error indicators are low). In that application involving a complex flow pattern, where velocities vary sharply over the computational domain, it is undoubtedly better to solve the advection step by applying a bicubic interpolator rather than the bilinear one. This allows to reduce numerical diffusion which otherwise would mask the influence of physical parameters of the experiment.

As far as we can judge, the applied open boundary conditions seem appropriate. They allow stable computations. Besides, velocity forecasts in the main flow, which, given the check point location, are most probably heavily influenced by upstream boundary conditions, are in the correct range, as the maximum computed wave height.

The model succeeds in reproducing the development of secondary circulations. However, numerical experiments highlight that the flow features and dynamics, namely the rythm and mechanism of secondary circulation development, their extent, speed and location, are heavily influenced by closed boundary conditions (slip/resistance along side walls) and eddy diffusivity. The available measurements are insufficient to allow a complete understanding of the flow evolution : most notably, the correct mechanism governing the emergence of secondary eddies cannot be identified.

By trial and error, it is possible to guess rough estimates of ad hoc slip parameter α and eddy diffusivity ε. Yet, both assumptions of constant and uniform wall resistance, and constant, isotropic, uniform diffusivity appear to be much too crude. The use of a detailed turbulence model could improve the forecasts. However, considering such things as the sensitivity of forecasts to slip condition, it would be useful to assess first of all model sensitivity to changes in the numerical parameters of the simulation. A local refinement of the computational grid close to the walls could for instance allow a smoother handling of the slip condition, and perhaps alleviate somehow the underestimation of recirculation velocities we observe in the accelerating phase of the flow.

A thorough quantitative comparison of our results with those reported in (Stelling & Wang, 1984) is not possible : models have been run under different conditions; results are not available at the same time interval; as pointed out above, different discretizations and numerical handling of boundary conditions can contribute to the discrepancies between the model outcomes. On the other hand, from a qualitative point of view, these different numerical experiments lead to the same conclusion as regards the role of slip conditions and eddy diffusivity, which is somehow comforting.

10.4 Final assessment of the performance of the proposed models

The tests presented in this chapter and in the previous one have helped us to investigate different aspects of the models proposed in chapter 8 :

1. Applications studied in section 9.2 and 9.3 concern unidirectional steady-state flows. The flow is controlled respectively, in the first case (backwater curve calculation in a sloping channel) by the balance between propagation and friction terms, in the second one (flow over a sill) by propagation and advection.

   (a) Numerical experiments highlight that in such cases, simplifications of the equations governing either the propagation or the advection step, as suggested in (Dan N'Guyen, 1988; Dan N'Guyen, 1993), are not advisable.

   (b) The implementation of a rather complex method to deal with the advection step, namely a characteristic method involving the application of a bicubic interpolator, is justified by the improvement it brings in the second test case. However, the use of simpler methods whenever the advection terms are relatively unimportant should not be neglected as it results in significant computational savings (sec. 9.2).

   (c) When the complete equations are considered, model results appear to be very accurate and compare favorably with other model outcomes (Galland & Hervouet, 1988).

   (d) Not surprisingly, model "QH" based on a "depth/unit discharge" formulation allows a better mass and flow conservation than model "UH" based on a "depth/velocity" formulation, which implies a splitting of the continuity equation.

2. The first application dealt with in this chapter (concentric wave propagation in a closed basin, section 10.1) allows first to test the performance of the algorithm used for the propagation step (relying on approximate factorization into one-dimensional operators, as suggested in (Dan N'Guyen, 1988)) in severe conditions, where the spatial and temporal gradients of flow variables are important. Secondly, it allows to investigate the dependence of model results on time step and implicitation parameters in case of very unsteady flow.

   (a) The factorization algorithm performs fairly well. Shortcomings plaguing classic ADI methods (for instance, spurious polarization) are not observed.

   (b) The resolution of the propagation step can be achieved iteratively, the successive sub-iterations aiming at refining the linearization of flow equations and alleviating the factorization error. This appears to improve slightly the algorithm behaviour for large time steps.

(c) In this case of quickly varying water-depths, the model based on "depth/unit discharge" formulation suffers some troubles. This is due to the fact that in such a model the estimation of advected discharges is plagued with a systematic error proportional to the depth variation during the time step. When this error is corrected, models QH and UH yield identical forecasts.

(d) There is no miracle : model results are dependent on the chosen time step. In order to limit this sensitivity, the implicitation parameter must be set to a value as close as possible to the optimal value $\gamma = 0.5$ which allows the finite difference approximation of flow equations to be second-order accurate (in time).

3. The next test case (tide in a harbour, section 10.2) leads to the following comments :

   (a) The applied boundary conditions, both at the open boundary and along closed ones, seem appropriate. Notably, the perturbation of the flow pattern induced by the salient corner in the harbour is minor.

   (b) Model results are satisfying (development of a recirculation consistent with what physics would suggest, good mass conservation).

   (c) Once again, the approximate determination of advected discharges lead to significant troubles in the "depth/discharge" model, unless it is corrected.

4. The last test case (separating flow in an expanding flume, section 10.3) is probably the most interesting as it deals with a flow where all terms (advection, diffusion, propagation) are important. From a purely numerical point of view, the resolution seems to proceed smoothly. Model results are qualitatively in agreement with the observations but, as in previous interpretation (Stelling & Wang, 1984), quantitative agreement is poor.
   In fact, this test makes us aware of the limitations of some simplifications, namely modelling the effect of turbulence by a simple constant diffusion operator and prescribing the tangential velocity at walls with the help of a uniform slip condition. Besides, it is fairly possible that the real flow has three dimensional features which are not negligible.
   It is suggested that a significant improvement of the model results in this case would require first to improve the model's content from a physical point of view (e.g. inclusion of a full turbulence model), secondly a more thorough investigation about the numerical handling of boundary conditions.

5. The computational cost of the proposed models is dependent mainly on the algorithm chosen to solve the advection step, secondly on the number of sub-iterations applied to solve the propagation step.
   When a characteristic method is applied to solve advection, most of the computational effort is devoted to the backtracking of fluid particles. Its cost is a rising function of local advection Courant numbers.

The number of sub-iterations required in the propagation step is equally dependent on flow conditions and on the threshold of convergence chosen by the user. It tends to increase as local propagation Courant numbers become larger.

In the applications discussed above, the computational cost per time step and per 1000 computational nodes varies between 0.25 and 0.40 s, except when advection is solved with the upwind method (cf sec. 9.2) : then, it falls to 0.08 s.

In conclusion, the performance of the proposed models appear to be satisfying. The outcome of the test cases lead us to contemplate with confidence full-scale applications in natural water bodies.

For the applications involving steady-state or slowly varying flow, the use of a model based on "depth/unit discharge" formulation should probably be preferred, as it allows ensuring a better flow preservation.

For applications involving quickly varying flows or flows where the influence of advection is dominant, this formulation can meet some troubles, as the solution of the advection step is plagued with a systematic error dependent on depth temporal gradients. This error can be corrected but this requires the use of an iterative loop, which raises the computational cost. Thus, in such a case, **the use of model based on "depth/velocity" formulation is more advisable.**

10.5 Résumé français : "Tests sur écoulements instationnaires"

Les tests discutés dans ce chapitre sont des tests de comportement puisqu'ils ne comportent pas de solution analytique. On peut globalement les qualifier de plus sévères que les tests du chapitre précédent. Ils traitent de situations plutôt éloignées des problèmes fluviaux mais ... qui peut le plus, peut le moins.

1. **Le premier problème, concernant la propagation d'ondes concentriques dans un bassin clos, est un bon test de la robustesse de la méthode de factorisation.** On est en effet en présence d'un écoulement trés instationnaire, vraiment multidirectionnel, où la surface libre est trés chahutée et évolue trés vite.
2. **Le second test, au sujet d'un bassin à marée, permet de voir comment se comportent les algorithmes en présence d'un coin saillant**, ce qui est toujours un problème délicat en analyse numérique. Par ailleurs, **c'est le premier cas où la diffusion de moment n'est pas négligée, ... ni négligeable puisqu'elle conditionne fortement les circulations.**
3. Enfin, **le dernier test est l'interprétation d'une expérience de laboratoire sur le développement d'un écoulement dans un canal à élargissement brusque.** En toute rigueur physicienne, ce type d'écoulement ne peut être décrit par un modèle sans représentation adéquate de la turbulence, ce qui n'a pas découragé quelques tentatives antérieures (Stelling & Wang, 1984; Dan N'Guyen, 1993). Cependant, ce problème permet entre autres d'explorer : (i) les limites d'application du modèle bidimensionnel de St-Venant (ii) l'influence des conditions limite à la paroi (iii) l'importance du terme de dispersion de quantité de mouvement (iv) les difficultés de l'expérimentation en relation avec la modélisation. **Il a donc un intérêt "pédagogique" indéniable**, raison pour laquelle nous avons souhaité lui donner un peu de publicité!

Propagation d'ondes dans un bassin clos (section 10.1)

Les conditions du test sont les suivantes (cf 10.1.1) :

- On considère un bassin carré, de côté 21 m, fermé, à fond plat ($z_f = -2.4$ m). Il n'y a pas de frottement (ni de diffusion de la quantité de mouvement).
- La surface libre est initialement déformée puisqu'elle décrit une gaussienne, centrée au milieu du bassin, symétrique, d'écart-type 1 m, de hauteur maximale 2.4 m.
  C'est bien sûr une situation instable. La gaussienne va s'effondrer. Compte tenu des symétries du problème, on devrait observer tout d'abord la propagation d'ondes concentriques. Ensuite, quand ces ondes vont se réfléchir contre les parois, la symétrie radiale va disparaître mais vitesses et hauteurs devraient garder une certaine symétrie vis à vis des diagonales et bissectrices (droites joignant le milieu des côtés) du bassin. Par ailleurs, la masse d'eau contenue dans le bassin est constante. **Ce sont ces propriétés de symétrie et conservation de la masse que nous allons vérifier.**
  On notera que la surface libre présente initialement une forte courbure (-1.2 au centre du bassin) et de fortes pentes (jusqu'à $\simeq 1$ m/m). Ceci constitue a priori une situation délicate pour l'application de la technique de factorisation (cf section 8.4.3).

- Les conditions limite à la paroi sont : (i) gradient de la surface libre à la paroi nul (relation discrétisée au second ordre) (ii) nullité des vitesses normales à la paroi. Les vitesses tangentielles sont laissées libres.
- Les pas d'espace Δx et Δy sont de 30 cm. On pratique des simulations avec un pas de temps Δt de 0.04, 0.08 et 0.12 s. La durée maximale de simulation est 3 s.
- Le comportement des modèles est évalué en étudiant surface libre et champs de vitesse sortis tous les 0.4 s ainsi qu'en enregistrant vitesses et cote à tous les pas de temps en 9 points choisis : le centre du bassin, les points situés à mi-chemin du mur et du centre du bassin sur bissectrices et diagonales. On suit également l'évolution des erreurs moyennes et maximales sur la résolution de l'équation de continuité (voir leur définition en 8.4.3).

On observe les résultats suivants :

1. **Il y a respect des symétries du problème, quelque soit la formulation, le pas de temps ou le paramètre d'implicitation choisis.**
2. **Les résultats sont sensibles au choix du coefficient d'implicitation de l'étape de propagation γ et du pas de temps** (ce dernier aspect ayant été investigué pour la formulation UH "hauteur-vitesse" uniquement).

   Plus le pas de temps est grand et le paramètre d'implicitation éloigné de 0.5, plus l'erreur de troncature attachée au développement semi-implicite des équations de l'étape de propagation est grande. Ceci se traduit par un amortissement des déformations de la surface libre et des vitesses (cf figures 10.4, 10.8 et 10.9).
3. **La conservation de la masse est bonne pour la formulation "hauteur-vitesse" UH, excellente pour la formulation "hauteur-débit" QH.** Pour $\Delta t = 0.04$ on a effectivement, quelque soit γ, une perte inférieure à 0.001 % en fin de simulation avec QH. Pour UH, la perte varie entre 0.39 ($\gamma = 0.6$) et 0.16 % ($\gamma = 1$).

   Quand on augmente le pas de temps, la perte de masse par pas de temps augmente (elle est de 0.004, 0.005 et 0.0054 % respectivement pour $\Delta t = 0.04, 0.08, 0.12$ s et $\gamma = 0.6$) mais ceci est compensé par la diminution du nombre de pas de temps. Ainsi la perte de masse finale pour $\Delta t = 0.12$ varie entre 0.26 et 0.09 % selon γ.
4. **En ce qui concerne la version UH, le passage à une résolution itérative ne semble pas justifié pour $\Delta t = 0.04$. Par contre, il est utile aux pas de temps supérieurs** (cf sec 10.1.2.3). Les mesures d'erreur sont nettement plus faibles (fig 10.7). De plus, l'itération permet de réduire la dépendance des solutions vis à vis du pas de temps.
5. Les solutions fournies par UH et QH ($\Delta t = 0.04$) ont des mesures d'erreur d'une qualité similaire mais sont néammoins trés différentes (section 10.1.2.2). **La source de ces différences est à rechercher dans la résolution de l'étape d'advection en formulation QH.**

Nous avons vu (cf sec 8.3.1.2) que, lors de la résolution de l'étape d'advection en formulation QH, on commet une erreur systématique sur l'évaluation des débits après advection. L'erreur relative est proportionnelle à l'incrément relatif de la hauteur d'eau durant un pas de temps. Nous avons évalué, lors de simulations avec QH, l'ordre de grandeur de cette erreur. Elle s'avère valoir en moyenne 1 à 2 % du débit advecté et au maximum 10 % (en début de simulation).

Nous avons développé une nouvelle version itérative de QH, où l'on évalue, puis corrige explicitement l'erreur de résolution de l'advection. On constate alors que les résultats de cette nouvelle version et de la formulation UH sont presque identiques (cf figure 10.5 et 10.6). La version itérée de QH est 30 % plus chère que la version originale.

En conclusion, ce test démontre **la robustesse de la méthode de factorisation.** Il illustre également **la sensibilité de la formulation "hauteur-débit" à la correction de l'erreur systématique de résolution de l'advection dans un cadre trés instationnaire.** Enfin, les solutions étant trés dépendantes du pas de temps, il souligne **la nécessité de choisir avec soin pas de temps et implicitation pour des écoulements instationnaires.** Nous considérons que cette remarque n'est pas restreinte à nos modèles mais qu'elle concerne plus généralement tout modèle basé sur une discrétisation temporelle au 1^{er} ordre des équations hydrauliques.

Bassin à marée **(section 10.2)**

Nous allons commenter plus brièvement ce test qui conforte les conclusions énoncées à l'issue du précédent.

Le problème est détaillé en section 10.2.1. Un bassin rectangulaire, pratiquement 2 fois plus long que large, à fond plat, est soumis à l'influence de la marée via un étroit canal d'amenée. Le fond est assez lisse (Strickler de 60), la diffusion de quantité de mouvement supposée constante et isotrope (0.1 m^2/s). On doit vérifier : la conservation de la masse d'eau à l'issue d'un cycle de marée, le développement d'une circulation au fur et à mesure que le bassin se remplit puis se vide, l'absence de perturbation des vitesses et hauteurs au voisinage du coin saillant qui marque la jonction du bassin et du canal d'amenée. La marée dure un peu plus de 12 heures, le pas de temps conseillé dans (Galland & Hervouet, 1988) est de 15 s. Compte tenu de l'importance du marnage, les nombres de Courant de propagation atteignent à peu près 10 aux pleines-eaux.

On observe (section 10.2.2) que :

1. **Si l'on ne corrige pas l'étape d'advection, la formulation QH diverge rapidement, le calcul devient instable.**
2. **Les circulations calculées sont cohérentes avec ce que la physique nous permet de supposer** (voir figures 10.13 à 10.15). Le coin saillant les perturbe peu.

3. **La conservation de la masse est satisfaisante.** La perte de masse est de 0.82 % et 0.74 % pour les versions "normale" et itérative de UH respectivement, de 0.69 % pour la version "corrigée" de QH.

4. L'augmentation du paramètre d'implicitation a pour conséquence un léger déphasage des solutions calculées par rapport à celle qui correspond à $\gamma = 0.6$. Ce déphasage entraîne en particulier une moins bonne conservation de la masse (perte doublée). De même la conservation de la masse souffre d'une augmentation du pas de temps. Au delà de $\Delta t = 20$ s, de toutes façons, toutes les versions (UH, QH, itéré ou non) sont instables.

En bref, **ce test permet de conclure encore une fois à l'intérêt de la formulation "hauteur-vitesse" pour des écoulements instationnaires.**

Écoulement dans un canal avec élargissement brusque (section 10.3)

Ce test est basé sur des expériences conduites au Laboratoire de Mécanique des Fluides de Delft par Koppel & Wang en 1981 et 1982. Les résultats sont présentés par (Stelling & Wang, 1984) qui, comme (Dan N'Guyen, 1993) un peu plus tard, tentent une interprétation numérique à l'aide d'un modèle bidimensionnel de type St-Venant. Les conditions expérimentales sont détaillées en section 10.3.1. Leurs caractéristiques majeures sont les suivantes :

- Le canal expérimental consiste un bassin de 5 m de long et 0.8 m de large, précédé d'un canal deux fois moins large de 10 m de long (cf figure 10.16). Le flot à l'aval est régulé par un déversoir en zig-zag. Les fonds sont plats. La rugosité correspond à un Chézy de 63. Au niveau expérimental et numérique on s'intéresse à la zone de 5m de long qui débute 1 m avant l'élargissement.
- A l'amont du canal d'amenée on applique un débit sinusoidal de période 150 s. Après propagation le long du canal, on mesure à l'entrée dans la zone étudiée une vitesse qui correspond à une série de Fourier à 3 composantes, de périodes respectives 150, 75 et 50 s. De même, à l'aval la cote mesurée peut être représentée par la superposition de 3 ondes sinusoidales.

  Les vitesses initiales sont nulles. La vitesse maximale observée est de 0.38 m/s. La hauteur initiale est de 10 cm, elle est ensuite au maximum de 12 cm.
- La diffusivité turbulente a été estimée indirectement en quelques points du canal et quelques instants. Elle semble varier dans une plage de 10^{-4} à 10^{-3}.m^2/s. Pour les besoins des simulations numériques, on supposera la diffusivité constante et isotrope.
- Les vitesses normales à la paroi sont nulles. On utilise pour le calcul des vitesses tangentes une formule paramétrée par un coefficient de glissement α, la valeur 1 correspondant à un glissement total, la valeur 0 à une viscosité totale (i.e vitesses tangentes nulles).
- On adopte des pas d'espace de 5 cm, un pas de temps de 0.125 s.

Les champs de vitesse ont été enregistrés toutes les 10 s, en environ 150 points, localisés principalement dans la zone d'extension du canal. On observe les phénomènes suivants :

- Un tourbillon naît immédiatement au voisinage de l'angle saillant. Durant la phase d'accéleration de l'écoulement (jusqu'à $t = 35$ s), il se développe en une zone de recirculation qui s'élargit avec le temps et dont le centre se déplace vers l'aval (figure 10.19).

- Une très faible recirculation se développe également dans la zone concave du canal, après élargissement.

- Pendant la phase de décélération du flot (45 à 65 s), le tourbillon principal semble se scinder en deux. Le premier tourbillon garde à peu près la même longueur mais s'élargit. Le second croît. La longueur totale de la zone de recirculation (longueur de décollement) continue donc à s'accroître mais moins vite que durant la phase d'accélération. On observe également que l'écoulement devient un peu sinueux.

Les résultats de nos expérimentations sont détaillés en section 10.3.2 (nb : pour ce test, les différences entre versions UH et QH sont négligeables). Elles confortent les observations de nos prédecesseurs, à savoir que **les résultats sont considérablement dépendants du type de conditions limite spécifées à la paroi et de la valeur choisie pour la diffusivité.** Quand ces paramètres sont correctement réglés, **les circulations calculées sont qualitativement proches des circulations observées. Cependant, du point de vue quantitatif** (taille des tourbillons, position de leurs centres) **c'est loin d'être parfait ! Ceci laisse à supposer que les hypothèses physiques sous-jacentes aux équations utilisées pour décrire l'écoulement sont un peu grossières** (l'utilisation d'un modèle de turbulence semblant en particulier souhaitable).

1. **La formation ou non d'un tourbillon secondaire dépend des conditions de frottement et diffusivité.** En particulier, quand on applique des conditions de glissement parfait à la paroi, on prédit le développement d'un seul tourbillon (section 10.3.2.1). Au contraire, dans une situation de viscosité totale, le modèle predit à l'aval du canal l'émergence de tourbillons qui n'ont pas été observés ! La meilleure concordance entre observations et simulations est atteinte pour des conditions de glissement partiel ($\alpha \simeq 0.5$).

   Si la diffusivité est fixée à une valeur trop élévée, on observe cette fois-ci une disparition trop rapide de la circulation secondaire (section 10.3.2.2).

2. Dans les gammes de valeur où l'on réussit à reproduire l'existence de deux tourbillons en phase de décélération du flot, nous avons, en étudiant les résulats de calcul toutes les secondes (soit à un intervalle 10 fois plus petit que celui entre observations), constaté que **le modèle prédit le développement successif de plusieurs circulations secondaires. Le mécanisme de formation de ces circulations dépend des conditions de frottement et diffusivité.**

- Dans la plupart des cas, les tourbillons secondaires semblent se développer au voisinage de l'angle saillant. Ils croissent en décollant de cet angle et vont fusionner avec la circulation primaire ... pendant qu'un nouveau tourbillon est en formation.
- Il n'y a que pour des diffusivités élévées (fixée à la valeur maximale observée) que l'on observe formation du second tourbillon par séparation en 2 du premier tourbillon (section 10.3.2.2).

En l'absence d'observations plus rapprochées il nous est impossible de savoir quel est le vrai mécanisme de formation des tourbillons !

3. **Frottement et diffusivité contrôlent le rythme des circulations secondaires**, à savoir le **début des cycles** de génération des tourbillons secondaires et leur **période**, ainsi que la **force des tourbillons** (i.e. importance des vitesses dans la zone de recirculation) et leur **extension.**

   En gros, plus glissement ou diffusivité sont importants, plus tard apparaissent les circulations secondaires et plus elles sont faibles . (nb : elles n'apparaissent d'ailleurs pas en cas de glissement parfait !).

Conclusions finales des tests (section 10.4)

Les tests présentés dans ce chapitre et le précédent nous ont permis d'explorer sous différents angles le comportement des différentes méthodes de résolution proposées au chapitre 8 :

1. Les tests étudiés en 9.2 et 9.3 concernent des écoulements uni-directionnels permanents. L'écoulement est contrôlé par l'équilibre entre propagation et frottement dans le premier cas, entre propagation et advection dans le second cas.

   - Les deux problèmes permettent de conclure aux dangers des simplifications du traitement de l'étape d'advection ou de propagation suggérées dans (Dan N'Guyen, 1988; Dan N'Guyen, 1993).
   - Quand on utilise les versions "complètes" des formulations "hauteur-vitesse" ou "hauteur-débit", les résultats sont trés bons, dans l'absolu et également vis à vis d'autres modèles (TELEMAC, (Galland & Hervouet, 1988)).
   - L'implantation d'un algorithme sophistiqué pour résoudre l'advection (méthode aux caractéristiques et interpolation bicubique) est justifiée par les améliorations apportées dans le second problème (passage du seuil). Toutefois, au vu des résultats du premier test (calcul d'une courbe de remous), on ne doit pas se priver d'implanter et utiliser une méthode plus simple, comme UPWIND (nettement plus économique !), dans les cas où les termes inertiels ont peu de poids.

- Comme on pouvait s'y attendre, la formulation "hauteur-débit" permet une meilleure conservation de la masse que la version "hauteur-vitesse", qui implique un éclatement de l'équation de continuité.

2. Les tests instationnaires du présent chapitre complètent cette vision des modèles.

   - Le premier (propagation d'ondes concentriques) met en évidence la robustesse de la méthode de factorisation et "revalorise" la formulation "hauteur-vitesse". En effet, dans des situations où la surface libre varie rapidement, une correction doit être apportée à l'étape d'advection dans la formulation "hauteur-débit". A défaut, les résultats se dégradent (à l'extrême, la simulation devient instable, comme dans le cas de la darse).
   - Les tests de la darse (bassin à marée) et du canal avec élargissement brusque montrent que les modèles proposés savent s'accommoder de singularités géométriques comme un angle saillant. Le mode de traitement des conditions aux frontières ouvertes apparaît également satisfaisant.
   - La résolution de l'étape de propagation au prix d'un cycle itératif apparaît susceptible d'apporter des améliorations dans des cas trés instationnaires (ondes concentriques). Elle semble en particulier réduire la sensibilité des calculs au pas de temps et au mode d'implicitation choisi.
   - Enfin le dernier test (élargissement brusque), bien que qualitativement acceptable, nous ramène les pieds sur terre. Il ne suffit pas d'avoir de bons algorithmes numériques, encore faut-il que les équations résolues reflètent bien la physique du problème !

3. Enfin, les coûts calcul semblent raisonnables. Il dépendent du mode de résolution de l'étape d'advection (et du pas de temps si l'on a recours à une méthode aux caractéristiques) et du choix de traiter ou non la propagation de façon itérative. Pour les tests étudiés, le temps CPU consommé pour 1000 noeuds de calcul en 1 pas de temps varie entre 0.25 et 0.4 s, sauf dans le cas où l'advection est résolue par la méthode explicite UPWIND : le coût est alors réduit à 0.08 s.

En conclusion, **les résultats des tests sont satisfaisants et nous permettent d'envisager avec confiance le passage à des applications grandeur nature. Pour des écoulements permanents ou lentement variables on préféra la formulation "hauteur-débit**, plus conservative. **Pour des écoulements fortement instationnaires on préfera la formulation "hauteur-vitesse"** si l'on veut s'éviter le souci de "corriger" l'advection telle qu'elle est traitée en formulation "hauteur-débit".

Chapter 11

Application to the Seine River

As mentioned in the very first chapter of this dissertation, the development of a surface water flow and transport model has been undertaken bearing in mind it would be applied in the frame of a research program, the Piren-Seine, devoted to the study of the Seine River and its water quality problems. More precisely, two-dimensional depth-averaged modelling is intended to provide understanding of the near and intermediate field mixing of point sources. This analysis should help interpreting measurements and correcting forecasts of one-dimensional water quality models, which appear to be the only feasible choice as far as planning and management tools are concerned (the full extent of the river area to be modelled corresponds broadly to its navigable part, from Montereau, approximately 100 km upstream Paris, to Poses, 200 km downstream of Paris, at the beginning of the estuary).

The present chapter provides an example of the model application to a Seine River reach. Section 11.1 introduces us summarily to the studied area while section 11.2 details the experiment we have been working on. The different steps of its hydraulic interpretation form the core of sections 11.3 to 11.5.

11.1 Main features of the studied area

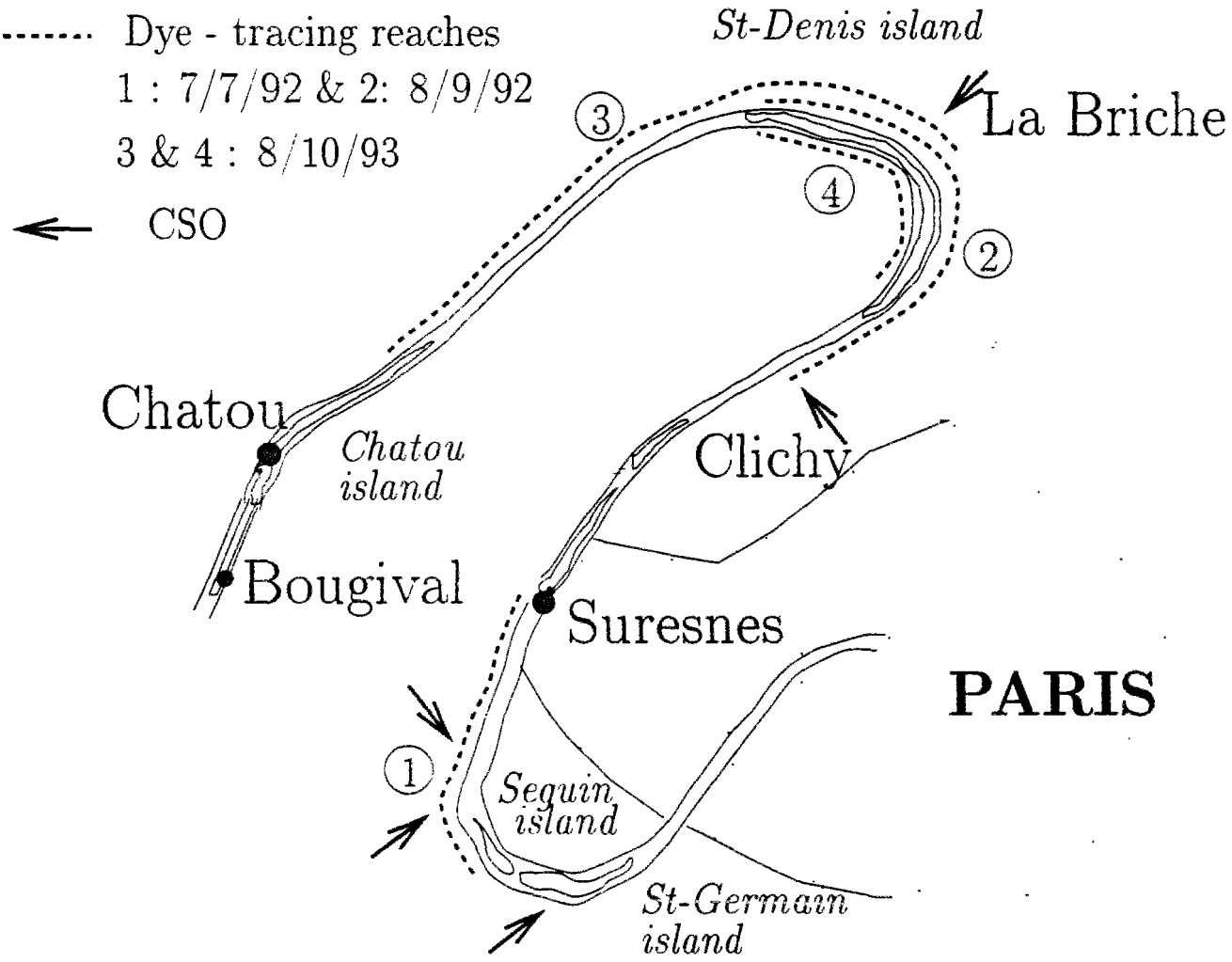


Figure 11.1: General map of the studied area

The CSOs problem The Seine River is a relatively small river heavily impacted by the intense anthropogenic activities in its catchment. Notably, it flows through Paris and its region, with approximately 10 million inhabitants. This urban concentration led to the development of large wastewater treatment facilities (e.g. the Achères plant, which treats 2.6 million cubic meters a day), fed by an extensive sewer network, most of whose parts were built many years ago (a century ago for the Paris inner-city) regardless of possible consequences of sewer overflows. Indeed, the most important pollutant point sources in the Seine River appear to consist of

sewer "products", whether they are "controlled" like sewage treatment plant effluents or occur randomly as storm overflows. For instance, the Achères discharge into the Seine is about 27 m^3/s which, even treated, perturbs the river ecosystem since it amounts to more than 25 % of typical low flow rates (Billen & Allardi, 1993). Similarly, the sum of storm overflows in Paris area may temporarily equal the river discharge (Simon *et al.*, 1994).

The area most perturbed by CSOs appears to be a 40-km reach of the Seine River between Paris (St-Germain island) and Chatou lock (cf figure 11.1). The two largest zones of CSO discharges from the Paris county are located there, at Clichy and La Briche. Phenomena linked to CSOs, like oxygen depletion and fishkills, are observed near Chatou, relatively far downstream ($\simeq$ 15 km) from the main CSOs outlets (Mouchel & Simon, 1993). In the frame of the Piren-Seine program, consequences of CSOs have been thoroughly surveyed, first by operating continuous monitoring stations of various quality parameters (oxygen, temperature, nitrogen, phosphorus ...) at the Suresnes and Chatou dams, secondly by following and sampling along the flow polluted clouds issued by the CSOs. However, the multiplicity of CSOs, their possible interaction and their persistent transverse heterogeneity (mixing length seems to be 5 to 10 km long, according to dye-tracing experiment achieved at the beginning of the eighties) contribute to make the interpretation of local measures a delicate task. As explained in chapter 2, this was the motivation for undertaking the development of a two-dimensional hydraulic and transport model : a tool able to represent faithfully dilution processes should allow a proper assessment and interpretation of other, biogeochemical, processes.

Hydraulic characteristics In low flows, the flow in the Seine is regulated by reservoirs in order to allow for such water uses as water supply and navigation. In Paris, summer discharges vary approximately between 80 and 120 m^3/s. CSOs induce very important transient perturbations to this flow regime (the most important are generally observed in June and September).

Fluvial navigation has two main consequences on the river :

1. The Seine river has been divided in reaches 20 to 50 km long, separated by navigation dams which regulate water depths : the objective is to ensure a sufficient water depth of about four meters. Consequently, during low flows, the river behaves more as a succession of quiet pools pouring into each other than as a natural flowing water body.
2. Because of regular dredging applied to maintain the wide navigation channel, the Seine bed is almost flat (slope between $5\,10^{-5}$ and 10^{-4}) and its cross sections are fairly regular (trapezoidal shape).

In the studied area, there are three locks : the upstream one is at Suresnes, the other ones located in the right and left arm off Chatou island (Chatou and Bougival dams, respectively 27.5 and 32 km downstream Suresnes). The average water depth in the reach varies between 3.8

and 6.5 m (the deepest parts are found upstream from dams). The river width varies between 140 to 220 m in reaches without islands and from 60 to 120 m, in each arm, when divided by an island.

When there are navigation dams on two adjacent arms, the balance of flow between each arm is completely controlled by the dams operation : such is the case for instance around the Puteaux island (immediately downstream Suresnes lock) and the Chatou island.
Otherwise, flow distribution should be dependent on each arm features (slope, bathymetry, bottom and bank roughness). In this second case, the flow divide has been seldom surveyed and is generally poorly known.
On the whole, it has been observed (even if often qualitatively) that there exist great variations in hydraulic conditions in the various arms (e.g. the residence times). These differences further increase when CSOs occur, because CSOs may influence only one arm of the river (for instance, La Briche overflows affect only the arm to the right of St-Denis island).

The dams, by heightening artificially the free surface elevation, keeping it nearly uniform (the water slope between successive dams is usually less than 10^{-6}), contribute to make the Seine River a slow flowing water course : the average velocity during low flows and dry weather conditions is about 0.1 m/s (regardless of reach differences). Besides, due to the regularity of cross sections, the velocity is fairly uniform across sections located *in straight stretches of the river.*

Specific experiments In order to appraise the mixing characteristics of the river, several dye-tracing experiments have been conducted in the frame of the Piren-Seine (cf fig 11.1) :

1. The first one took place in a 6 km long reach upstream from the Suresnes lock and focused on evaluating the representativeness of a continuous monitoring station implemented in the middle of Suresnes dam. Indeed, several significant CSO outlets are located 5 to 6 km upstream the dam, on the left bank, and could possibly influence the monitoring point. Average flow during this experiment was 110 m^3/s.
2. The second one was designed to evaluate the impact of an overflow from Clichy, where the dye was injected. The dye cloud was surveyed till the downstream end of the St-Denis island, namely $\simeq$ 8 km downstream the injection. The average flow was about 170 m^3/s which is significantly more than typical low flow discharges but fairly close to flows that can be observed during major CSOs events. The right arm conveyed 60 % of the total flow and the whole dye cloud appeared to flow through it. *This is the experiment we shall investigate more thoroughly.*
3. In the third experiment dye was injected slightly upstream of the Briche outlet. The objective was to simulate the fate of an effluent from La Briche and to survey the mixing

downstream of the island of the right arm and left arm flows : during CSOs the right arm should be more heavily polluted than the left arm, so that their junction could induce some reduction of pollutant concentrations. This experiment was perturbed by a sudden flood occurrence, with discharges achieving values superior to 300 m^3/s. This notably delayed the completion of vertical mixing of the dye.

4. The fourth experiment was intended to provide some hydraulic information about the left arm off St-Denis island. As it is not open to barge navigation, this stretch is poorly known (no soundings available). Dye was injected at the upstream end of the island and its transit time till the downstream end was surveyed.

The third and fourth experiments took place on the same day. Gauging operations were achieved during all experiments except the first one. On the other hand, aerial video surveys of the tracer cloud were made during the first (Suresnes) experiment.

The sampling design for the experiments, except the last one, was similar : the rhodamine B cloud was surveyed at different cross sections, spaced between 1 and 3 km apart. Four to eight points were monitored in these cross-sections.

- In fact, the monitoring sections correspond to bridges, where it was easy to locate precisely the sampling points. These various bridges have several passes. Measurement points have been set so that they are spread evenly across each pass. The number of points in a pass depends of its width. Usually the passes next to the banks are narrower, which explains why, generally, only one measurement was located in them. As the river width and the bridges geometry vary, the distance of monitoring points from the river banks was not constant.
- The total number of samplings performed across a section was also limited by practical considerations. Samplings were performed either from a boat, either from the bridges, which stand some twenty to thirty meters above the water. In the last case, we used special "fishing rods". Manipulation of the fishing rods or sampling from the boat (sometimes interrupted or delayed by the passing of a barge), displacement between monitoring points, numbering and exchange of sampling bottles required approximately fifteen minutes with 5 or 6 points across a section. At each section, we tried to find some balance between a good spatial description (as many points as possible) and a good temporal description (a high sampling frequency), so that 6 measurement points per section was about the maximum number we could manage.
- Dye was injected in the upper 50 cm of the water column. Consequently, at the section closest to the injection, sampling was done at two depths, at the surface and 3 meters deep, in order to check the uniformity of tracer concentrations along the vertical.

The first stages of interpretation of these experiments stress the following points :

- For discharges typical of low flow conditions (first and second experiment), vertical homogeneity is achieved in the first kilometer downstream of the injection. Trials performed with a simplified dispersion model led to estimate that the vertical diffusivity ranged between 10^{-3} and $1.4\,10^{-3}$ m^2/s in the first experiment and from $1.4\,10^{-3}$ to $1.8\,10^{-3}$ m^2/s in the second one (Simon, 1992).
  It is worth noting that, in the case of CSOs, the vertical mixing should be achieved much sooner, due to the configuration of weirs and the magnitude of CSOs discharges.
- The distance required for cross-sectional mixing of bank-released tracers in the Seine River in the Paris area seems to be approximately five kilometers. These findings are in agreement with previous experiments (Bujon, 1983). This mixing length might even be shorter for overflows which possess an initial momentum and direction when entering the river : this should accelerate both their vertical and transverse mixing.
- Besides initial overflow discharge conditions, advective phenomena (namely differential advection and possible transverse currents in bends) appear to play a dominant part in the development of transverse mixing, as documented by the in-situ aerial surveys. This underlines that a detailed hydraulic modelling is needed, should we want to estimate accurately the longitudinal and transverse diffusivities.

11.2 Clichy dye-tracing experiment (8/9/92)

11.2.1 Experimental setting

Figure 11.2 gives an outline of the sampling design.

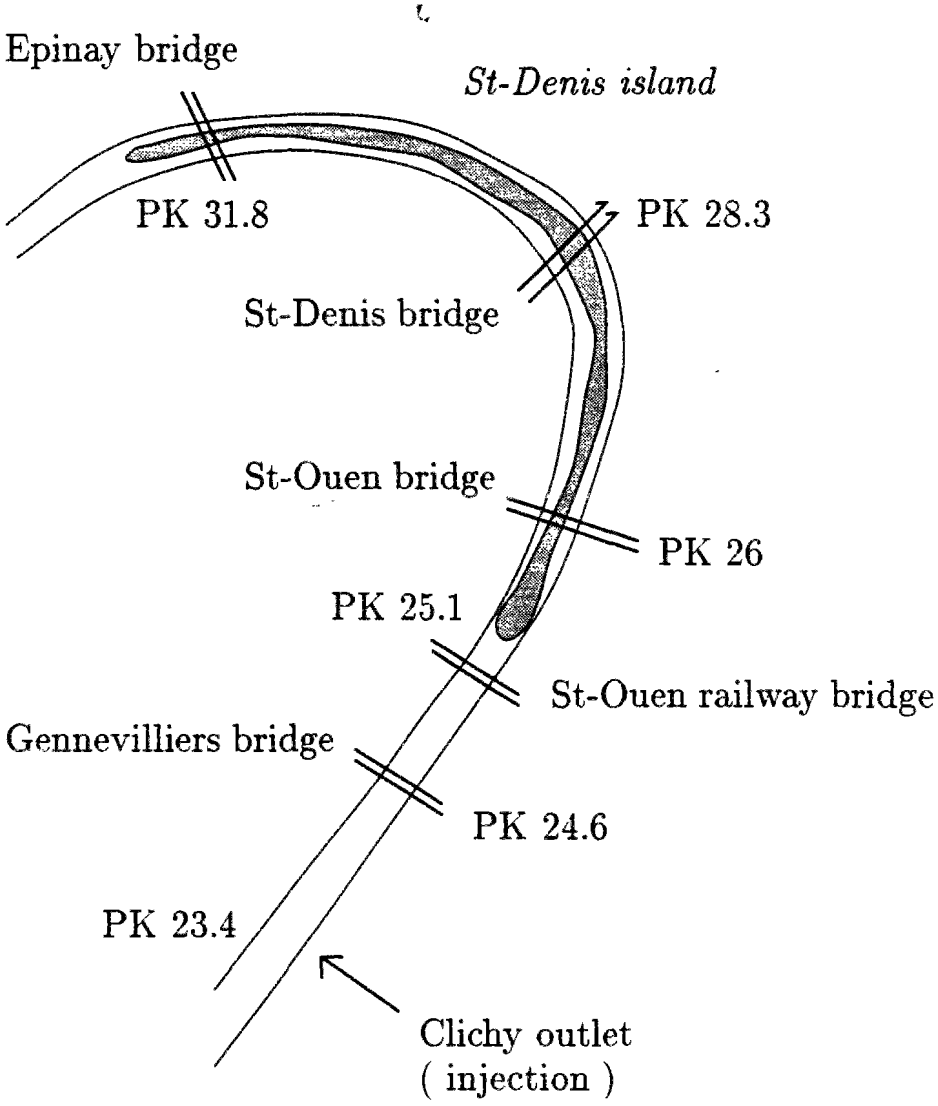


Figure 11.2: Scheme of measuring cross-sections

The 15 kilograms of tracer (rhodamine B) were injected at a constant rate next to the Clichy CS0 outlet, on the Seine River right bank, between 3h10 and 4h10 a.m. The exact locations of the monitoring sections are given in table 11.1. Symbol PK refers to the system of reference points (Kilometric Points) used by the Service de Navigation de la Seine (SNS). For instance, the St-Denis bridge is at PK 28.270, that is to say 270 m downstream of the reference point at PK 28. The distance between consecutive reference points is not always strictly equal to 1 km.

The distance between sections has been estimated along the bank. It may be slightly different when considering the central axis of the river (up to 10 or 20 m apart).

Table 11.1: Location of monitoring sections. Reach Clichy - Epinay

Section	Kilometric Coordinate (PK)	Distance from previous section (m)	Distance from injection (m)
Clichy outlet	23.380		0
Gennevilliers bridge	24.605	1215	1215
St-Ouen railway bridge	25.100	520	1735
St-Denis island upstream end	25.460		2095
Right arm			
St-Ouen bridge	26.050	935	2670
St-Denis bridge	28.275	2150	4820
Epinay bridge	31.735	3260	8080
St-Denis island downstream end	32.830		9135

Table 11.2 details the repartition of monitoring points across the surveyed sections. As regards the first and second section (Gennevilliers and St-Ouen railway bridge), we didn't mention other stations located in the left half of the river, as no rhodamine was observed there. All samplings were surface samplings except under Gennevilliers bridge : there, at the two stations closest to the right bank, rhodamine was surveyed also 3 m below the surface.
No rhodamine flowed through the arm left to the St-Denis island.

Table 11.2: Location of measuring points

Section	River width (m)	Distance of measuring points from the right bank (m)	Sampling frequency (mn)
Gennevilliers	150	16, 33, 60	15
St-Ouen (railway)	170	10, 35.7, 50.2, 79.8, 94	15-20
St-Ouen	114	11.2, 41.5, 74.7, 101.7	10-15
St-Denis	101.5	11.2, 34.4, 47.1, 57.9, 82.5	10-20
Epinay	114	13, 36, 55.5, 75, 98.5	15-20

11.2.2 Flow conditions

Just upstream of Paris, the Seine joins with a major tributary, the Marne River. The flow during the experiment was surveyed first with the help of two automatic gauging stations located at Alfortville (Seine) and Noisiel (Marne), respectively 30 and 60 km usptream of the Clichy injection point. There, discharges estimates are available every two hours, with a 5 % error margin for this range of flow rates. As detailed in (Simon, 1992), the flow in the experimental reach can be estimated by summing the Alfortville and Noisiel discharges, taking into account the propagation delay between the stations and the reach, then subtracting the amount of water extracted by potable water plants ($\simeq$ 6 m^3/s).
According to this data, the flow inside Paris stayed in the range 167 to 178 m^3/s during the experiment. The flow at the Suresnes dam should be a bit lower ($\simeq -1.2$ m^3/s) because of another water intake located slightly upstream of the dam. This discharge level is significantly higher than observed during summer 1992. This is due to a sudden raise in the Marne discharge, caused by drastic operations on one of the navigation locks : the Marne flow rate stayed above 100 m^3/s from September 7, 8 p.m., to September 8, 6 p.m., while it reached only 56 m^3/s at 8 a.m. Sept. 7 and had fallen down to 46 m^3/s at 8 a.m. Sept. 9.

Besides, during the dye-tracing experiment, the flow was gauged in each arm around the St-Denis island. Because of boat traffic, the measurements lasted the whole morning. The total flow estimate is 168 m^3 s. The right arm conveyed 60 % of the flow (i.e. 100 m^3/s). Estimates from the gauging operation and monitoring stations are close (less than 6 % apart), which validates both kind of data.

The flow within the experimental reach is also dependent on the regulation of Suresnes dam ($\simeq$ 6 km upstream Clichy), Chatou and Bougival dams (respectively $\simeq$ 12.5 and 17 km downstream Epinay bridge). The regulation of Suresnes and Chatou dams was modified twice during the tracing experiment, with the following consequences on free surface elevations :
(nb : all elevations in this chapter are given with respect to a french coordinate system, the NGF, Nivellement Général de la France orthométrique)

	Suresnes lock	
Hour	**Upstream elevation**	**Downstream elevation**
3h 00	26.58 m	
5h 00	Modification	
6h 30	Modification	
7h 00	26.49 m	23.75 m
17h 00	26.43 m	

Hour	Chatou lock Upstream elevation
8h 00	23.63 m
9h 00	Modification
11h 00	Modification
12h 00	23.56 m
17h 00	23.48 m

There was no operation at Bougival, which is regulated (with a bottom sluice gate) during the low flows so that the discharge in the left arm off Chatou island remains approximately constant, at the level of 60 m^3/s.

11.2.3 Observed pollutographs

The water samples were analyzed with the help of a fluorimeter. The detailed decription of the device and of its calibration method is given in (Simon, 1992) (appendix B). For this dye-tracing experiment, rhodamine concentrations are estimated with a precision of $\pm 0.1\,\mu g/l$ as regards low values, $\pm 0.3\,\mu g/l$ for high values. Figures 11.3 to 11.5 display the resulting pollutographs.

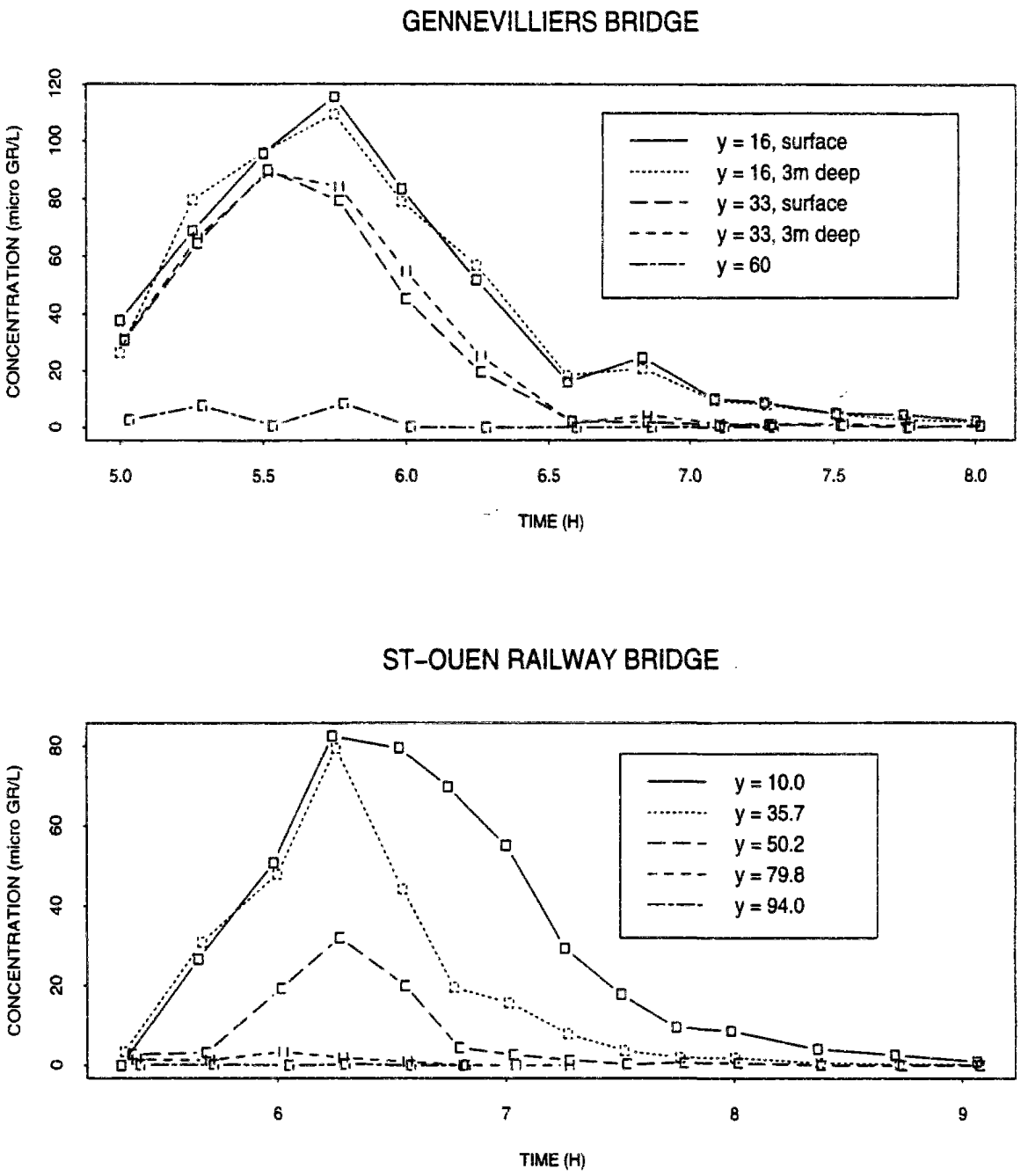


Figure 11.3: Rhodamine concentrations at Gennevilliers and St-Ouen (railway) bridges

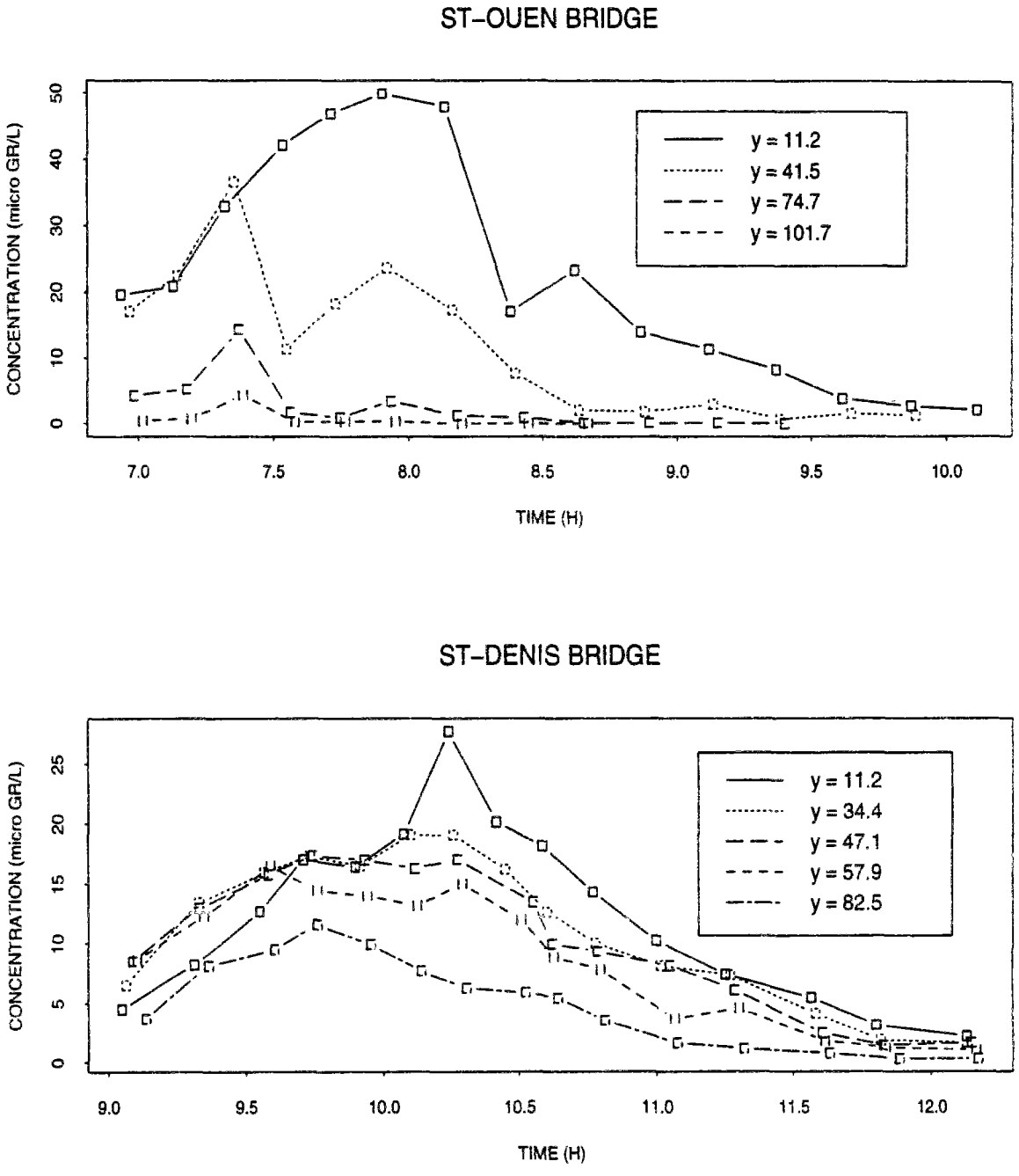


Figure 11.4: Rhodamine concentrations at St-Ouen (road) and St-Denis bridges

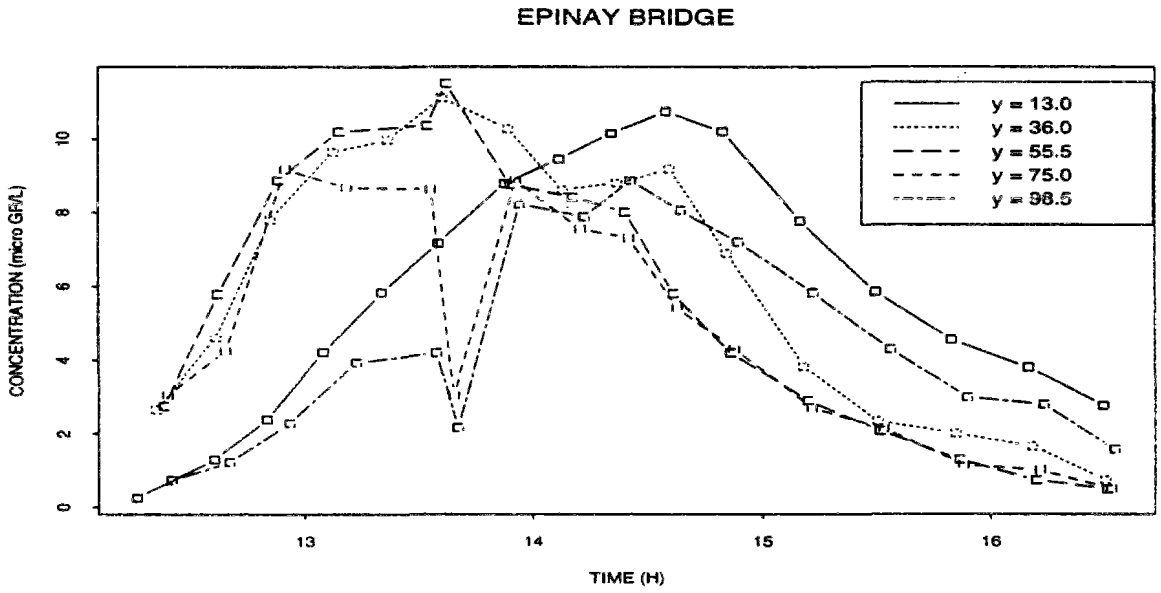


Figure 11.5: Rhodamine concentrations at Epinay bridge

- The high flow rate was unexpected for two reasons. First, there was some trouble with the automatic gauging stations on the eve of the experiment. Secondly, once the stations were back in function, there was no access to their data during the night. The sampling operations had been scheduled based on lower flow estimates. Consequently, the sampling teams were somewhat surprised by the dye cloud arrival, notably at the first stations, where the night, or the dim light of the morning, didn't allow to detect easily the first traces of dye and correct the sampling plan. This explains why the dye cloud beginning has not always been recorded.
- At Gennevilliers bridge, the rhodamine has spread only across the right third of the cross-section. On the other hand, the tracer appears to be evenly distributed in the water column.

  After the passing of most of the rhodamine cloud, a colored trickle about 2m wide was observed to linger along the right bank, probably because of some trapping induced by bank irregularities or vegetation. At 8 a.m., concentrations there were still notable ($\simeq 12\,\mu g/l$). Yet, we focused on the main tracer cloud dynamics and didn't survey any longer this tail of dye.
- At St-Ouen railway bridge, a few hundred meters upstream of the St-Denis island, as at St-Ouen road bridge, transverse mixing has not developed much. At the last section, the pollutographs display two local maxima, for some unknown reason.
- 2 km downstream, at St-Denis bridge, transverse homogeneity is well advanced. The ratio of rhodamine peaks on the right and left side of the river has fallen to 2.4, whereas it was more than eleven at the preceding section. This acceleration of transverse mixing may probably be ascribed to the curving of the river.
- At the most downstream section (Epinay bridge), the transverse repartition of rhodamine appears to be no longer dependent on the bank of injection, but mostly controlled by the differential advection across the section. For instance, the pollutographs near both the left and right banks exhibit some delay with respect to central ones, due to the fact that velocities in the middle of the river are higher than along the banks. Rhodamine maxima are in the same range (9-11 $\mu g\ l$) at each sampling point.

  However, pollutographs at the two stations closest to the right bank display an anomaly, consisting of a sharp decrease at $t \simeq 13h\ 40$. The fact that this is observed at two different locations seems to indicate that it is not simply an error measure. We have no explanation to suggest for this anomaly, except perhaps that a sudden side discharge might have caused this perturbation (there are several sewer outlets along the river bank in that area ... but the weather was fair).

In spite of the above mentioned anomalies, the data appear on the whole to be of good quality.

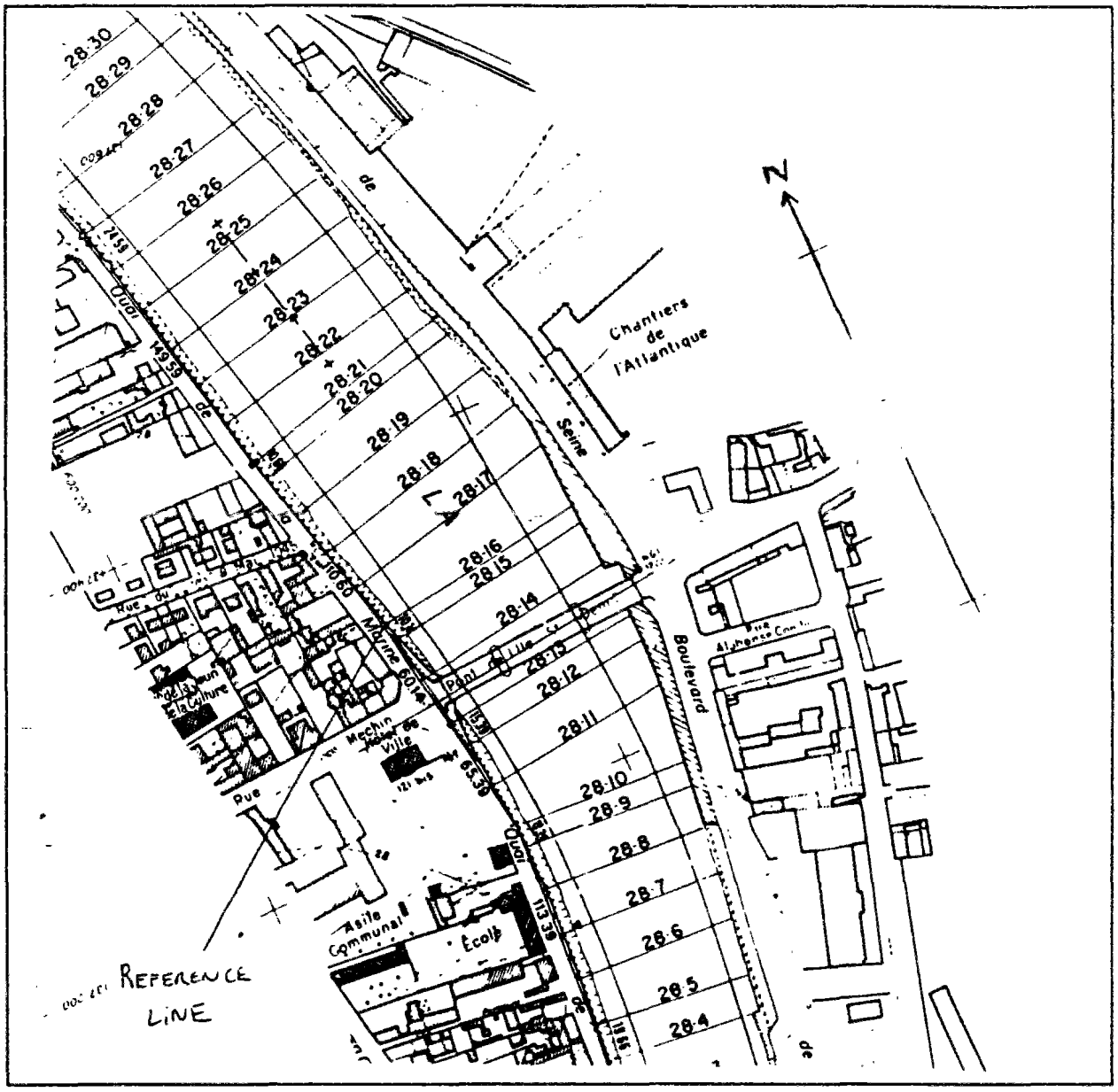


Figure 11.6: Plan of transverse profiles of the Seine River (St-Denis bridge surroundings)

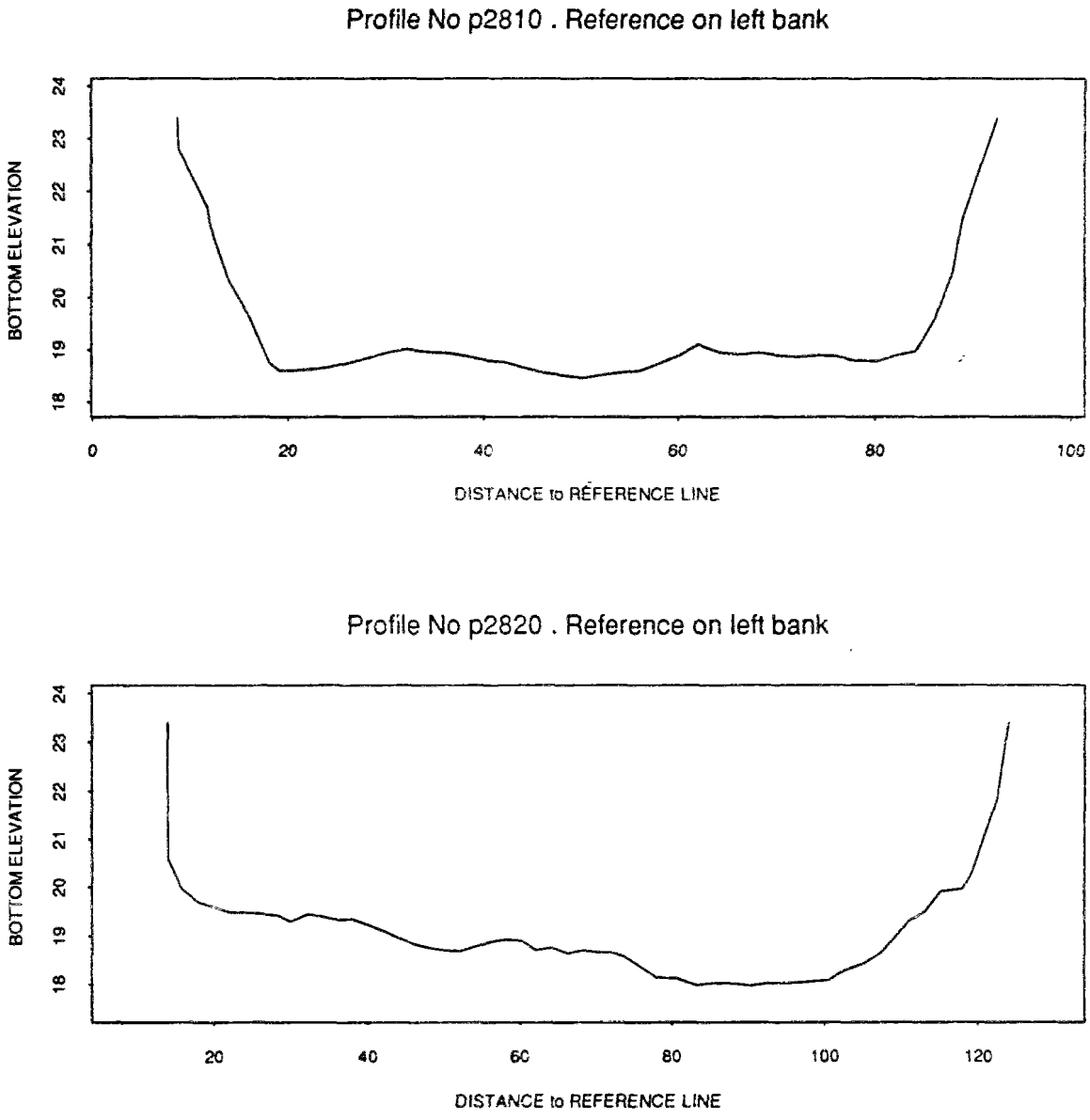


Figure 11.7: Examples of cross sections of the Seine River

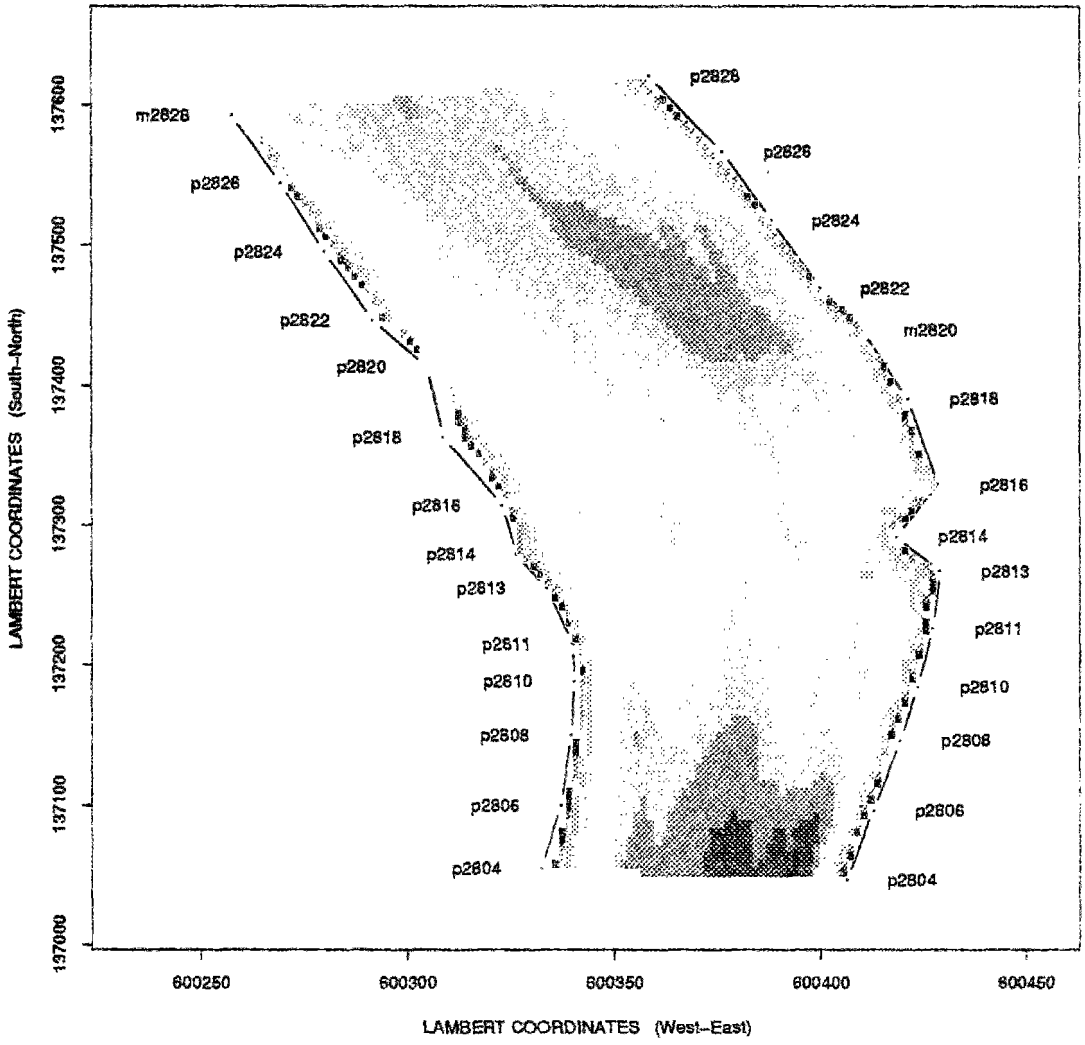


Figure 11.8: Example of Seine River bathymetry

11.2.4 Bathymetric data

A hydraulic model can be run accurately on some stretches of river only if supported with appropriate bathymetric data. While 1-D models are used to yield average estimates of flow velocity, 2-D models are supposed to produce a detailed description of flow patterns. This can be achieved only if the bathymetric description is very precise.

Table 11.3: Bathymetric data available from Clichy to Chatou and Bougival navigation locks

Area	date of soundings	soundings frequency
Upstream St-Denis island		
PK 22.0 to 23.2 & PK 24.0 to 25.1	1974	every 25 m
PK 23.2 to 23.9	NO DATA	
PK 25.1 to 25.5	1989	every 25 m
Right arm off St-Denis island		
PK 26.4 to 27.0 & PK 28.7 to 28.9	1974	every 25 m
PK 27.5 to 28.0	NO DATA	
other zones	1988/89	every 25 or 50 m
Left arm off St-Denis island		
NO DATA		
Between St-Denis and Chatou islands		
PK 33.0 to 34.1	1974	every 25 m
PK 34.1 to 35.4	1992	9 cross sections over 1300 m
PK 35.4 to 39.5	1963/64	every 25 m
PK 39.5, 39.9, 40.2	1989	3 cross sections
Right arm off Chatou island		
PK 40.2 to 45.3	1989	$\simeq$ every 500 m
PK 44.3 to 44.45	1992	every 25 m
Left arm off Chatou island		
PK 40.0 to 43.3	1990	every 50 m
PK 43.5 to 45.0	1989	every 500 m

The Seine river bathymetry is surveyed by the SNS. Systematic soundings of the Seine River downstream Paris were achieved in the sixties (one cross-section every 25 m). Since then, there

was no general updating of these data but some specific stretches of the river have been sounded more recently, generally when some public works affecting the river banks or bed - like bridge construction or lock maintenance - was planned in that area. On the other hand, some data eventually got lost ! Table 11.3 provides a summary of data available between the Clichy outlet and the Bougival and Chatou locks (for more details, see (Simon, 1992), appendix C). Data collected before 1988 are mere drawings of the cross sections and must be digitized. Each cross section is described by 20 to 60 points.

Soundings are located with respect to a reference line located alternatively on the left or right bank of the river (the "polygonale") (cf figure 11.6). The bed elevation is recorded according to the distance to this reference line (cf figure 11.7). In order to exploit the data, it is necessary to express them in a common coordinates sytem, the Lambert coordinates of whose x axis and y axis are respectively parallel to the West-East and South-North direction (cf figure 11.8).

The following problems are encountered :

1. Data are lacking for a few stretches (particularly the arm left to St-Denis island, not surveyed as it is not open to barge navigation).
2. Some of the soundings are dated (20 to 30 years old).
3. Some of the recent data have been collected sparsely (one to several hundred meters between each cross section).
4. Finally, some transverse profiles are incomplete : only the part of the Seine covered by the navigation channel has been surveyed.

We have adressed these different issues as follows :

Old soundings In several reaches of the river, soundings taken in different years were available. Their comparison shows that the temporal evolution of the Seine bed is quite moderate except for singular points, immediately upstream of bridge piers or channel division, where the sediment deposition is fast (0.2 m/year). Therefore, even old soundings data may generally be trusted.

Missing or sparse data Missing cross sections have been obtained through interpolation, preserving the river width that can be deduced from aerial plans (scale of 1/2000). The relevance of this reconstitution has been checked with the help of one-dimensional hydraulic modelling, comparing computed to observed transit times in several situations.
(nb : examples of such reconstitution are provided in appendix H.1.2)

Incomplete transverse profiles The probable shape of the cross section is deduced from the shape of the closest complete profiles (or from the oldest soundings of this same cross-section) and from other kinds of data. For instance, aerial plans of the Seine channel also include a description of its immediate surroundings (like cadastral surveys). These informations, eventually checked with the help of field visits, allow us to guess the type of

the river bank (whether it is a natural, gently sloping bank, or a vertical embankment). Once the most probable shape has been defined, the profile is completed trying to keep consistent with this general shape.

On the whole, the Seine cross-sections are fairly regular and smooth. They are slightly more irregular in the vicinity of hydraulic singularities such as bridges. There, it can be observed that the piers tend to induce upstream deposition and downstream scouring. This is illustrated for instance on figure 11.8 which displays Seine bed levels around St-Denis bridge, which passes between profiles P28.13 and P28.14 (deepest parts correspond to dark green areas, shallowest to red and purple ones). However, these perturbations have limited extent.
Finally, our aim is not to achieve a detailed study of the flow patterns next to the bridges but to forecast mixing on a reach several kilometers long. Consequently, we have not taken into account bridges piers in our representation of the Seine river.

In summary, the bathymetry of the Clichy-Chatou reach was described using 380 selected cross sections (approximately one every 50 m).

11.3 First stage of hydraulic interpretation

Obtaining an acceptable agreement between measures and forecasts will require adjusting several parameters of our transport and flow model, namely the Strickler coefficient (K_s) which conditions the magnitude of bed resistance to the flow and the diffusivities controlling transport of mass and momentum by horizontal dispersion. Mass and momentum diffusivities are generally assumed to be proportional, linked by the Prandtl constant (cf 2.2.2 & 2.4.5) whose usual range in fluvial situations is [0.5, 1].

In order to delimit the probable range of Strickler values, it is useful to resort to a one-dimensional St-Venant hydraulic model. In most fluvial situations, characterized by low Froude numbers, the dominant forces are propagation and friction (cf 8.2) and their balance defines the free surface slope for given flow rate and bathymetry. Backwater curve calculations allow the calibration of the Strickler coefficient in a 1D model. This value is generally close to the appropriate value to be used in 2D simulations. We have thus decided to apply the PROSE model developed in the frame of the PIREN-Seine program (Even & Poulin, 1993).

Besides the calibration of Strickler coefficient, the use of PROSE has other benefits :

- Once PROSE is calibrated, it can be applied to compute average transit times in the reach. Comparing them to transit times observed for the rhodamine cloud allows us to check for instance the adequacy of our reconstitution of missing bathymetric data (e.g the 700 m

stretch around Clichy overflow, the 500 m stretch unknown in the right arm of St-Denis island, cf table 11.3).

- Secondly, running our 2D model requires us to prescribe consistent upstream and downstream hydraulic conditions. Our knowledge of water levels within the reach is poor, as mentioned in section 11.2.2 : water levels are known only at the Suresnes and Chatou dams, and only at three different times of the day. It would be cumbersome to model with the two-dimensional tool the whole reach between the injection point at Clichy and the Chatou and Bougival dams when we are in fact interested only in what occurs within a 15 kilometers shorter reach located between Clichy and our last monitoring section at Epinay bridge. The application of PROSE can provide us with water level forecasts at the downstream tip of the St-Denis island, that we shall use as downstream boundary conditions for our 2D model.

11.3.1 Estimation of Strickler rugosity coefficient

Our approach to K_s estimation is fully detailed in appendix H.1. Hereafter, we shall only recall its main steps.

The task of calibrating properly the Strickler is rather tricky for the following reasons :

1. Surface profiles in the Seine River are poorly known, since water levels are only recorded at dams (and at some bridges of Paris) and, furthermore, that the recordings are only stored for set times (generally at 7 a.m. and 5 p.m.).

2. Due to the dominant part played by navigation dams in the regulation of the river hydraulics, there is no one-to-one, unique, relationship between surface slope and flow rate in a given reach.

3. Lastly, whole parts of the Seine are unknown (all the arms which are not open to barge navigation).

We have first tried to estimate the Strickler coefficient for several "reference" situations, relative to flow rates of 100, 200, 300, 400 and 500 m^3/s . The corresponding water level differences in the Suresnes-Chatou reach (namely 5, 26, 56, 87, 118 cm) are deemed to be statistically representative of Seine flow conditions (Even & Poulin, 1993). The Strickler coefficient was calibrated according to the following working hypothesis :

(a) The Strickler coefficient is considered to be uniform all over the reach. This seems a reasonable assumption as the Seine features (cross sections, bed materials) appear to be fairly uniform downstream of Paris.

(b) The unknown left arm off St-Denis island is approximated based on an average bed slope defined from transverse profiles closest to the upstream and downstream tips of the island and on the respect of typical arm widths as measured on aerial maps of the river. 6 different reconstitutions are tested, corresponding to three different bed slope values and two different shapes (rectangular or trapezoidal) for the cross sections (cf section H.1.2).

(c) The right arm off St-Denis island is described first as in PROSE (12 sub-reaches of average length 590 m). Secondly, its geometric description is refined after a closer examination of the cross section variations (26 sub-reaches of average length 270 m).

The results of this calibration exercise are reported in table 11.4. The different approximations of the arms geometry have little influence on the Strickler calibration (cf H.1.3) : $\pm$ 0.5 with respect to the values indicated in table 11.4. On the other hand, once equal Strickler coefficients are assumed in each arm, it is their geometric description which conditions flow balance, whatever the flow rate. According to the assumed bathymetric reconstitution, the right arm appears to convey between 54 and 57 % of total flow.
Due to the rather large number of assumptions formulated when calibrating the Strickler, we do not consider the values reported in table 11.4 as reliable estimations of K_s variation as a function of the flow rate but reckon that they merely define the probable range of the Strickler coefficient in the reach. This range of values is slightly lower than what has been reported about rivers with features similar to the Seine ones, namely $40 \leq K_s \leq 45$ (Carlier, 1986).

Table 11.4: Strickler estimation for reference situations in the Suresnes-Chatou reach

Q (m^3/s)	100	200	300	400	500
K_s	34	29	28	32	33

As usual our knowledge of the free surface profile on the very day of the dye-tracing experiment is poor (cf 11.2.2). We have looked for the Strickler value which allows us to reproduce the free surface profile observed at about 7 a. m., namely 23.75 m downstream of Suresnes dam (this one being regulated so that the water level upstream of the dam is 26.49 m) and 23.63 m upstream of Chatou dam. The Seine flow is set to 170 m^3/s in Paris, which reduces to 168.8 m^3/s at Suresnes, because of the water intake located upstream of the lock.
A uniform Strickler value of about 35-36 allows us to reproduce the water level difference between the reach extremities (cf H.1.4), but not the observed flow repartition around St-Denis island (namely 59.5 % of the discharge into the right arm). In order to achieve a correct flow balance, rugosity must be increased in the left arm : this is translated into a reduction of $\simeq$ 20 % of the Strickler coefficient in the left arm with respect to the value adopted in the right arm.

Considering that the features of the Seine bed are fairly uniform, it is delicate to justify such difference. Consequently, this diminution of the Strickler appears to be a mere technical trick which somehow compensates for our approximate description of the river bathymetry, notably of the left arm.

Finally, **our starting point for calibrating bed roughness in the frame of two-dimensional hydraulic modelling will be to take the Strickler coefficient equal to 36 all over the studied reach.** We may expect to be obliged to make some adjustment to this value since the PROSE model and ours rely on a different description of the river, much more detailed in the latter case.

11.3.2 Estimation of downstream water level

Since only the consequences of dam regulation are known, but not the details of their operation, it is difficult to contemplate a detailed simulation of the free surface profile evolution throughout the day. Consequently, we decided to simulate the flow in the Suresnes-Chatou reach for three cases only : when the dams are regulated according to the 7 a.m. and 5 p.m. informations and for an intermediate situation at noon (then, the water level upstream of Suresnes is assumed to be 26.46 m). As regards the flow discharge, we may consider it is constant : the reported variations are small (Simon, 1992), they have the same order of magnitude as the uncertainty in the flow measurements. Thus, the discharge is always set to 170 m^3/s in Paris (then, 168.8 m^3/s at Suresnes). For each case, we computed the transit times between monitoring sections (cf table 11.5) and the water level at the downstream tip of St-Denis island.

1. In the column "*observed transit time*", we indicated a range of values rather than a single one. Indeed, as the distribution of rhodamine across the river section is quite uneven, it is not straightforward to define the center of mass of the whole cloud of dye. We computed the range of transit times between two successive monitoring sections by evaluating and substracting the first-order temporal moments of observed pollutographs. For instance, 3h 43 is the transit time of the central part of the cloud between St-Denis and Epinay bridges while 4h 20 is the transit time evaluated from bank pollutographs.
2. As regards PROSE results, we noticed that, the lower the downstream elevation prescribed at Chatou dam, the lower was the surface profile in the whole reach and thus, the smaller the wet sections and the larger the transfer velocities. However, these velocity variations are quite moderate and have little influence on computed transit times.
3. We could not expect the agreement between PROSE and the observations to be perfect since PROSE computes velocities averaged over the whole cross section whereas, due to the transverse heterogeneity of the dye cloud and the differential advection within a cross-

Table 11.5: Observed vs. computed transit times between monitoring sections (8/9/92)

Section	Transit time observed since previous section	Transfer velocity (cm/s) since previous section	Transit time computed by PROSE
Injection			
Gennevilliers bridge	2h 00 - 2h 12	15.0 - 17.0	1h 34 - 1h 37
St-Ouen bridge (railway)	0h 39 - 0h 44	20.0 - 22.5	0h 38 - 0h 39
St-Ouen bridge (road)	1h 11 - 1h 26	18.5 - 22.0	1h 16 - 1h 19
St-Denis bridge	2h 11 - 2h 40	22.0 - 27.5	2h 18 - 2h 23
Epinay bridge	3h 43 - 4h 20	21.0 - 24.5	3h 48 - 3h 58

section, different parts of the cloud travel at different velocities. However, the agreement is quite good, except between the injection and Gennevilliers brige.

The large discrepancy between observed and computed transit times from the injection point to Gennevilliers bridge can be ascribed to two major causes. First, the dye cloud stayed close to the injection (right) bank where velocities should be significantly lower than the average one. Secondly, Clichy is located within an area where no bathymetric information is available. There, considering the apparent regularity of the river width, the bathymetry was approximated by linear interpolation between the last profiles known upstream (P23.08 at PK 23.2) and downstream (P23.37 at PK 23.9) of this zone. It is possible this is too crude.

On the other hand, the fair agreement between computed and observed transit times from St-Ouen road bridge to St-Denis bridge denotes by the way that our reconstitution of the 500 m unknown stretch upstream PK 28 (cf table 11.3) is appropriate. It also confirms that (twenty-year) old soundings (e.g. from PK 26 to PK 27) still reflect correctly the river geometric features.

The free surface elevation computed at the downstream tip of St-Denis island depends on the downstream Chatou condition : its value is respectively 23.67, 23.60, 23.53 m for 23.63, 23.56, 23.48 m observed at Chatou (conditions at 8 a.m., noon, 5 p.m.). In fact, the surface profiles computed for the different combinations of open boundary conditions are nearly parallel

($\simeq$ 7 cm apart). A variation of 7 cm of the water depths modifies but slightly the extent of the river wet sections ($\simeq$ 1 %) : this explains why PROSE results are barely different from one simulation to another (cf point 2 above). Considering the shape of Seine cross sections, the slight variation of the water level should similarly have little influence on the flow patterns (transverse repartition of velocities, their respective magnitude) except very close to the banks.

Consequently, **we decided to undertake the calibration of the two-dimensional model for one set of open boundary conditions only, corresponding to the intermediate state of the surface profile, namely**

- **elevation at the island downstream tip : 23.605 m**
- **elevation at the island upstream tip : 23.658 m**
- **constant flow rate : 170 m^3/s in Paris, thus 168.85 m^3/s at Suresnes lock**

11.3.3 Analysis of the dominant phenomena

11.3.3.1 Order of magnitude of diffusivities in the Seine River

Estimation of mass diffusivities in the Seine River has already been attempted, either with one-dimensional or pseudo two-dimensional models, never - to our knowledge - with the help of a complete two-dimensional model. The trouble is that *diffusivity estimation is dependent on the accuracy of the flow pattern description.* For instance, figure 11.9 recalls the main mechanisms which control the longitudinal spreading of a pollutant cloud as it flows downstream. Longitudinal dispersion appears to be caused by differential advection, i.e. by the fact that velocities differ from their mean, whether it is a temporal (turbulent diffusion phenomenon), a depth-average (cf fig 11.9 (b)) or a cross-section (cf fig 11.9 (a)) mean. Similarly, transverse dispersion is to be ascribed mainly to secondary currents whose representation is beyond the grasp of two or one-dimensional models. If the hydraulic model fails to describe faithfully the differential advection or transverse currents within the cross section, this will be somehow compensated when calibrating the diffusion coefficients. Consequently, *we have used previous interpretations of dye-tracing experiments only to provide us with an order of magnitude of the diffusivities.*

In pseudo two-dimensional model KALPLAN (Bujon, 1983), the Seine is assumed to be a series of rectangular, uniform, sub-reaches. The cross-section averaged velocity is uniform within each sub-reach. Transverse velocities are null. The transverse distribution of longitudinal velocities is trapezoidal. The trapezoidal shape has been chosen for the following reasons : (i) it was closer to the flat parabolic velocity profile observed in straight reaches of the Seine than assuming an uniform velocity distribution (ii) this form allowed an analytical solution of the two-dimensional advection-dispersion equation. KALPLAN has been calibrated with

various experiments, led since the beginning of the eighties. The diffusivities applied downstream Paris are the following (Bujon, 1992, personal communication). The longitudinal and tranverse diffusion coefficients are respectively set to 3.5 and 0.05 m^2/s for a flow rate of 90 m^3/s, to 8 and 0.07 m^2/s for 175 m^3/s. They are interpolated between these bounds for intermediate discharge values.

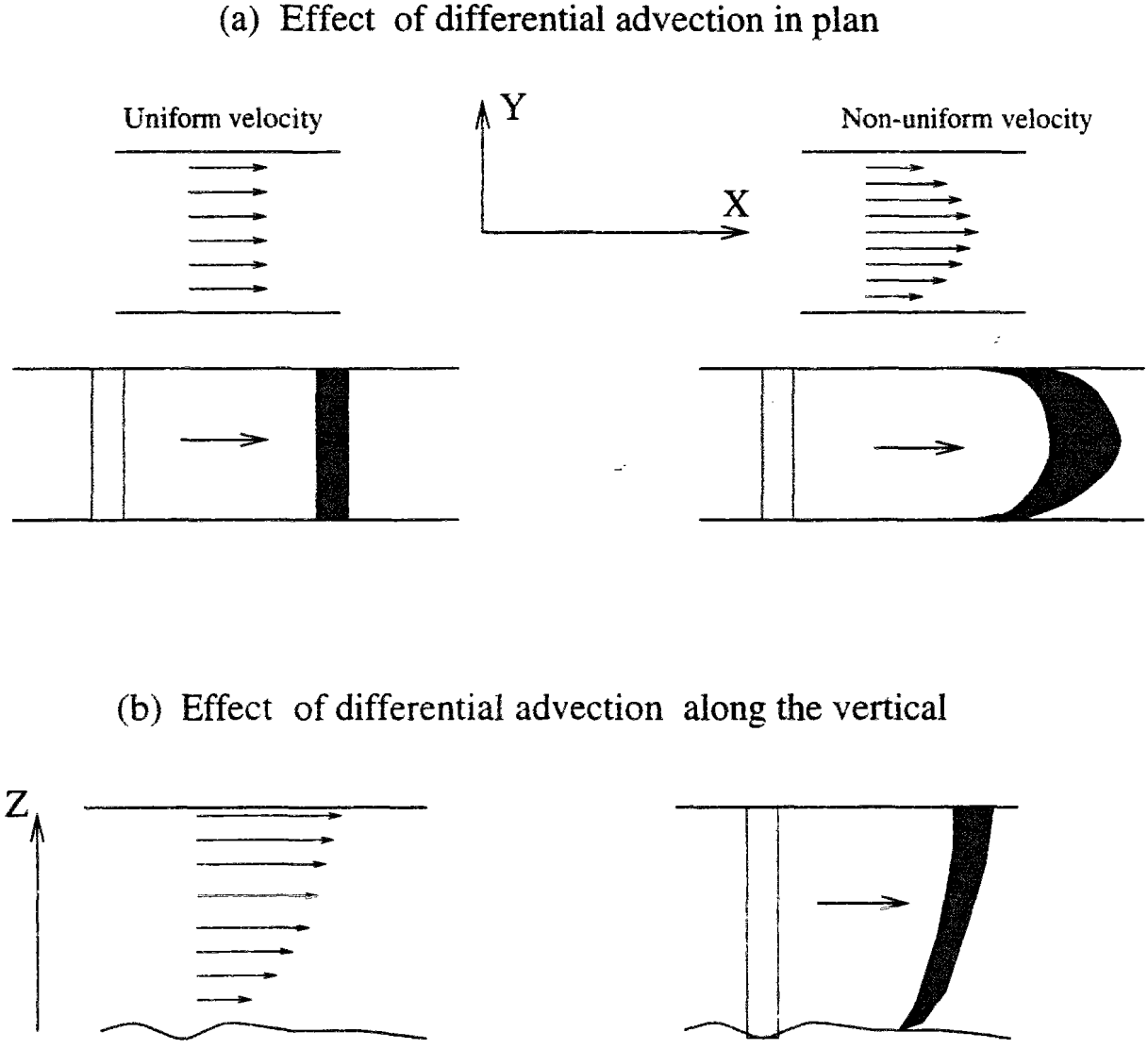


Figure 11.9: Main mechanisms causing longitudinal dispersion

In pseudo two-dimensional model TULIPE (Théry *et al.*, 1993a; Théry *et al.*, 1993b), the cross-section averaged velocity and free surface elevation are provided by a one-dimensional St-Venant model, the PROSE model we already mentioned. Thus, the geometry of the river need not be simplified (yet, the precision of its description depends on PROSE space steps, approximately 500 m long ...). Then, the velocity profile is built according to the following

hypothesis : (i) velocities are parallel to the river longitudinal axis (same as in Kalplan) (ii) energy loss depends only on bed and banks friction (ii) the slope of the energy line is uniform across a section. Thus, the ratio of local velocity to cross-section averaged velocity turns to be a power function of the ratio of local to average depths. This method for building stream tube models was proposed twenty years ago (Holly, 1975) and has since then met reasonable success (Yotsukura & Sayre, 1976; Holly, 1979; Harden & Shen, 1979; Cunge *et al.*, 1980b; Holly & Nerat, 1983; Holly *et al.*, 1990). TULIPE was applied to two reaches of the Seine upstream of Paris and of the junction with the Marne River. Dye-tracing experiments corresponding to three different flow rates per each reach were available. In the first one (10 km long), which is fairly straight, average longitudinal diffusion coefficient varies between 0.5 and 10 m^2/s (flow rate between 50 and 300 m^3/s). In the second, fairly meandering, 16 km reach, longitudinal diffusivities are larger : from 5 to 40 m^2/s (flow between 50 and 200 m^3/s). Both reaches are significantly narrower than downstream Paris so that, for the same flow rate, velocities are much larger than they would be downstream (they are approximately twice as big). Average transverse diffusivities belong to the range 0.05 to 0.15 m^2/s (first reach), 0.15 to 0.30 m^2/s (second reach).

We shall comment hereafter the observed dependency of these diffusivities upon hydraulic parameters such as friction velocity and water depths. For the time being, let us look only at the order of magnitude of diffusivities in the Seine River, as suggested by pseudo two-dimensional models :

- The scale of longitudinal diffusion coefficient is 1 to 10 m^2/s in fairly straight parts. It seems to be larger in a meandering reach. However, this is perhaps due to the fact that the approximate velocity profiles assumed in stream tubes models are farther apart from the reality in bends than in straight reaches, so that the diffusion term compensates for more "errors" as regards advective transport. Indeed, in a model such as TULIPE the maximum velocity is always next to the point of maximum depth. On the other hand, in bends, the point of maximum velocity tends to move from the inner bank (bend entry, effect of tranverse free surface slope) to the outer one (bend exit, action of transverse circulation) (Uan, 1979).

- The transverse diffusivity is twenty to a hundred times smaller than the longitudinal one : 0.05 to 0.15 m^2/s in straight parts, twice larger in meandering parts. This increase is undoubtedly representative of the increasing strength of secondary currents in bends.

We recall that mass and momentum diffusivities have the same order of magnitude in a river (cf introduction to 11.3).

11.3.3.2 Adimensional analysis of physical factors

Now that we have an estimation of the respective orders of magnitude of flow velocities, water-depths, diffusivities and roughness coefficients of the Seine River, it is possible to perform the kind of adimensional analysis of flow equations we introduced in 8.2.

The average (cross-section) depth in the right arm off St-Denis island is in the range [3.7, 4.7]. It is slightly deeper upstream of the island (from 4 to 5 m) (cf (Simon, 1992), appendix C). Finally, we consider the typical range of water depths in the studied reach (from Clichy to St-Denis island downstream end) to be [3.5, 5] m. Consequently, the wave celerity varies between 6 and 7 m/s. On the other hand, the average flow velocity lower and upper limits appear to be respectively 15 and $\simeq$ 30 cm/s (cf table 11.5). Consequently, the Froude number is extremely small, between 0.02 and 0.05. The relative weight of the propagation terms with respect to the inertial ones is inversely proportional to the square of the Froude number : the former ones should be increasingly dominant over the latter ones ($400 \leq p_1 \leq 2500$).

Typical length scales of the studied reach are 1 km in the flow direction, 100 m in the transverse direction. We recall that the respective weight of the diffusion terms is "measured" by ratio $p_2 = \nu/VL$ where ν, V and L are diffusivity, velocity and length scales (cf 8.2). According to one-dimensional PROSE simulations, a typical scale for the longitudinal velocity is $\simeq$ 0.2 m/s. According to KALPLAN (Bujon, 1983), a typical scale for the longitudinal diffusivity would be 10 m^2/s : then, p_2 equates 0.067, which means horizontal diffusion of momentum is rather unimportant with respect to advective and overall propagation terms. This observation is confirmed on other reaches. Indeed, we also evaluated ratio p_2 for the reaches upstream of Paris where TULIPE was applied. p_2 depends on the flow rate and varies respectively between 0.01 and 0.02 (straight reach) and from 0.03 to 0.10 (meandering reach).

From the work reported in section 11.3.1, we deduce that, broadly speaking, the Strickler roughness coefficient is in the range [30, 40]. Consequently, the product $p_3 = gL/K_s^2 H^{4/3}$ (H typical depth scale), which represents the weight of friction terms, should vary between 0.7 and 2.

In conclusion, it seems that in the frame of the Seine River application : **inertial (advection) and friction terms have roughly the same order of magnitude, horizontal diffusion of momentum is significantly smaller (at the utmost, a tenth of the preceding terms), propagation terms are by far the dominant ones in the flow equations.** The smallness of diffusion terms also suggests that a correct treatment of the advection part of the advection-dispersion equation is essential in the simulation of dye dispersion.

This adimensional analysis leads us to drop the diffusion terms in the flow equa-

tions. As they should have very little or no influence at all on the computation of flow patterns, it seems irrelevant to keep in our working equations those terms we cannot identify properly and which merely induce some extra computational effort. **The analysis also confirms that, since the propagation terms are dominant, it is important to a have a proper description of the river bathymetry, as it conditions heavily the flow depths repartition.**

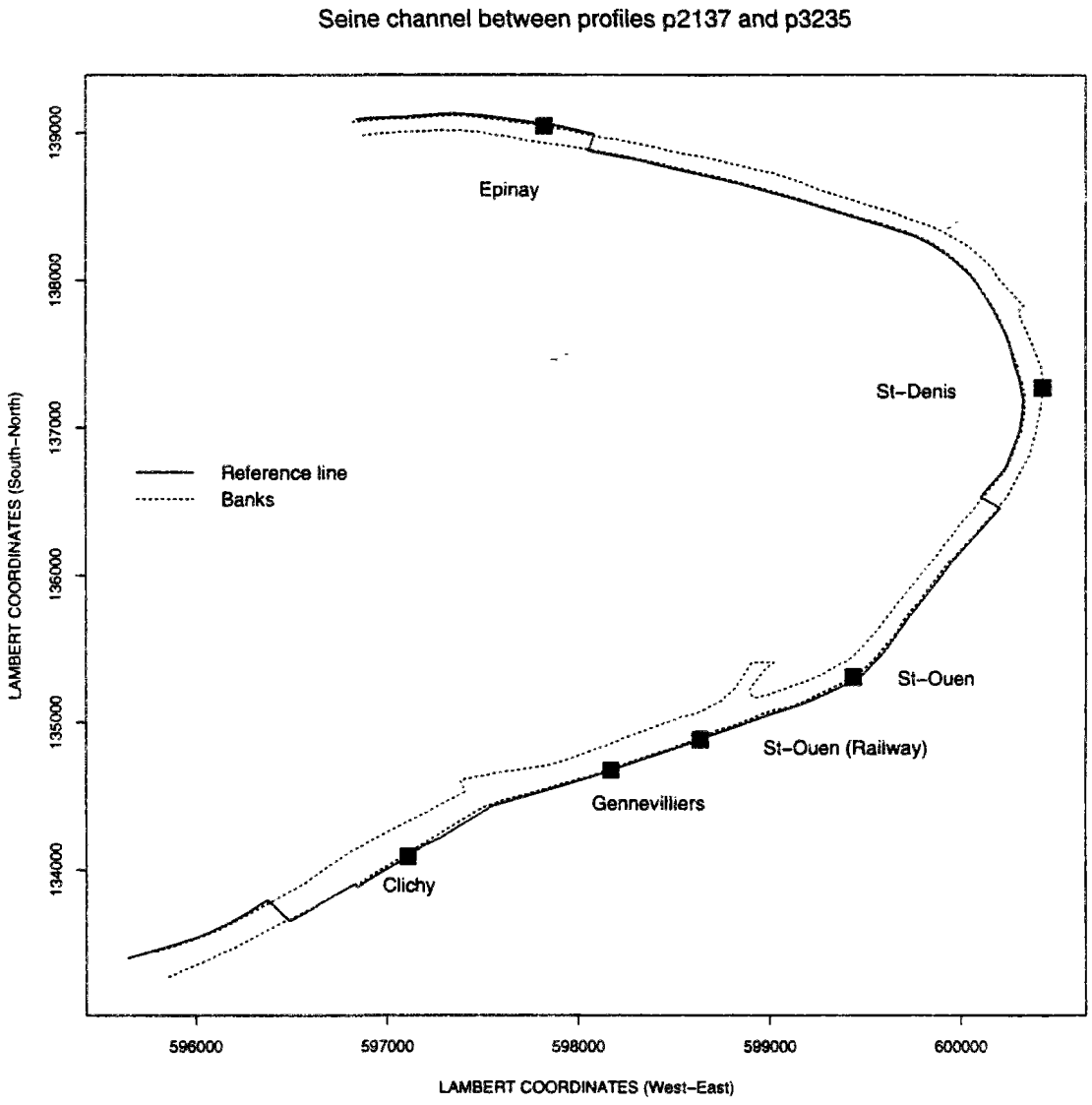


Figure 11.10: The dye-tracing experiment reach

11.4 Two-dimensional hydraulic modelling

11.4.1 Discretization of the studied domain

The studied reach is drawn on figure 11.10 in the Lambert Coordinate system. Its upstream limit is located about 4 km downstream of the Suresnes lock, starting after two small islands (island of Puteaux and la Grande Jatte). There are in fact two "downstream limits". The farthest one is the downstream end of the right arm off the St-Denis island. We have chosen not to model the unknown left arm. Consequently, its beginning constitutes another downstream open boundary of the computational domain (cf details in appendix H.2).

We have worked in two stages. In the first one, we modelled the reach only till a section located approximately 200 m downstream of the St-Ouen (road) bridge. Previous to the discretization, we changed to a coordinate system whose origin is located at point (596000,133400) and whose X-axis forms a 30 degrees angle with the West-East direction. In that way the studied part of the river (which is fairly straight as can be checked on fig 11.10) is approximately parallel to the new X-axis : this eases the discretization. The space steps in the $x-$direction and $y-$direction (the transverse one) are respectively set to 25 and 5 m. The domain includes thus approximately 5200 computational points.
We chose to work first on this smaller area in order to analyze the flow repartition around the island, its sensitivity to the Strickler value and boundary conditions (cf section 11.4.3).

Then, we have been extending our computational domain in order to include the whole of the right arm. We have been using another coordinate system (origin (600000,137500), rotation of 70 degrees). In this system, the Seine right arm takes the form of an inversed "L", whose upstream (upstream St-Denis bridge) and downstream halves are approximately lined up with the X and Y axis respectively. This change of coordinate system is once again intended to simplify the discretization and achieve a more exact approximation of the river geometry. As for the first domain, we tried to use space steps about 25 m long in the flow direction and 5 m wide in the transverse direction. Of course, as the main flow direction changes, there is, in the curve, a "transition area" where the space step in the $x-$direction is progressively reduced (with a geometric ratio of 0.95) while the space step in the $y-$ direction is increased (geometric ratio 1.05). Finally, the entire reach represents about 13000 computational nodes.

11.4.2 Conditions of hydraulic simulations

Model version We are dealing here with a flow where the water-depth variations within a time step (and during the whole simulation by the way) are really slight : there should be no

trouble linked to the use of a model version whose dependent variables are the unit-width flow discharges and the free surface elevation. We worked thus with a "depth/discharge" formulation as this formulation behaves generally better than the "depth/velocity" one as regards flow and mass conservation (cf chapter 9). Of course, no simplification of the propagation step equations is allowed.

Initial conditions In order to achieve a steady-state flow, we chose the following strategy. Initially, the river was supposed to be at rest (flat free surface elevation, null flow). At the upstream limit, the flow was raised progressively to its chosen level (in 3600 s for the first domain, 900 s for the second, bigger, one) then remained constant. In this stabilization phase, the velocity field was stored at fixed intervals and the flow repartition was subsequently computed at several sections along the river (approximately every 100 m). Computations were stopped when this flow repartition appeared to reach an equilibrium.

Choice of time steps We have by no means tried to optimize the time step but deliberately chosen small ones which would ensure a safe, even if slow, convergence. We have seen in the preceding chapters that the stability and accuracy of the factorization method applied to the propagation step equations depends on the magnitude of the propagation Courant number C_p, namely the ratio $\sqrt{gH}\Delta t/\Delta l$. We have seen above (sec. 11.3.3) that the wave celerity $\sqrt{gH}$ is about 6-7 m/s in the Seine. Considering a mesh size Δl of 25 m in the direction of wave propagation, we obtain that C_p is approximately 3 for $\Delta t = 10$s. When working on the first domain, we have indeed chosen to work with $\Delta t = 10$ then 5 s (stabilization phase). In the second domain, the mesh size in the flow direction is smaller than 25 m in the transition area corresponding to the description of the river curve : it can be as small as 5 m. Consequently, we have used an even smaller time step, 2 s. In the first domain applications, the steady-state situation was achieved after approximately 450 iterations in the stabilization phase. Flow needs more time to develop into the second, longer, reach : convergence requires about 1400 iterations. From a computational point of view, our method to obtain the steady-state velocity field is obviously far from being efficient. We could have tried to start the simulation with an approximate flow distribution closer to the final one than the "river at rest" situation. We could have looked for a more efficient combination of time steps. However, we decided to focus on the results of the simulation, on their hydraulic interpretation, and to care little for CPU time improvement.

Advection solution With such small time steps, the advection Courant numbers are very low (no more than 0.2). Besides, advection terms are not the dominant ones in the flow equations (cf 11.3.3). Consequently, we solved them with the simple, economic, upwind algorithm.

Open boundary conditions At the downstream limits, the boundary condition consists of a fixed water level, obtained with PROSE simulations. A posteriori examination of the forecast surface profiles discloses that in most sections of the river (notably in the curve) the free surface elevation is not uniform. However, the variations are small (generally less than 0.3 mm from one bank to another). Consequently, the fact of assuming a uniform elevation across the downstream sections should have little or no consequence on the flow computation.

At the upstream boundary, our model requires us to prescribe not only a global flow rate but also the flow distribution across the section. An approximate repartition of the discharge is obtained following the method used in stream tube models (model TULIPE, (Théry *et al.*, 1993a), cf 11.3.3.1). In order to limit as much as possible the influence of this approximation, we have been careful to locate our inflow boundary far upstream ($\simeq$ 1.5 km upstream Clichy) of the area we are really interested in.

Solid boundary conditions At solid boundaries, we impose the normal gradient of the free surface elevation to be zero. Along solid boundaries, the normal component of the flow velocity is canceled while the tangential component is set free, i.e. is computed with the help of one-sided differences once the water levels have been forecast (cf section 8.6.2).
As we are working in a rectilinear, not curvilinear, coordinate system, the boundary of our computational domain is not parallel to the true banks. We assume this explains why in two or three limited spots of our computational domains (very shallow spots) we have found the aspect of the velocity field to be not fully satisfying.

11.4.3 Results

Strickler calibration We have first looked at the influence of the Strickler coefficient tuning on surface profile slopes. For that purpose we worked on the shorter domain. Free surface elevations at both open boundaries (left and right arm) are prescribed by the PROSE model : they equate respectively to 23.658 and 23.653 m. The difference between the water levels at the island tip and at a test section located 2600 m upstream the island should be about 14 mm (14.3 exactly !).

Table 11.6 sums up the main flow features observed for different Strickler values. First we can observe that flow conservation is good but not perfect. The most appropriate value seems to be $K_s = 40$, a value which is more in agreement than 36 with the expected range of roughness coefficients for a river such as the Seine. However, the water level difference is not dramatically sensitive to 10 % changes in K_s values and we are dealing there with differences that would be hardly measurable on the field.

The Strickler coefficient influences somewhat the flow repartition : as friction in the right arm is decreased (by increasing K_s), this one conveys slightly more flow, all other conditions being the same.

We have estimated all over the computational domain the differences between velocity modulus forecast according to different K_s values and made an analysis of their magnitude and location. A ten percent variation of the Strickler induces only on average a 1.5 % variation in velocity modulus (the lesser the friction, the higher the average flow velocity). This confirms that friction is not an essential factor controlling flow development in the Seine, as forecast by our adimensional analysis. Discrepancies are generally more marked in the right arm, as K_s influences the flow rate there. However, points where velocity differences are superior to 5 % are few (about 150 computational nodes) and clearly located in specific sites : the local widening of the Seine 500 m downstream Clichy ($\simeq$ 50 computational nodes, see flow pattern on figure H.9) and other nodes spread all over the computational domain but systematically in very shallow areas (ten to forty centimeters deep, close to the banks). In such shallow zones, friction, not surprisingly, plays a very important part in controlling velocity magnitude.

In summary, given the uncertainty in the surface profile data in the Seine, we cannot conclude for sure about what is the "good" Strickler value : **the most adequate value appears to be 40 but values in a ten percent wide range around 40 are also acceptable.** This incertitude about K_s has little consequence for the simulated flow pattern. It might play a more important part when dealing with the estimation of diffusivities. Indeed, these ones are usually related to friction velocity at the river bed. These friction velocities are proportional to depth-averaged velocity and inversely proportional to K_s. Consequently, a ten percent variation of the Strickler yields approximately a ten percent variation of the friction velocity, even while the depth-averaged velocity is undergoing only a 1 % variation.

Table 11.6: Influence of Strickler value on flow features

Strickler value	Level Difference (mm)	Flow rate (m^3/s) at island tip	Flow conservation (%)	Flow in right arm (m^3/s)	Flow in right arm (%)
36	17.0	166.2	-1.6	102.1	61.4
40	14.6	168.0	-0.5	103.5	61.6
45	12.3	169.8	+0.6	107	63

Study of flow repartition Assuming from now on that the adequate Strickler coefficient is 40 we have been studying the sensitivity of flow repartition to the downstream boundary condition. The water level difference between the island tip and our downstream boundary in the right arm (200 m after the St-Ouen bridge) is once again feeble : 5 mm ! We have applied

to it a 20 % (1 mm) variation. Besides, as detailed in app. H.2, two sets of recent soundings (1988 & 1989) are available as regards the area of the island tip : they indicate a local but large change in the river cross sections due to dredging works. We could not a priori discard a subsequent influence on flow repartition. Consequently, simulations have been performed with both sets of bathymetric data.

Table 11.7: Sensitivity of flow repartition to boundary conditions and bathymetry

Downstream water level	Flow rate (m^3/s) at island tip	Flow conservation (%)	Flow in right arm (m^3/s)	(%)
Soundings of 1988				
23.652	168.05	-0.47	107	63.7
23.653	167.95	-0.53	103.5	61.6
23.654	167.60	-0.74	100	59.7
Soundings of 1989				
23.652	168.69	-0.09	107.6	63.8
23.653	168.66	-0.11	104.1	61.7
23.654	168.30	-0.32	100.7	59.8

Considering the results displayed in table 11.7, it appears that flow repartition is slightly more influenced by the prescribed water level in the right arm than it was by Strickler modifications. On the other hand, changes in the local bathymetry around flow separation have minor influence on the discharges conveyed by each arm.

When we look at the differences induced by the variation of the downstream free surface elevation, we observe that the influence on the velocity field upstream the island is negligible (less than 0.4 % on average) and that notable discrepancies occur only in the right arm, because of the difference between flow rates.

As regards velocity field, forecasts with the same boundary condition but different bathymetry are only locally different, obviously at the location where soundings differ : there, the difference may be impressive (up to 60 %) because of the large variation in bottom elevations (up to 1.2 m from one year to the next).

All simulated flow repartitions (tables 11.6 and 11.7) are of the same order of magnitude as the observed one (60 % of the total discharge conveyed by the right arm).

Comments on the forecast velocity field Finally, we turned to the modelling of the whole reach, where a Strickler value of 40 allowed once again a correct approximation of the free surface

slope. As mentioned above, this leads to the forecast that 103.5 m^3/s flow through the right arm.

Different parts of the velocity field are displayed in figures 11.11 to 11.15 (and figures H.9 to H.11, appendix H.3). In these figures, we mentioned, besides the velocities (one every 2 computational nodes), the limits of the computational domain, the location of true banks and the names of surrounding transverse profiles.

Upstream of the island, the transverse velocity profiles appear to be very flat (cf 11.11), i.e. velocities across a section are nearly uniform. The only singularity is the local widening close to Clichy and already mentioned above, which constitutes a nearly stagnant area (cf figure H.9). The flow separation is illustrated on figure 11.12.

In the right arm, where cross sections are 40 to 60 % narrower than upstream of the island, velocities are larger and their transverse gradient is more important (cf figures H.10, 11.13 and 11.15), even in fairly straight parts of the arm (e.g. around Epinay bridge, fig 11.15). Velocity profiles have a rather marked parabolic shape. Once again, singularities are few (e.g local narrowing, fig H.11). The more marked is a slight recirculation (fig 11.14) that develops at a widening which corresponds to the arrival into the Seine of a small secondary channel, closed by a lock. This junction is located about 500 m downstream of the St-Denis bridge and close to the various sewer outlets of La Briche area (it is slightly noticeable on figure 11.10).

On the whole, flow patterns have a correct aspect. We shall see now, through the interpretation of the dye-tracing experiment, if they reflect reality too ! ...

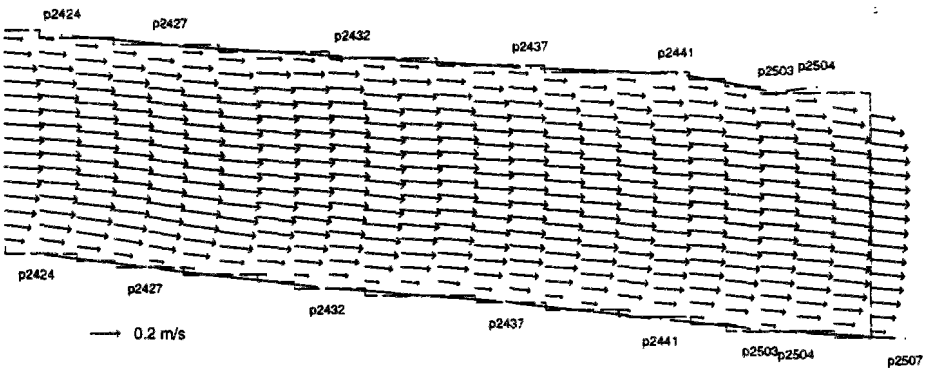


Figure 11.11: Flow pattern between Gennevilliers ($\simeq$ profile P2424) and St-Ouen railway bridge ($\simeq$ profile P2505)

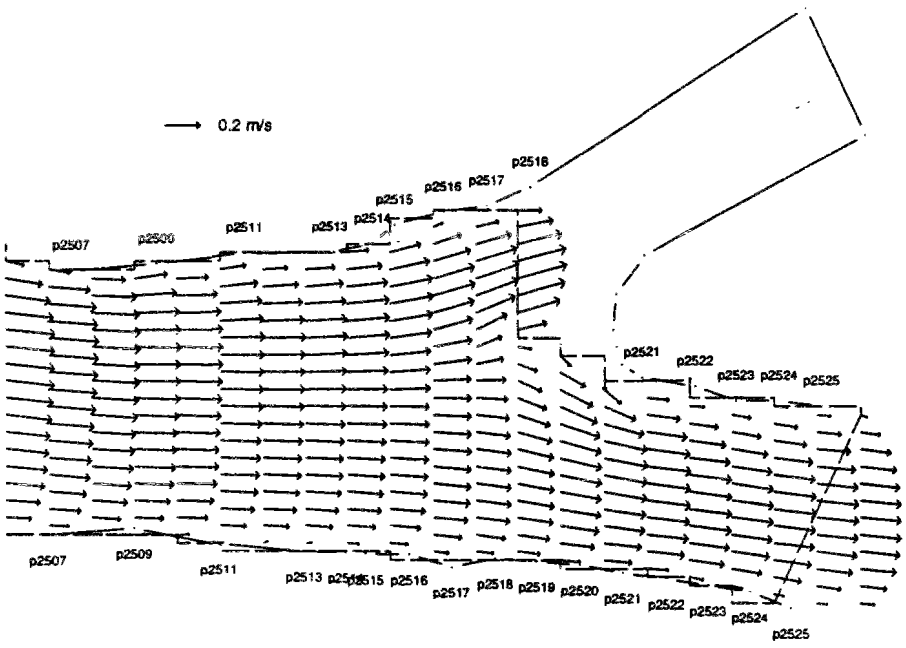


Figure 11.12: Flow separation around the St-Denis island

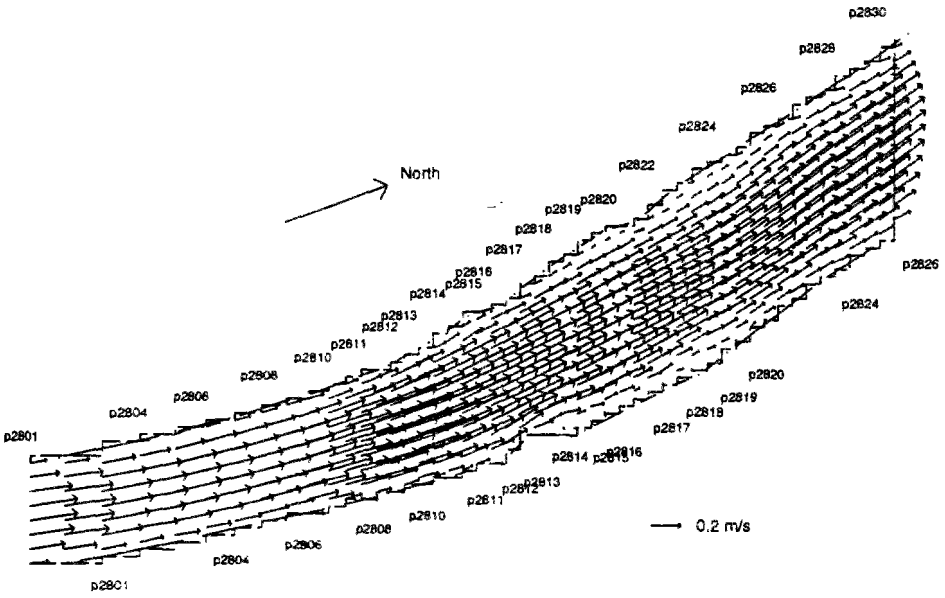


Figure 11.13: Flow pattern in the vicinity of St-Denis bridge ($\simeq$ profile P2813)

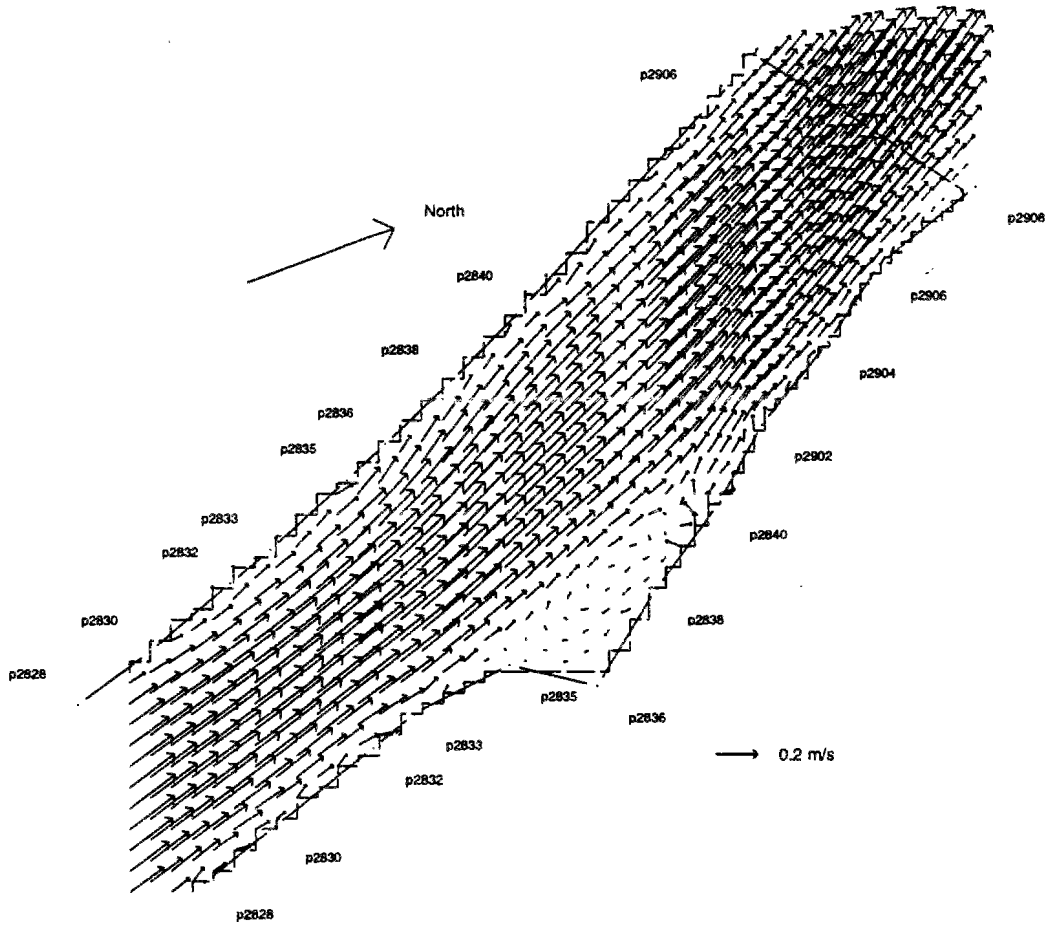


Figure 11.14: Flow pattern at the lock of La Briche

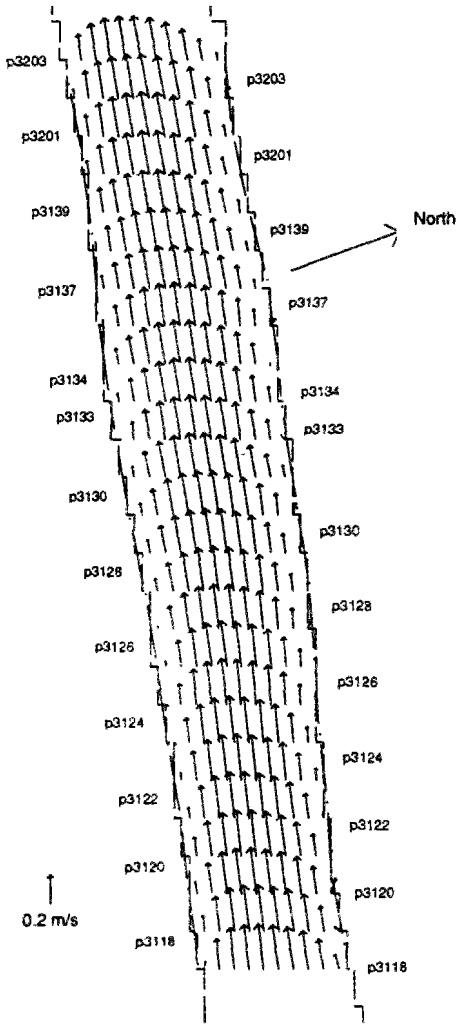


Figure 11.15: Flow pattern in the vicinity of Epinay bridge ($\simeq$ profile P3133)

11.5 Interpretation of the dye-tracing experiment

11.5.1 Working hypothesis

11.5.1.1 Inflow concentration field

We are working with a depth-averaged model, unable to deal with vertical gradients. Unfortunately, the dye was not injected all over the water column but, let's say, in the fifty centimeters closest to water surface. This is very different from what occurs with sewer outlets (the Clichy one, for instance, enters the Seine through a kind of lateral channel, of the same depth as the river). Measurements performed at the Gennevilliers bridge have demonstrated that the vertical homogeneity is nearly achieved there. However, how can we deal with Clichy-Gennevilliers reach ? In fact, several options were available :

- to try and apply a pseudo three-dimensional model (two-dimensional for the flow, adding vertical diffusivity as regards concentration);
- to assume instantaneous vertical mixing at the injection, calibrate whatever longitudinal and transverse diffusivities we could in order to reproduce the observations made at Gennevilliers bridge ... and discard the so estimated diffusivities;
- to consider the Gennevilliers bridge as our upstream boundary as regards the dye transfer modelisation and take advantage of the measurements there in order to "build" an inflow dye field.

We chose the last option, considering it would probably not induce more incertitude and errors than the two other ones and that it was the more straightforward to apply. The concentration across the Seine under Gennevilliers bridge is thus obtained through linear interpolation between the measuring stations. As regards the concentration between the right bank and the measuring point closest to it, it was either assumed to be constant or extrapolated. Choosing one or another method bears little consequence : it yields (feeble) modification of the forecasts only at St-Ouen railway bridge at the point closest to the right bank, none at other points.

When the dye transverse distribution is so reconstituted, the total dye mass flowing through the Seine at Gennevilliers appears to be about 13 kilograms, 2 less than the injected mass. This denotes that there was some dye loss. This is plausible : measurements were stopped before the arrival of late pockets of dye and before the flowing of tails along the bank was achieved. However, it also probably reflects that our reconstitution is somewhat too crude.

11.5.1.2 Simulation conditions

We have worked with the steady-state velocity field forecast with Strickler coefficient set to 40. We used the same computational grid as for flow modelling.

To model the concentration transport, we have applied the backward characteristic method selected at the conclusion of part II (Akima bicubic interpolator) for advection-dominated situations. Use of a characteristic method allows us a flexible choice of the time step. We need to choose it big enough so that errors introduced by repeated interpolations are limited and small enough so that the inflow concentration field is precisely described. After some trials, we chose to work with $\Delta t = 360$s (6 min). Then, advection Courant numbers are about 3 and above.

11.5.1.3 Formulation of diffusion coefficients

Horizontal diffusion of momentum has been neglected in the computation of velocity field as it appeared to have negligible weight with respect to such terms as wave propagation (cf sec 11.3.3). However, horizontal diffusion of mass cannot be so immediately and easily discarded when computing mass transport as it is the only force which balances advective transport by the mean (depth-averaged) flow.

Transverse diffusivity In rivers, it is customary to relate transverse diffusivities ϵ_y to the product U^*h, where U^* denotes the bottom friction velocity and h the water-depth (cf B.3). This is justified by the fact that horizontal diffusion is notably induced by turbulence, which is mainly controlled by bed shear, at least in straight parts. The term U^*h appears to be the product of velocity and length scales typical of bed shear turbulence. Expressing ϵ_y as being proportional to U^*h is thus consistent with the eddy-viscosity approach in modelling mass or energy fluxes due to turbulence (cf section 2.2).
Yet, apart from bed shear turbulence, other mechanisms can contribute to transverse mixing (Holley & Abraham, 1973) : (i) secondary currents, especially helical motion associated with bends (ii) turbulence caused by various obstacles to the flow (boulders, groins, etc ...), if they are present. Secondary circulations are dependent on the precise features of the river (e.g. bathymetry, magnitude of the flow). For instance, (Lau & Krishnappan, 1977) observe on straight reaches that, all other things being equal (depth, friction velocity), transverse diffusivity decreases as the river width increases, because secondary currents are generally weaker in wider sections.
This explains why there is no "universal" value suggested for the proportionality coefficient α_y ($= \epsilon_y/U^*h$). However, its plausible ranges are fairly well known (cf sec. 2.3.2.2 and app. B.3.2) :

- *straight laboratory flumes or man-made channels :* $0.1 \leq \alpha_y \leq 0.3$ (average 0.23)
- *slightly irregular or gently meandering natural rivers :* $0.4 \leq \alpha_y \leq 0.8$ (average 0.6)
- *sharply curving channels (laboratory and field) :* $1 \leq \alpha_y \leq 4$

Longitudinal diffusivity Since rhodamine transport computations rely on a depth-averaged two-dimensional hydraulic model, longitudinal diffusion should account here only for the effect of turbulence and vertical non-uniformities. Then, Elder's analysis of wide open channels should apply (cf app. B.3.1) : it leads to the conclusion that ϵ_x is also proportional to U^*h. The proportionality coefficient α_x is about 6 (more exactly 5.93).

However, when longitudinal diffusivity is estimated in the frame of one-dimensional models, where it is bound to integrate the effect of transverse heterogeneity of the velocity, U^*h turns to be a poor explicative parameter, with α_x taking very large values (one to several hundred) without a definite pattern. One formula successful in reducing the level of uncertainty attached to ϵ_x forecast is Fischer's formula (cf B.3.3) : $\epsilon_x = \beta_x \left(\overline{U}^2 W^2\right) / \left(\overline{U}_* \overline{h}\right)$, with $\beta_x = 0.011$, where overbars denote cross-section averages, W the river width. This formula was found (by Fischer and independent researchers) to agree with observations within a factor of four of so, which is not bad ...!

Previous estimations in the Seine River

TULIPE applications In the recent applications of stream tube model TULIPE (Thery *et al.*, 1993b), the transverse diffusivity was represented with formula $\epsilon_y = \alpha_y U^*h$. As regards longitudinal diffusivity, since the stream tube model is somewhere halfway between a full 2D one (Elder formula) and a 1D one (Fischer formula), we decided to try the calibration of ϵ_x according to both formulae. Table 11.8 summarizes our results concerning the proportionality coefficients α_y, α_x or β_x.

The data linked to the first experiment in the straight reach (flow $Q = 49$ m^3/s) are rather dubious so that we shall focus on the other experiments. We may notice that the α_y values are consistent with observations reported in literature : α_y is about 0.7 in the first (straight) reach, between 2.5 and 3.5 in the second, meandering, one. On the other hand, neither α_x nor β_x values agree with the values recommended in their respective formula and besides, their dependency on flow rate is rather erratic.

KALPLAN data For a flow (175 m^3/s) very close to what we observed during the Clichy dye-tracing experiment ($\simeq$ 175 m^3/s), KALPLAN would use constant and uniform diffusivities, respectively 8 and 0.07 m^2/s in the longitudinal and transverse directions.

Our hydraulic simulations provide us with an order of magnitude for velocities and depths

Table 11.8: Results of stream tube model calibration

First (straight) reach			
Q (m^3/s)	49	145	280
α_x	30	5	60
β_x	0.00003	0.00005	0.00015
α_y	6	0.65	0.67
Second (meandering) reach			
Q (m^3/s)	55	140	210
α_x	100-125	275	125-150
β_x	0.0007	0.0016	0.0007
α_y	3-3.5	2.5-2.75	3.5

in the experimental reach. Their average values would be respectively $0.20 - 0.25$ m/s and 4.5 m approximately. For such values, and a Strickler coefficient of 40, friction velocity U^* is about $1.2 - 1.5$ cm/s and product U^*h about 0.055-0.07 m^2/s. Consequently, the proportionality coefficients α_y and α_x between diffusivities assumed in KALPLAN and explicative variable U^*h would be approximately 1 and 125. The order of magnitude is acceptable for α_y. The high value obtained for α_x stresses once again, as in TULIPE, that assessing the longitudinal dispersion with pseudo two-dimensional models is not easy.

Our choice As we are relying on a full two-dimensional model, the longitudinal diffusivity ϵ_x should account only for dispersion induced by vertical gradients of the velocity, and should thus be close to the value given by formula $\epsilon_x \simeq 6U^*h$. The resulting plausible range is then $0.33 \leq \epsilon_x \leq 0.4$ m^2/s. If we compute the Peclet number $U\Delta x/\epsilon_x$ in the longitudinal direction (which provides a measure of the relative strengths of advection and dispersion at the scale of the computational grid) we find it is around 16 (setting $\Delta x = 25$m). This indicates that longitudinal diffusion should play a part far less important than differential advection in spreading the rhodamine cloud in the flow direction. Consequently, *we begun our study by setting longitudinal diffusivity to zero.* In fact, we felt no need for then re-introducing it, as the model always correctly estimated the longitudinal spreading of the dye cloud (see following figures). Consequently, we focused only on the tuning of the ratio α_y. It is worth noting that the Clichy-Epinay reach consists first of a fairly straight part, followed by a large curve. We may expect α_y to be somewhat dependent on that drastic change in the river geometry. That is what we are investigating hereafter.

We have not been trying to reach a perfect, an optimal, adjustment of the diffusivities. We

have merely been checking that their range was consistent with results previously reported in the literature. This limited goal justifies why we relied only on visual comparisons of simulations and observations.

11.5.2 Comparison of forecasts with measurements

(nb : at a given monitoring section, measuring points are referred to with letters. They are alphabetically ordered according to their distance to the right bank, A denoting the point closest to the bank. On all figures, measurements are marked with square dots, solid lines connecting them.)

1. **St-Ouen railway bridge**

   Measures outline that the dye cloud stays close to the injection (right) bank. Indeed, significant levels of concentration are observed only at points A, B and C which are respectively located 10 m, 20 % and 30 % of the river width apart from the right bank. At the contrary, only traces are noticeable at point D in the middle of the cross section.

   Here, forecasts are displayed for 4 different values of proportionality coefficient α_y which tunes the strength of transverse diffusion. When no diffusion is assumed ($\alpha_y = 0$), the peak concentration (at point A) is clearly overestimated. A better agreement is obtained when introducing some diffusion, setting α_y to 0.23 as suggested by Elder, or to 0.4 or 0.6, which are typical values recorded for gently meandering reaches (cf figure 11.16).
   Forecasts yielded by these different α_y are very close except at point D (cf fig 11.17), the farthest from the injection bank ... and also the less relevant as regards dye fluxes through the cross section. This likeness can be ascribed partly to the fact that this monitoring section is close (500 m downstream) to our effective upstream boundary (Gennevilliers bridge) so that discrepancies due to different diffusivities had few time to develop.

   In all cases, the model predicts the arrival of the dye cloud at point A with a slight delay. It also overestimates the concentration at intermediate point C, which is possibly partly due to an inexact description of the entering concentration field.

   The positive points are that, assuming a slight diffusion typical of that kind of reach, both concentration maxima and the longitudinal spreading of pollutographs are fairly reproduced.

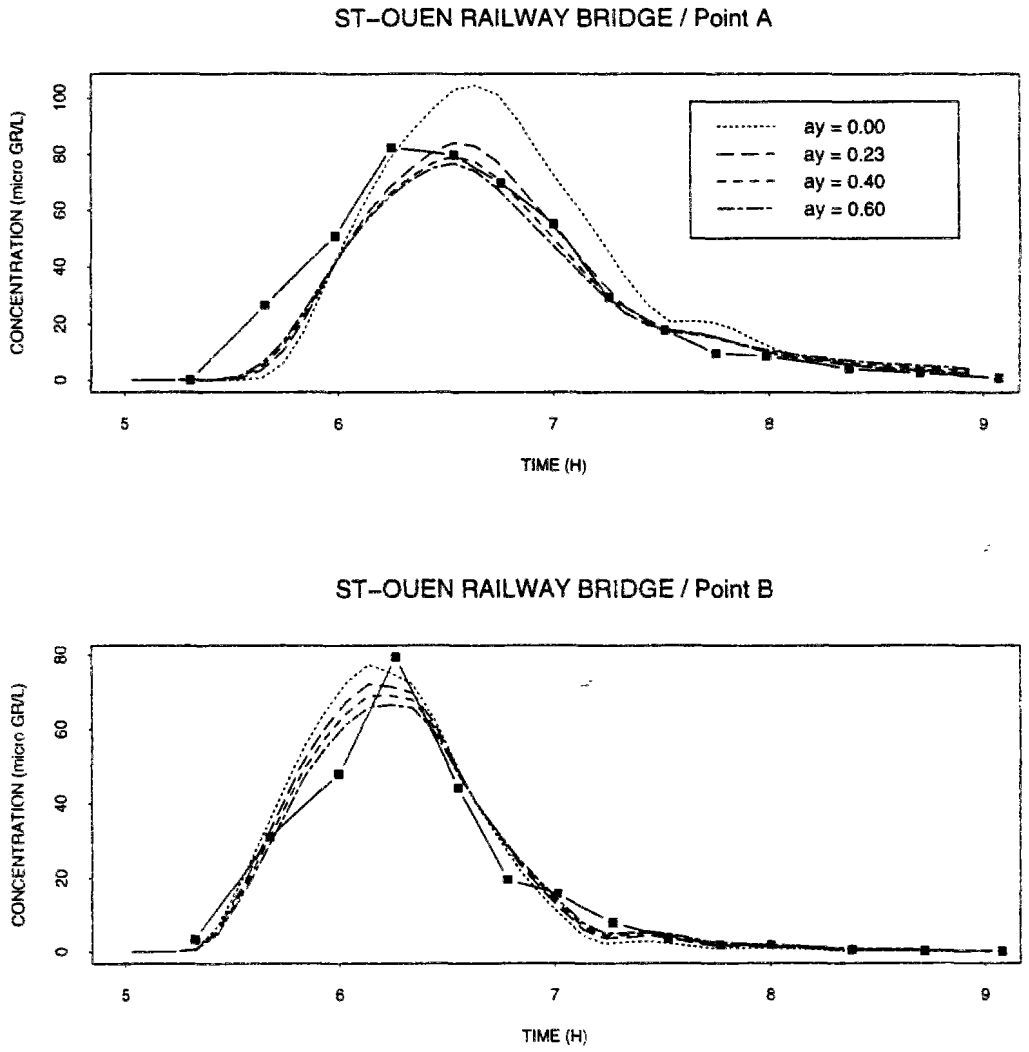


Figure 11.16: Measures and forecast pollutographs at railway bridge (10 and 35.7 m off right bank)

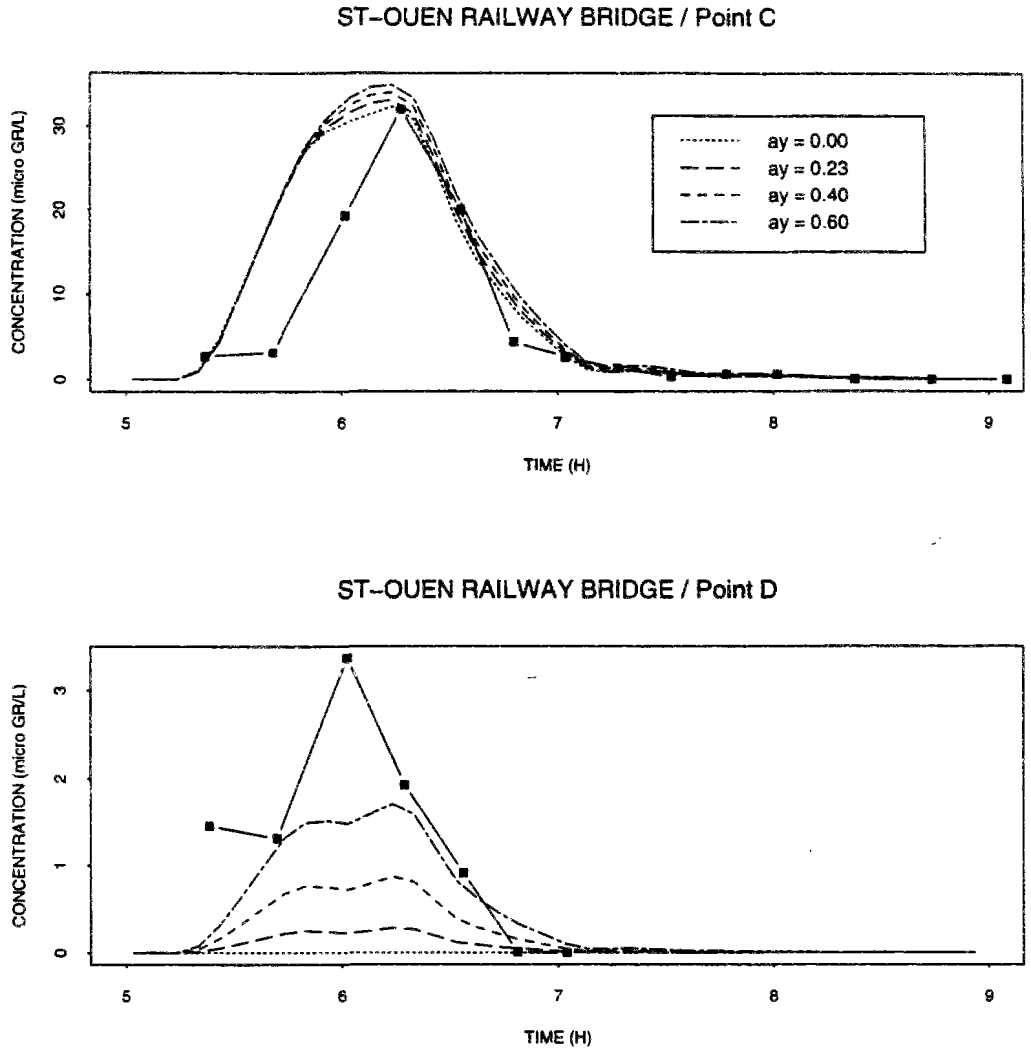


Figure 11.17: Measures and forecast pollutographs at railway bridge (50.2 and 79.8 m off right bank)

2. **St-Ouen road bridge**

This is the first monitoring section in the right arm. Here too the dye cloud flows mostly through the right half of the cross section (nb : points B and C are located respectively at one third and two thirds of the river width). At points B and C, we observe the passing of two local maxima, the first one at about 7h 20, the second one at 8h approximately. This could not be simulated.

When no transverse diffusion is introduced, the model not only forecasts too large concentrations along the banks but yields larger errors as regards the dye transit times. *The more the diffusivity is increased* ($\alpha_y = 0.6$), *the better the agreement between simulations and observations at the points where most of the dye seems to pass (points A and B, figure 11.18).* However, the model keeps on forecasting systematically a too early arrival of the dye at point B, which may notably be due to an imperfect estimation of the flow discharge conveyed by the right arm.

The model also tends to overestimate the extent of the spreading of the dye cloud across the river as can be observed at points B and C (even if, in the last case, the model fails to reproduce the extremum, cf fig 11.19).

Figure 11.20 displays the kind of transverse distribution that the model forecasts (here for $\alpha_y = 0.4$). It predicts that the peak will indeed flow between A and B monitoring points. In fact, we can be fairly sure that real transverse profiles were different ... as there is a bridge pier between A and B ! We didn't take into account the bridge piers when modelling the river. Yet, it is probable they influence locally the streamlines and perhaps induce preferential flow through one or another pass. In any further interpretation of this dye-tracing experiment, we could try to quantify this influence.

Finally, as at the railway bridge, *the longitudinal extent of the dye cloud is correctly evaluated.*

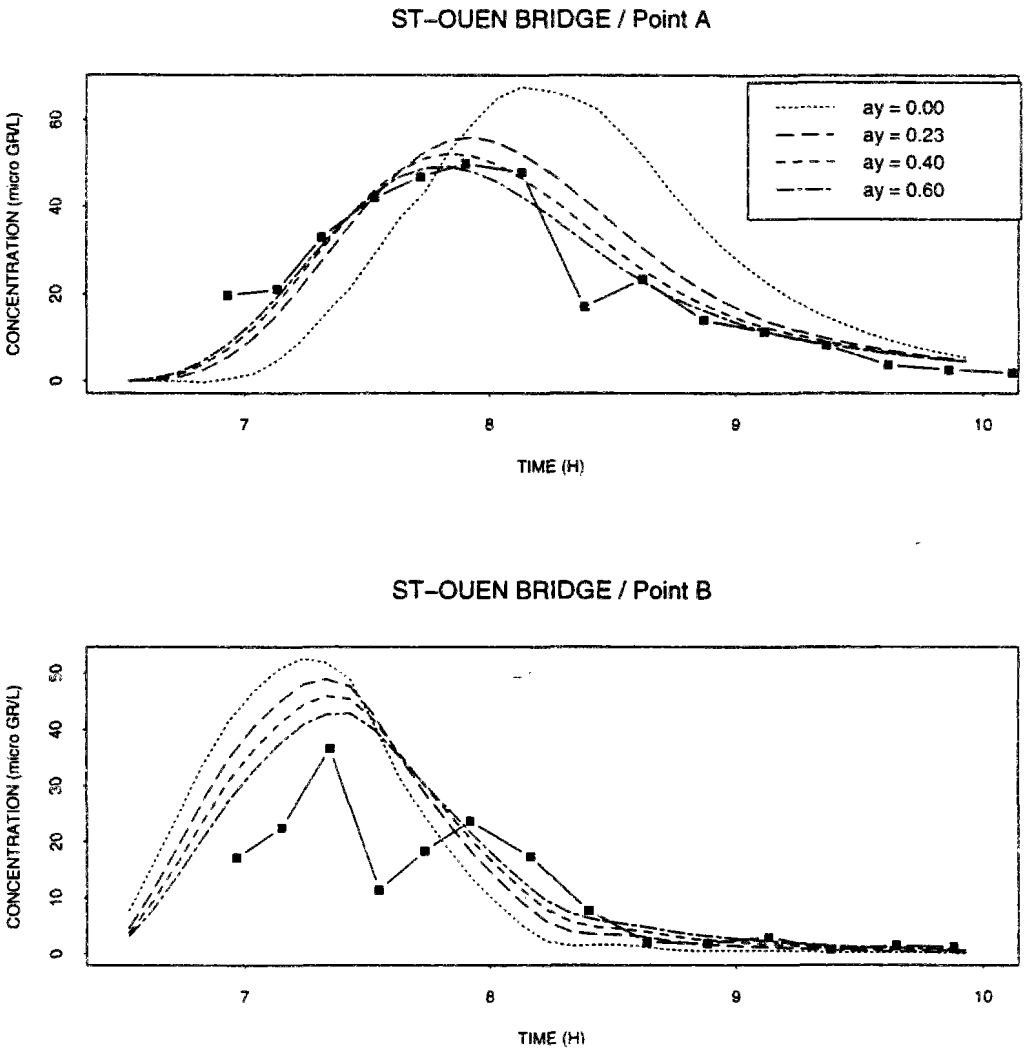


Figure 11.18: Measures and forecast pollutographs at St-Ouen bridge (11.2 and 41.5 m off right bank)

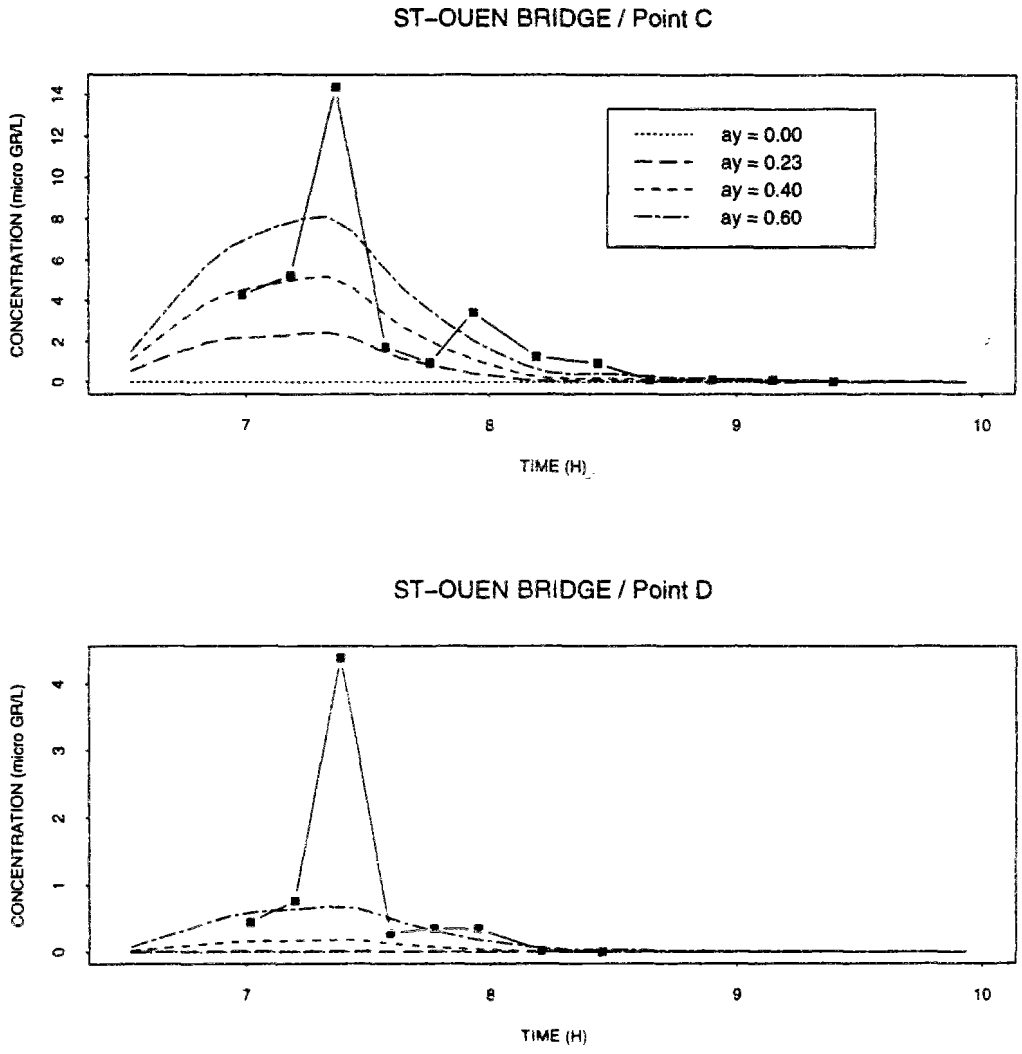


Figure 11.19: Measures and forecast pollutographs at St-Ouen bridge (74.7 and 101.7 m off right bank)

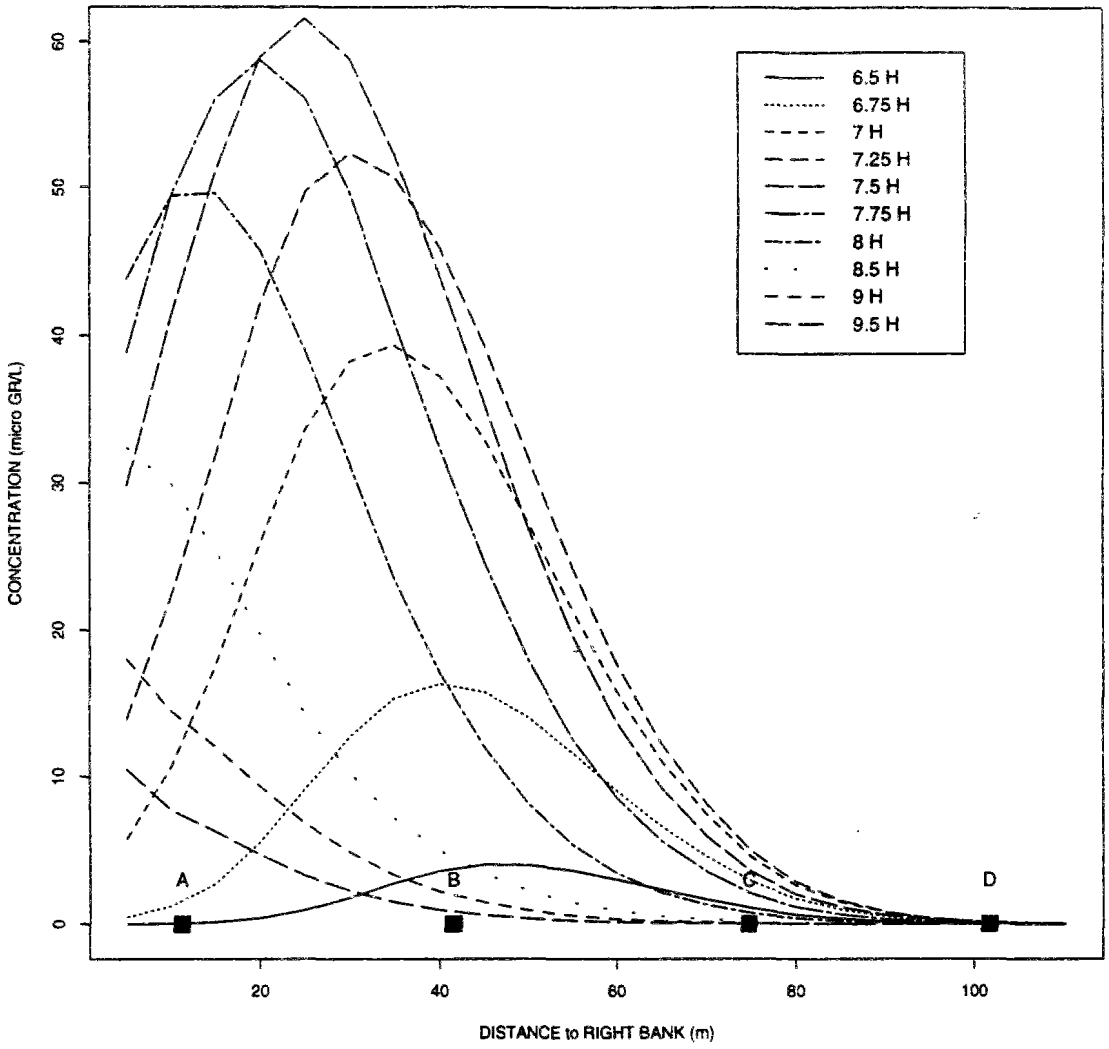


Figure 11.20: Temporal evolution of the dye repartition across the Seine at St-Ouen bridge

3. **St-Denis bridge**

This bridge is located within the large bend described by the Seine in the second half of the Clichy-Epinay reach. Measures show that transverse homogeneity has made considerable progress. The ratio of concentrations along the injection bank (figure 11.21) and along the opposite (left) one (figure 11.23) equates now merely 2 whereas, in upstream monitoring sections, no dye was detected close to the left bank. Similarly, remaining differences between right bank (point A, fig 11.21) and central (points B to D, fig 11.22) rhodamine levels are feeble.

Now, let's look at the model forecasts. In previous sections, agreement was good. But from now on, things are getting a little more complicated.

At the two preceding monitoring sections, the best agreement was reached when setting α_y between 0.4 and 0.6. Forecasts obtained with these values at St-Denis bridge are not satisfying : while results at point A are still correct (fig 11.21), such is not the case in the central part of the river (points B to D, fig 11.22). There, forecast pollutograms display too important extrema and are too early with respect to observed ones. Can this be ascribed entirely to an overestimation of the discharge ? Further trials demonstrate the opposite.

In fact, values of α_y in the range [0.4, 0.6] are recommended for fairly straight reaches, while the Seine describes a curve beginning somewhat upstream of St-Denis bridge. In curves, α_y has been reported to take much higher values (2-3, even more). *We studied what occurs when we change α_y value after the St-Ouen road bridge, increasing it to 1 or 2 in the rest of the right arm. While not canceling all discrepancies, this clearly improves the agreement between simulated and observed pollutograms.* Notably, when a stronger transverse diffusion spreads more evenly the dye cloud across the river, so that the dye is less concentrated in the river central, fastest flowing part, observed and forecast transit times turn to be closer. On the other hand, concentrations along the left bank become overestimated (figure 11.23).

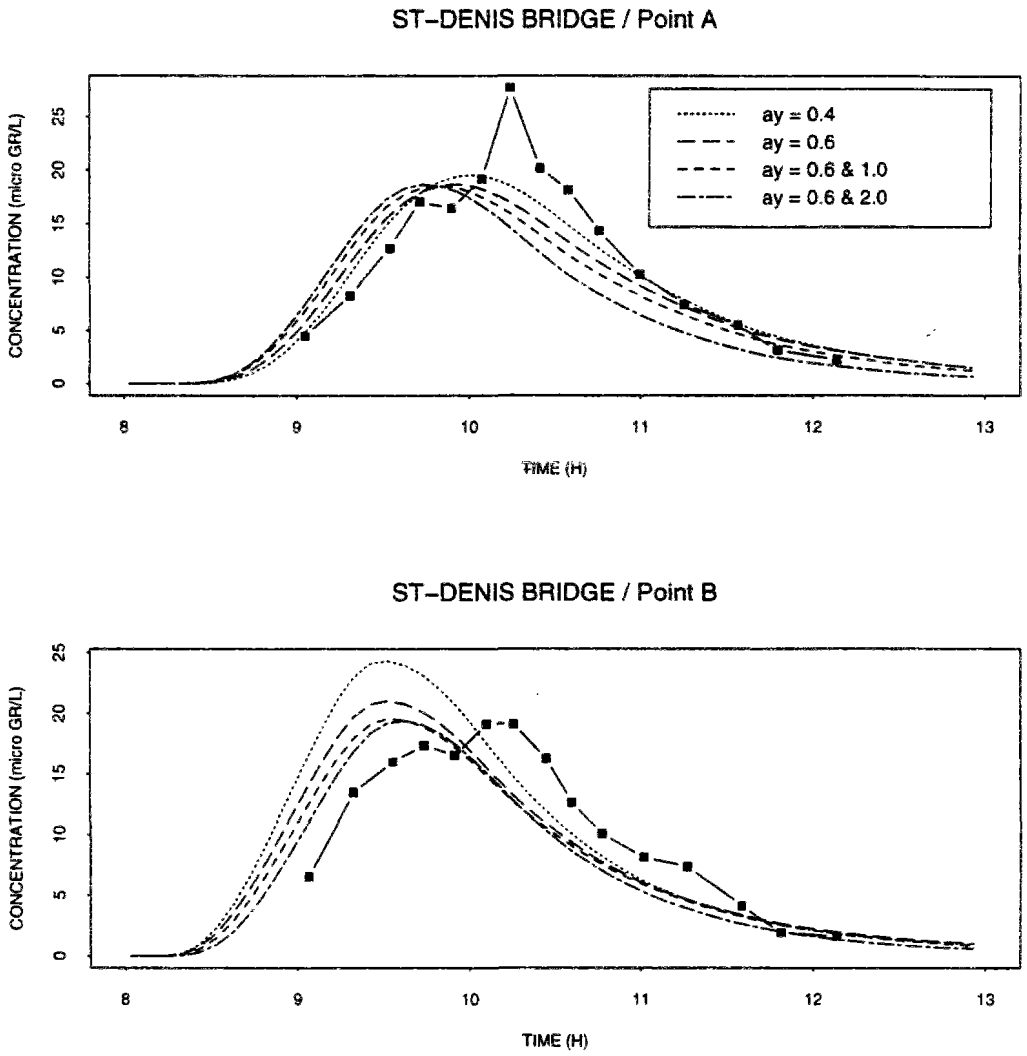


Figure 11.21: Measures and forecast pollutographs at St-Denis bridge (11.2 and 34.4 m off right bank)

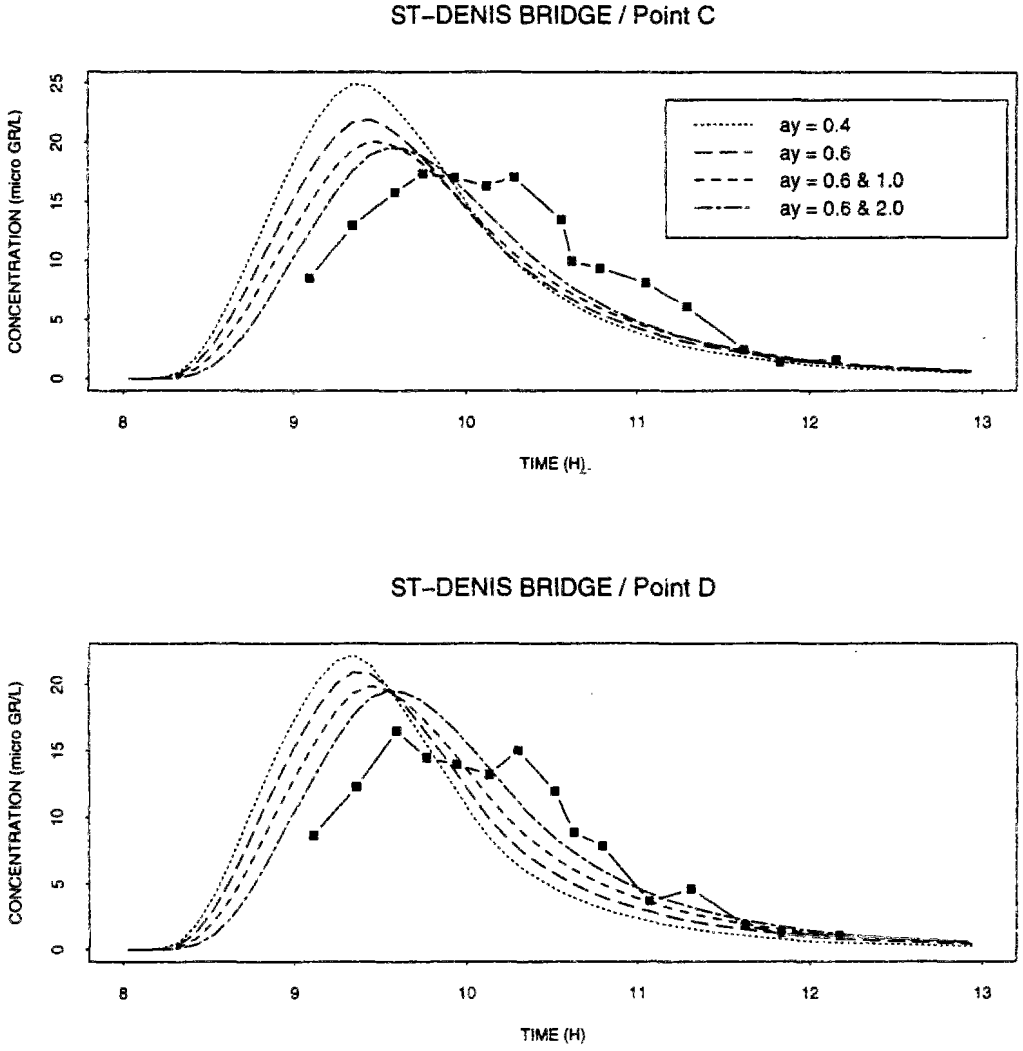


Figure 11.22: Measures and forecast pollutographs at St-Denis bridge (47.1 and 57.9 m off right bank)

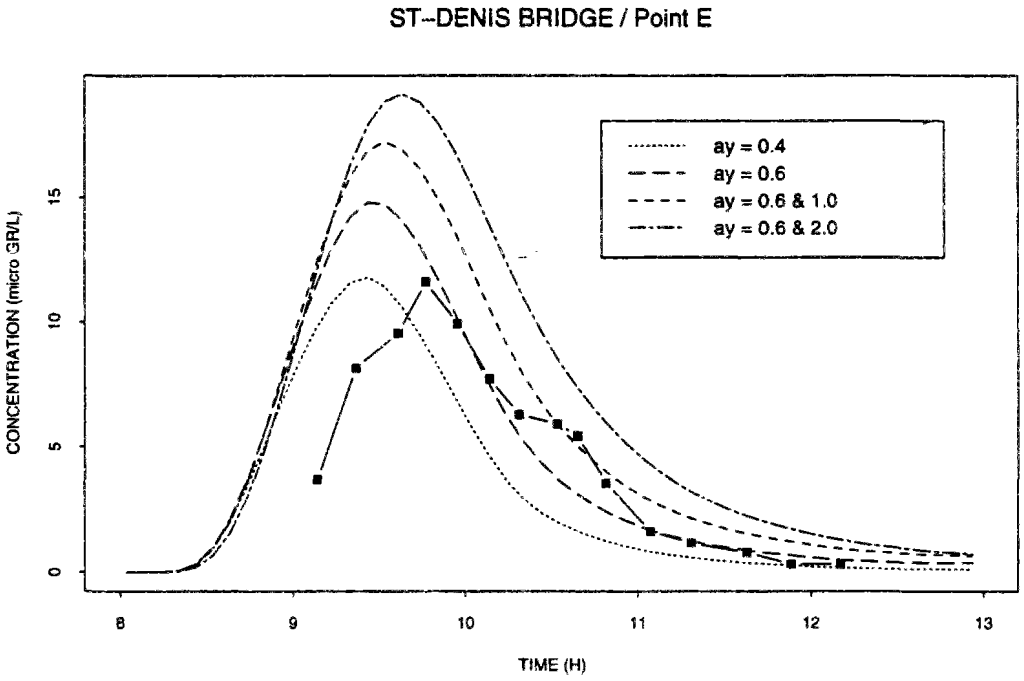


Figure 11.23: Measures and forecast pollutographs at St-Denis bridge (82.5 m off right bank)

4. Analysis of the dye cloud development

We shall now follow the development of the dye cloud from another point of view. Instead of observing the temporal evolution of concentrations at fixed locations, we shall look at the dye spatial repartition at given times, as it is forecast by the model.

Figures 11.24 to 11.34 are snapshots of the dye cloud, taken at different times, in different parts of the experimental reach. In each figure is included a small map of the reach, which helps to locate what part (drawn in red) is displayed. Seven ranges of concentration, evenly distributed, are visualized : the darker shade of red corresponds to areas where the concentrations exceed $\simeq$ 85 % of the cloud peak concentration while the lightest blue indicates concentrations inferior to 15 % of this maximum. The arrows give the respective scales in X and Y directions. Indeed, in some figures we used a smaller scale across the river than along its longitudinal axis, in order to provide a better glance at the dye spreading.

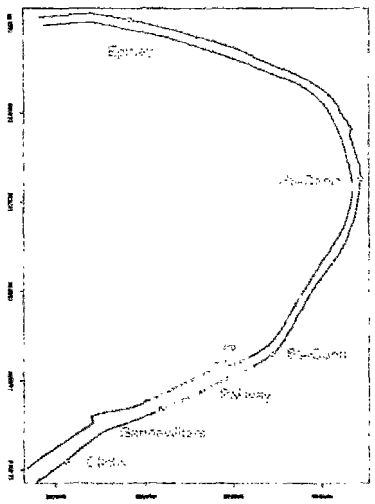
(a) Figures 11.24 and 11.25 represent the extent of the rhodamine cloud respectively at 6h30 and 7h, setting $\alpha_y = 0.6$. They illustrate the dye transfer in the first, straight, part of the experimental reach. We notice that the cloud remains "stuck" to the injection bank and that transverse concentration gradients are large. (nb : for smaller values of the coefficient α_y, the shape of the dye cloud is similar, except that the cloud is somewhat narrower and the transverse gradients even larger)

(b) In fact, the repartition of the dye across the river really evolves only as the cloud enters the bend (cf figure 11.26, corresponding to a simulation performed with $\alpha_y = 2$). There, the dye detaches itself from the bank and spreads all over the river section.

(c) Once the dye cloud has entered the mainstream, the areas of strongest concentrations are located no longer along the right bank but in the middle of the river, where velocities are the largest (cf figures 11.28, 11.30 and 11.32).

(d) The detachment of the rhodamine cloud at the bend entrance is forecast whatever the proportionality coefficient α_y and the subsequent magnitude of the transverse diffusion. This can be checked on figures 11.27, 11.29 and 11.31 (obtained when setting α_y to the lesser value of 0.6). Yet, we may note that, with feebler diffusion, the shift of the maximum concentrations from the bank to the middle of the river occurs a bit lately (fig 11.29 vs fig 11.28) and, of course, transverse gradients are more marked. On the other hand, the dye front appears to progress at a greater speed.

(e) Finally, let's have a look at the rhodamine as it flows under Epinay bridge. On figure 11.33 ($\alpha_y = 2$), we can observe that the cloud has taken the typical shape forecast by dispersion theory. Once again, we note that, when transverse diffusion is decreased ($\alpha_y = 0.6$, figure 11.34), the model predicts the persistence of stronger concentration gradients and a quicker displacement of the bulk of the dye towards the downstream end of the reach.



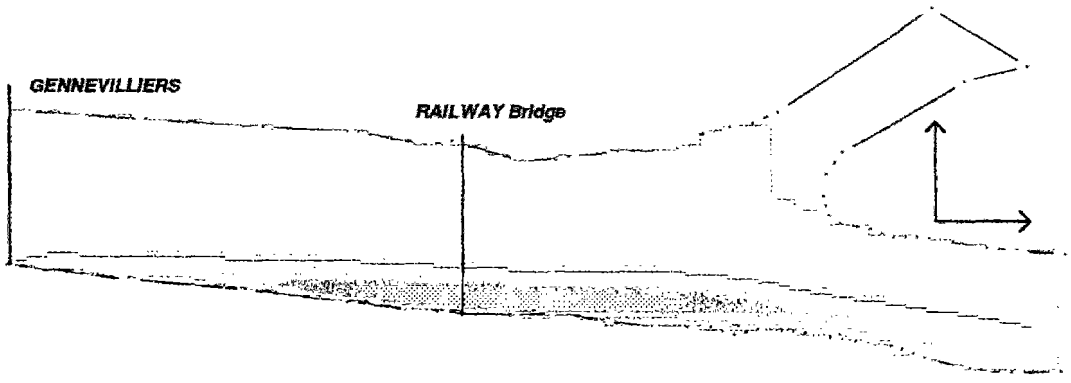


Figure 11.24: Rhodamine levels at $t = 6.5$ h, forecast with $\alpha_y = 0.6$

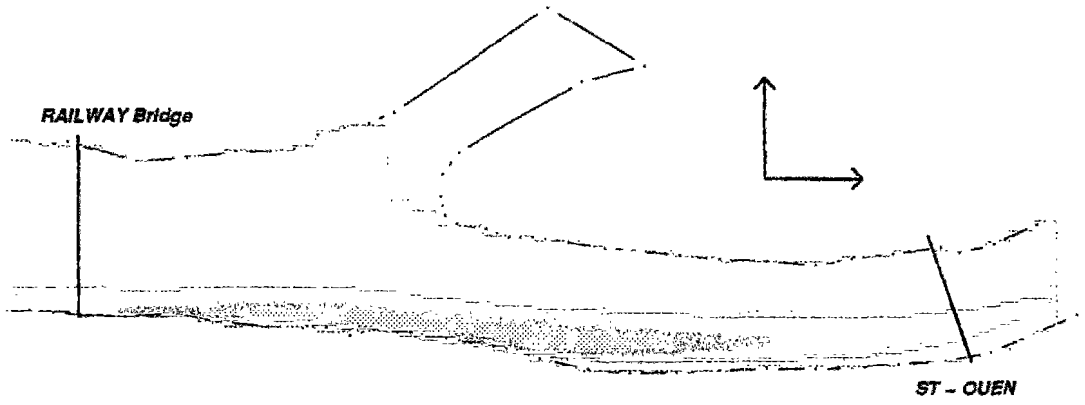


Figure 11.25: Rhodamine levels at $t = 7$ h, forecast with $\alpha_y = 0.6$

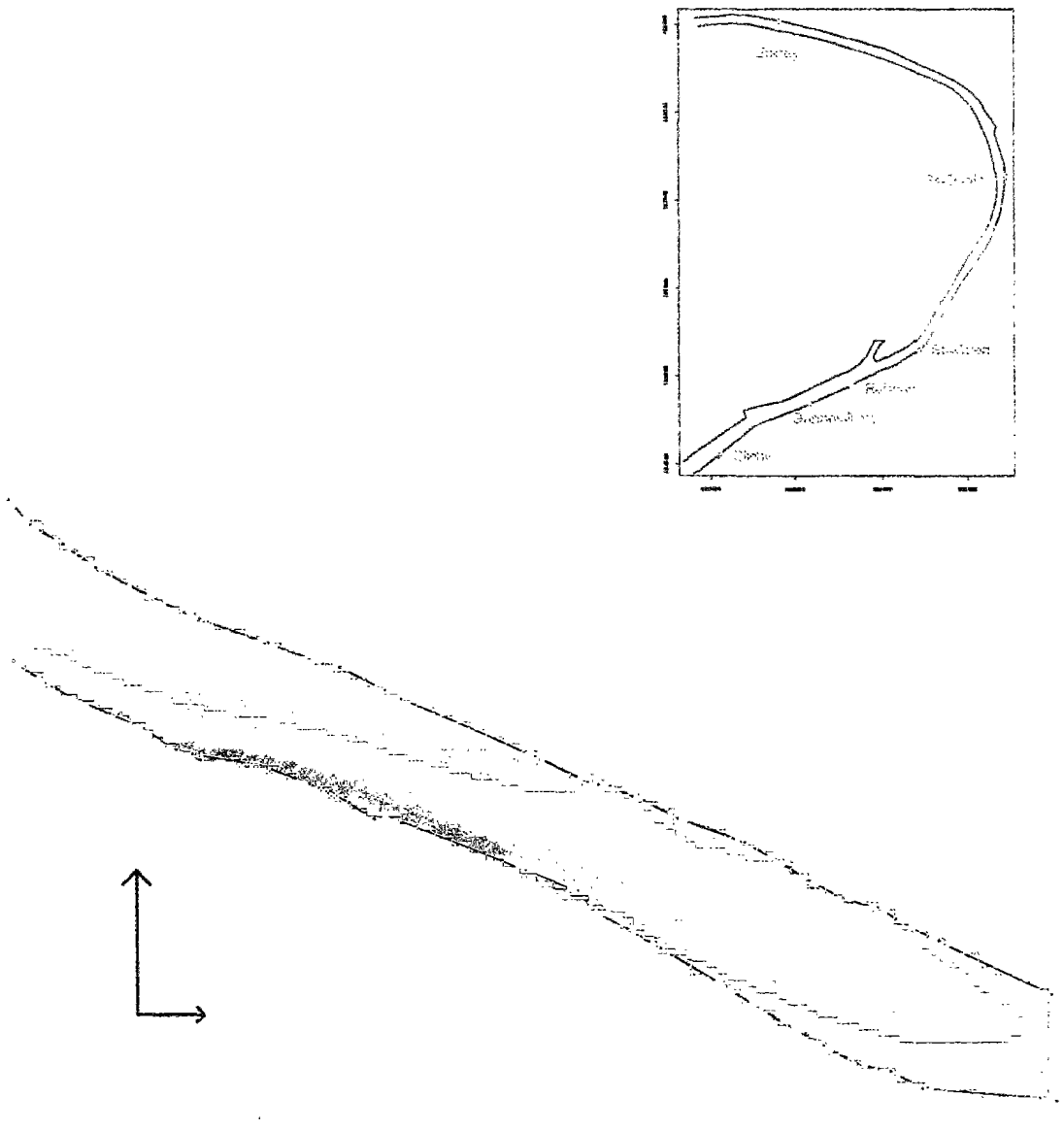


Figure 11.26: Rhodamine levels at $t = 8.5$ h, forecast with $\alpha_q = 2.0$

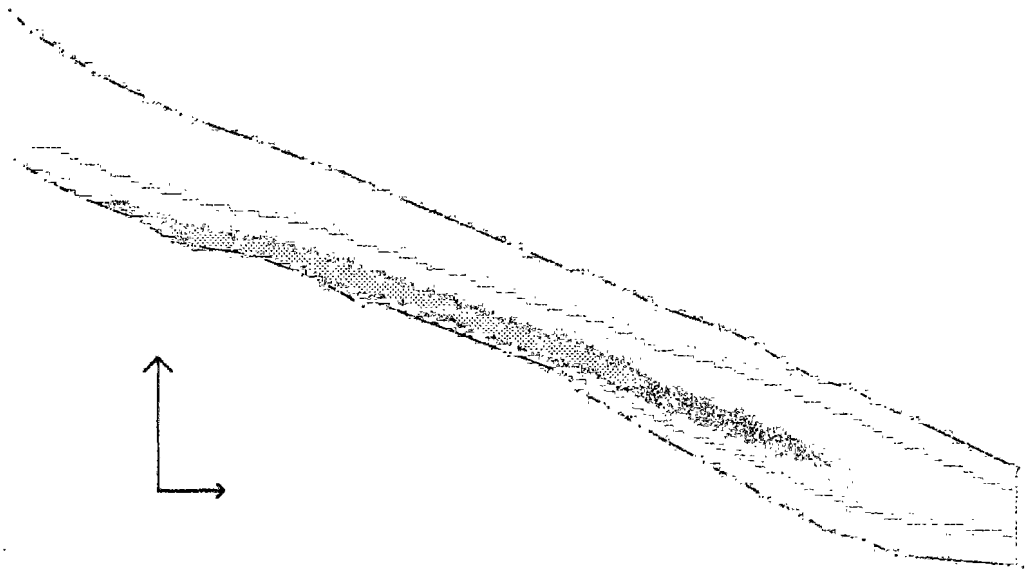


Figure 11.27: Rhodamine levels at $t = 8.5$ h, forecast with $\alpha_y = 0.6$

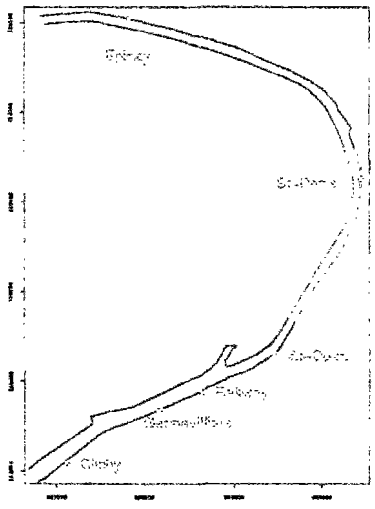


Figure 11.28: Rhodamine levels at $t = 9$ h, forecast with $\alpha_y = 2.0$

Figure 11.29: Rhodamine levels at $t = 9$ h, forecast with $\alpha_y = 0.6$

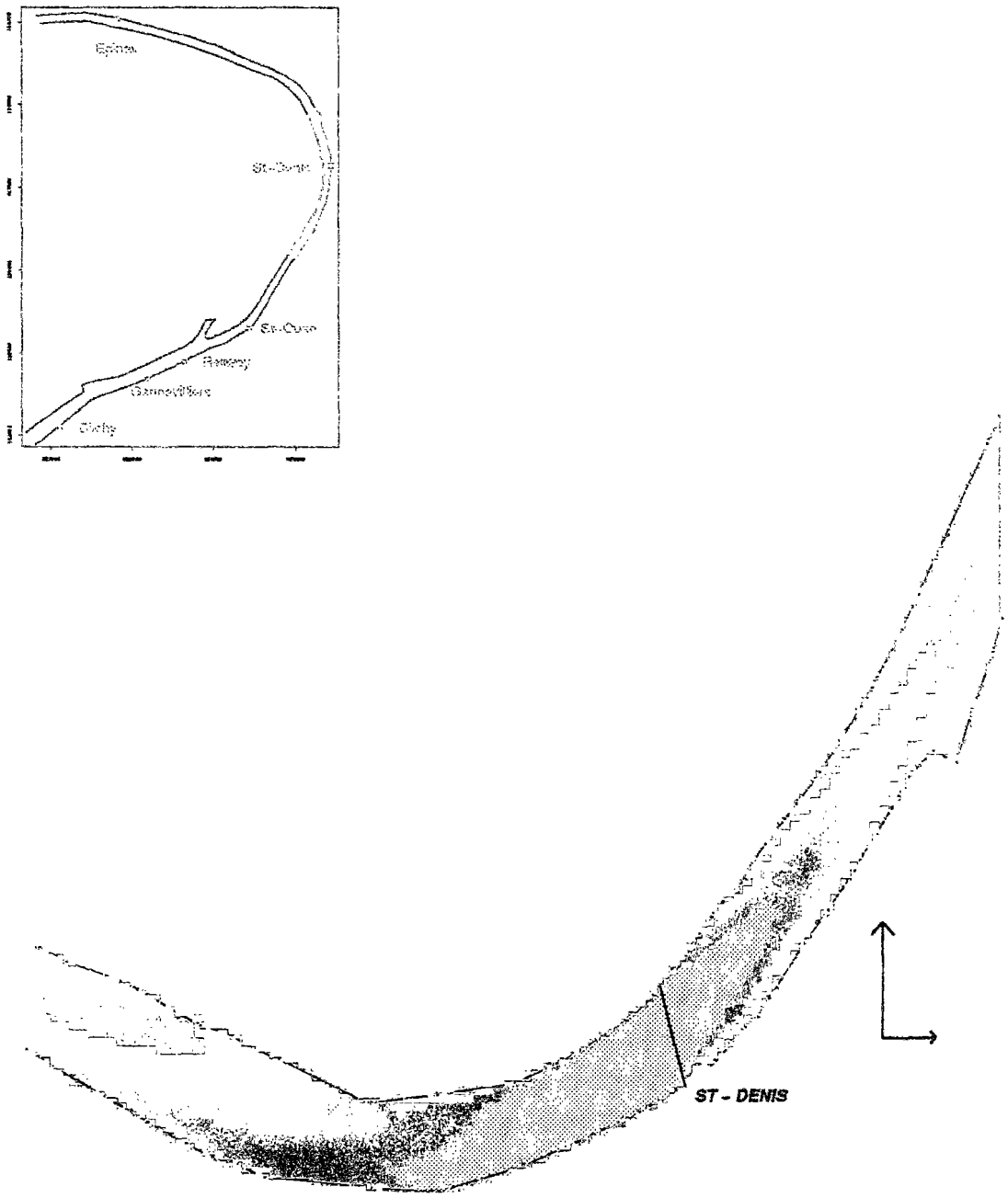


Figure 11.30: Rhodamine levels at $t = 9.5$ h, forecast with $\alpha_y = 2.0$

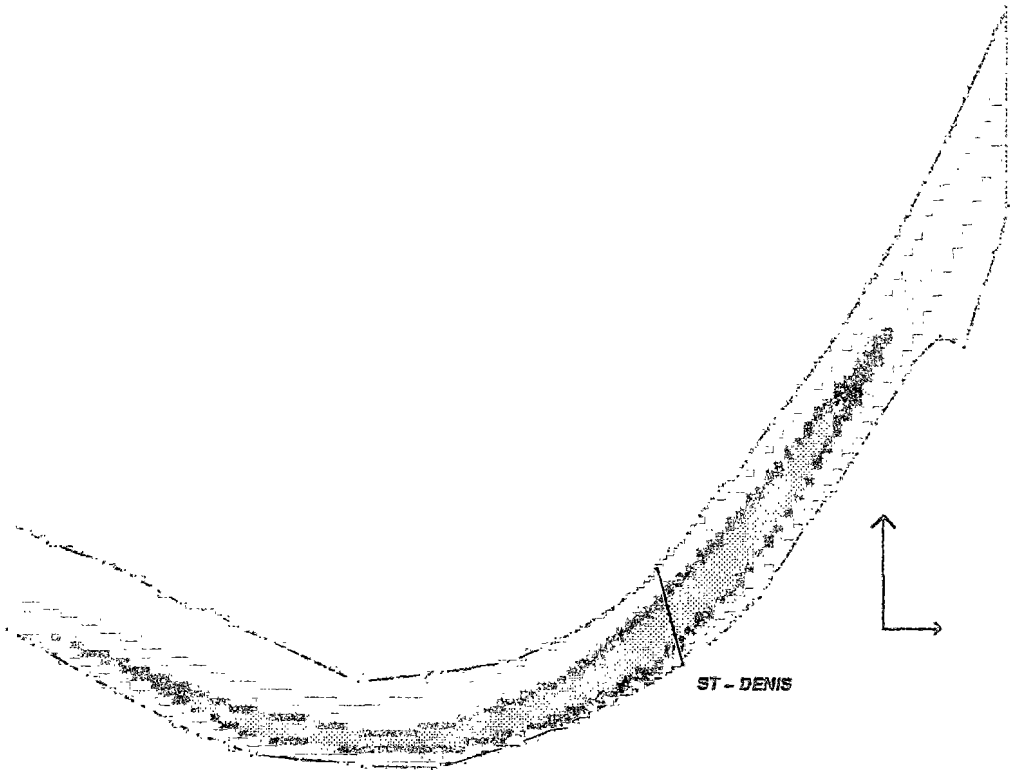


Figure 11.31: Rhodamine levels at $t = 9.5$ h. forecast with $\alpha_y = 0.6$

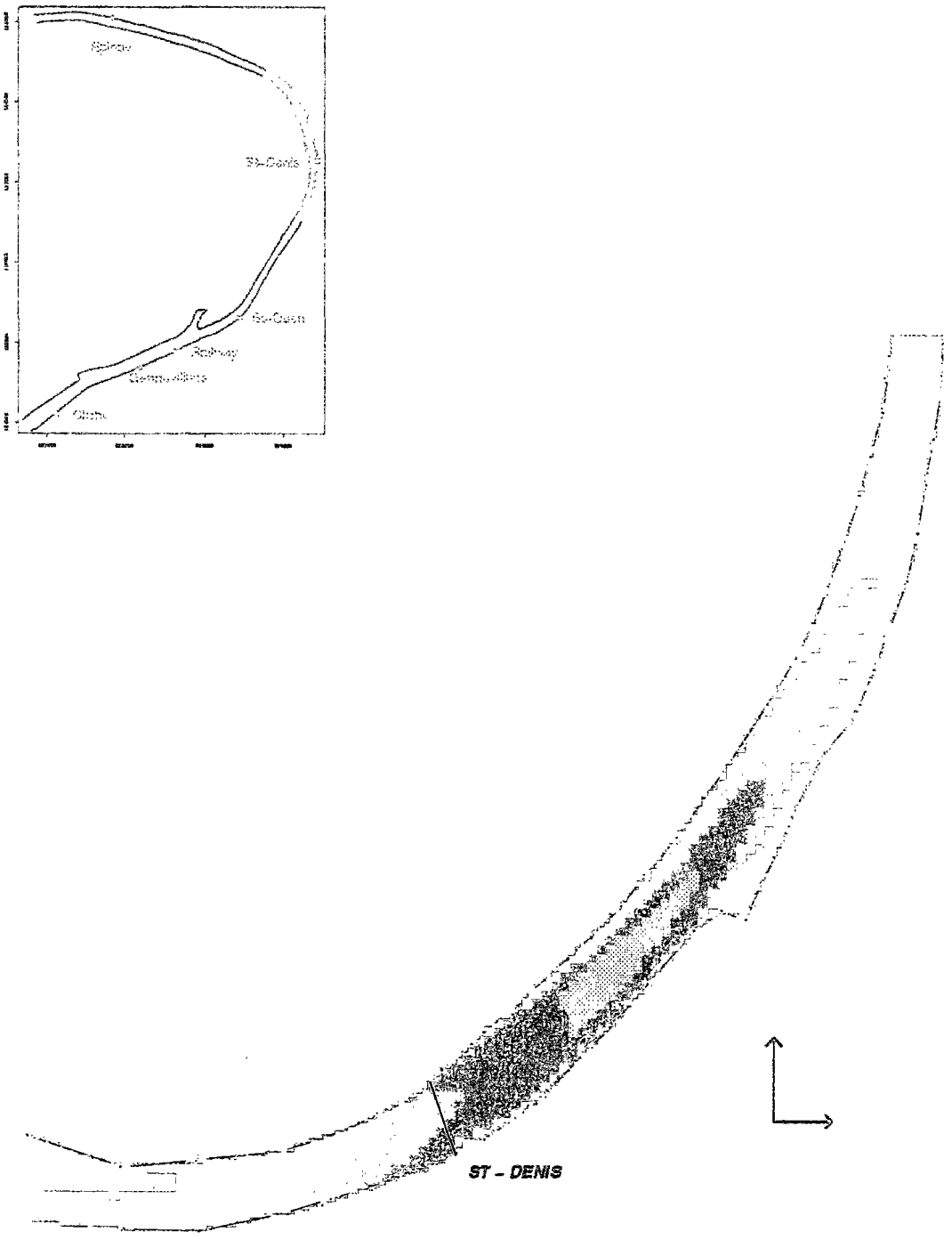


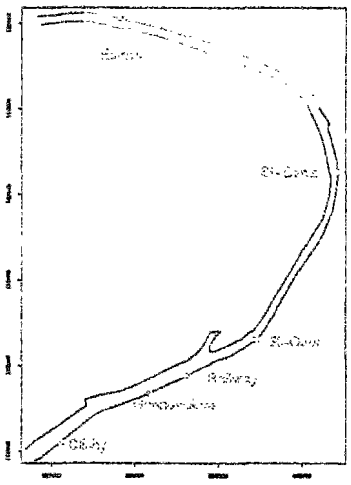
Figure 11.32: Rhodamine levels at $t = 10$ h, forecast with $\alpha_y = 2.0$

5. **Epinay bridge**

The comparison of simulations to observations at this bridge confirms that maintaining α_y to the 0.6 value typical of fairly straight reaches is inadequate. With this coefficient, the dye appears to arrive much too early at the bridge. Concentrations in the central part of the river are also overestimated. As could be guessed when comparing dye isolines corresponding respectively to $\alpha_y = 0.6$ (fig 11.34) and $\alpha_y = 2$ (fig 11.33), the phase shift is reduced when α_y value is raised. For $\alpha_y = 2$, the advance of forecast over observed pollutographs in the river central pass is approximately 30 minutes (figures 11.35 and 11.36) : this represents 5 % of the total transit time since the dye injection. This error is in the same range than the uncertainty about the measure of the flow rate.

However, increasing the transverse diffusion brings but a very slight improvement as regards the reproduction of the dye arrival along the banks (point A, upper half fig 11.35, and point E, fig 11.37). This is probably due to inaccuracies plaguing the estimation of the velocity field : perhaps the size of the computational grid mesh ($\simeq$ 5 m across the section) is too large to allow a really fair description of velocity gradient in the immediate vicinity of banks.

Finally, with α_y set to 2, both peak concentrations and longitudinal spreading of the cloud are well approximated.



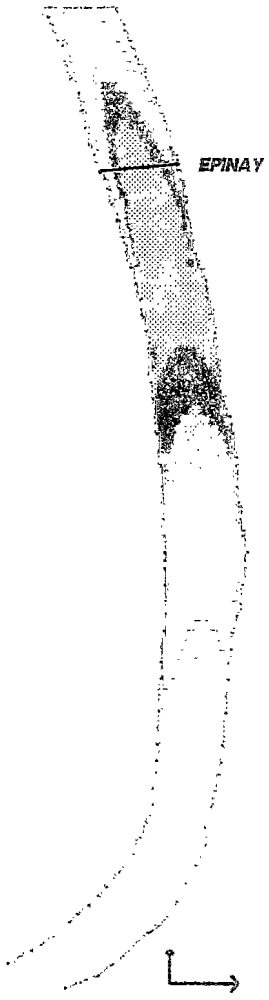


Figure 11.33: Rhodamine levels at $t = 13$ h, forecast with $\alpha_v = 2.0$

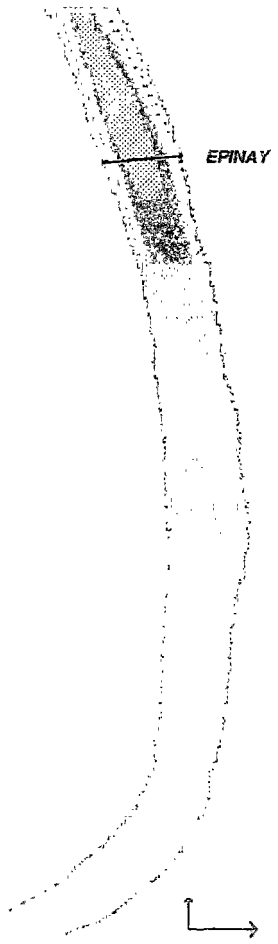


Figure 11.34: Rhodamine levels at $t = 13$ h, forecast with $\alpha_y = 0.6$

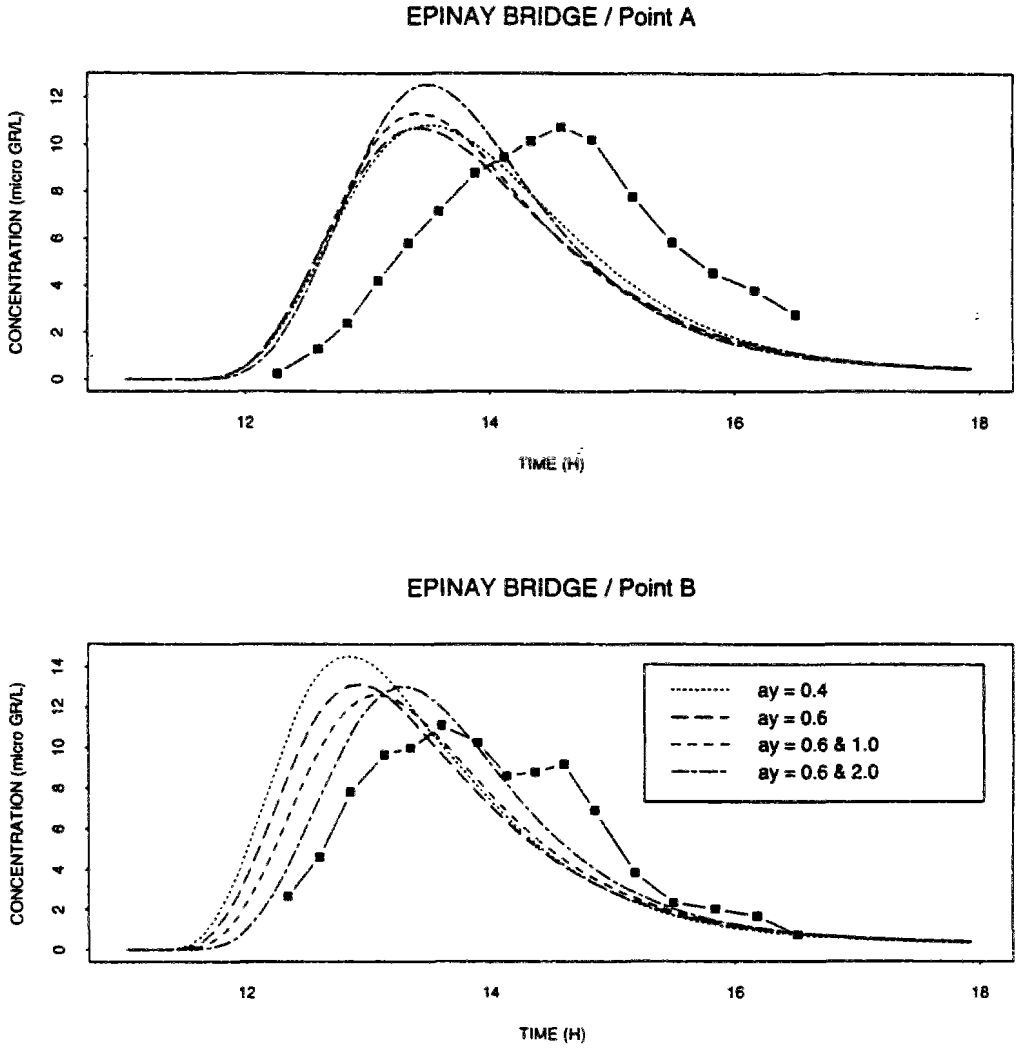


Figure 11.35: Measures and forecast pollutographs at Epinay bridge (13 and 36 m off right bank)

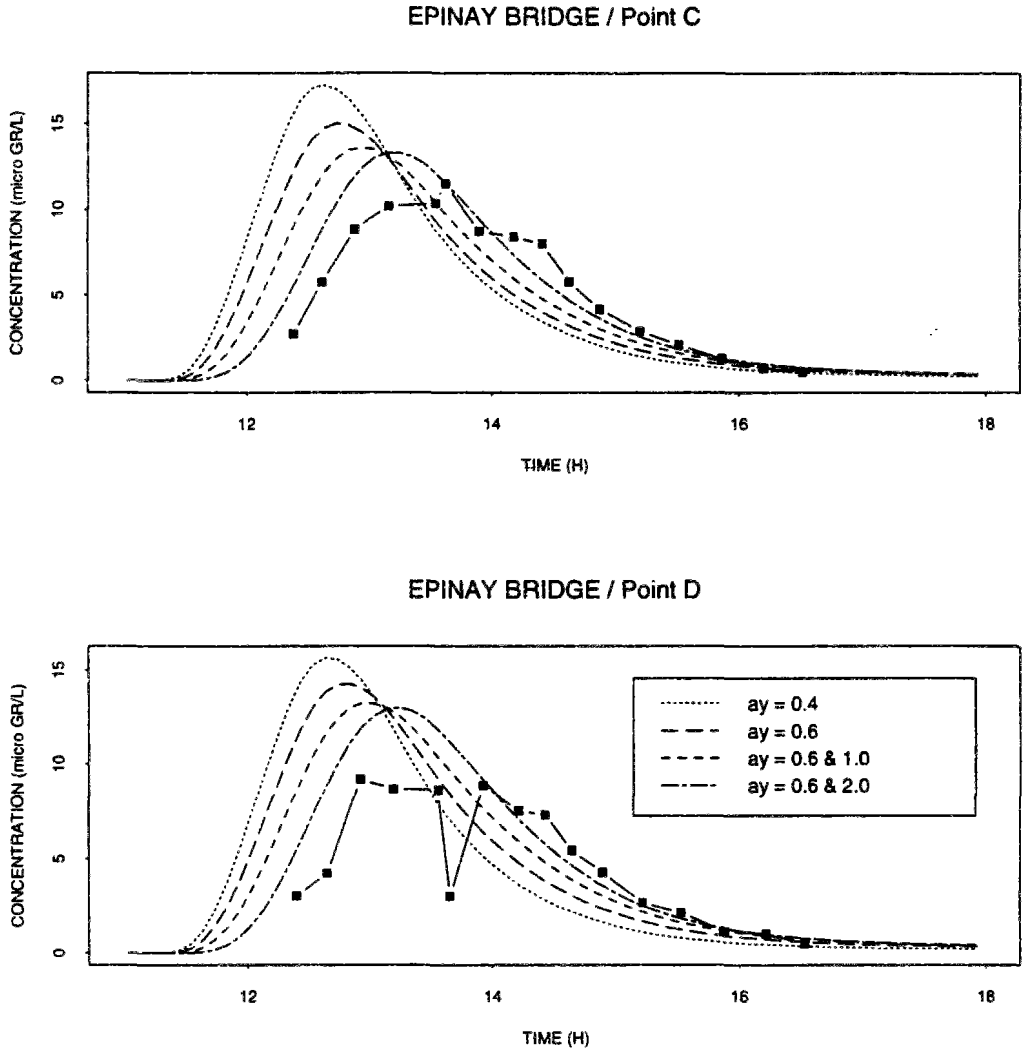


Figure 11.36: Measures and forecast pollutographs at Epinay bridge (55.5 and 75 m off right bank)

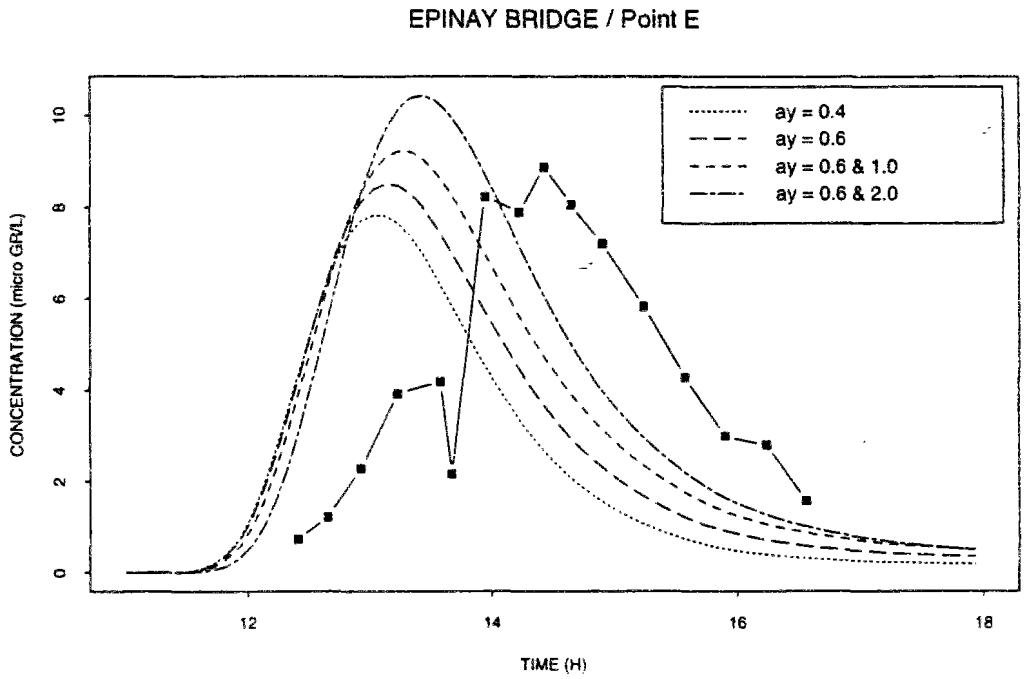


Figure 11.37: Measures and forecast pollutographs at Epinay bridge (98.5 m off right bank)

11.6 Conclusions of the Seine River application

The comments detailed in the above section 11.5.2 can be summarized as follows :

1. **We obtained unfailingly a good approximation of the dye cloud longitudinal spreading, without any introduction of some longitudinal diffusion.** In the kind of situations we are studying, longitudinal spreading results mainly from the differential advection phenomenon. Thus, this good agreement between forecasts and measurements is an indirect confirmation that the computed velocity field is fairly correct : the forecast transverse gradients of velocity cannot be too far from the real ones.

   Of course, everything is not perfect ! We observe a systematic advance of forecast over monitored pollutographs throughout the Seine right arm off St-Denis island. It is thus highly probable that there is some error in our estimation of the flow conveyed by this arm. However, this error is, as regards the central part of the river, in the same range than incertitude plaguing the flow measurements themselves.
   Finally, the phase shift in the vicinity of the banks could probably be reduced should we switch to a finer discretization allowing a better approximation of velocity gradient at these solid boundaries.

   We may assume that these encouraging results are due both to the soundness of the various algorithms embedded in our model (which were extensively tested !) and to the care devoted to the constitution of a comprehensive bathymetric data base.

2. **The good agreement between forecasts and measures** (as regards transit times, spreading, maximum concentrations) **is observed when setting the transverse diffusivity to values typically recommended in the literature.**

   We could not find an unique, uniform, value of calibration coefficient α_y (ratio of transverse diffusivity by product U^*h) able to generate accurate forecasts over the whole reach. Is it a failure ? No ... this merely outlines that relating transverse diffusivity only to bed shear induced turbulence is not sufficient. In the first half of the reach, which is rather straight, the adequate α_y is somewhere between 0.4 and 0.6 : this is consistent with values reported in the literature. In the second, curving, half, where transverse gradients of velocity are more marked, α_y must be increased in order to maintain agreement with observations. Once again, the most adequate value ($\alpha_y \simeq 2$) is in the range identified by previous experiments in bends.

3. In conclusion, **relying on a velocity field produced by a complete two-dimensional model reduces the need for calibrating the transport model.** Indeed, in order to obtain accurate forecasts, we needed to adjust only the transverse diffusivity ϵ_y. We not even had to look for a very precise tuning of this parameter.

Since less effort is required to calibrate this flow and transport model, we can assume subsequently that its predictive power will exceed that of simpler models such as KALPLAN or TULIPE, where the longitudinal diffusivity notably appeared to vary rather wildly, without any clear dependency on the magnitude of flow and other hydraulic variables.

In short, this first full-scale application of our flow and transport model allows to validate its foundations, both from the numerical and the physical points of view. We have there a model which has been developed with an eye at forthcoming applications to the Seine but which applicability is by no means limited to this river only. The various tests performed as regards advection or propagation algorithms certainly provide a rather thorough assessment of the model potential, which should help other modellers to apply it.
However, much remains to be done :

- The most straightforward and easiest improvement would be to implement other formulations of the diffusivity. Notably, some relate the α_y coefficient to local velocity gradients, increasing it naturally when gradients are more important (Leclerc, 1994, personal communication). This could potentially both induce better forecasts and improve the power of prediction of the model, by further reducing the need for tuning coefficients.

- Secondly, we have not devoted a single thought to the optimization of the model application (choice of the discretization, of the time steps, appropriate initialization of the simulations). It is obvious that if the model is to be further applied, either to the interpretation of other dye-tracing experiments, or to the simulation of storm overflows, some progress must be made.

  Notably, while we used a 1D model to reduce the scope of the modelled reach, we must think about extending the discretization to cover the whole reach between Suresnes and Chatou. We must be aware that this task can be complicated by the fact that we are working with rectilinear coordinates, which are not ideally suited for covering meandering reaches.

  Besides, finding simpler, quicker ways to initialize the model under steady-state flow conditions should be a top priority.

  Finally, we can expect to run the hydraulic model with larger time steps but, due to the magnitude of wave celerity, these will probably remain smaller than the time steps best suited for computing dissolved compound transport. A slight modification of the code will allow us to compute simultaneously flow and transport using different time steps, thus applying costly particle backtracking only once in a while.

- Lastly, but not least, no model (from one to three-dimensional ones) can be expected to be fully predictive in this reach while we know so little about the left arm, which hinders a correct estimation of the flow repartition around the island. Unfortunately, achieving

soundings in this area does not depend on us ! Besides, realizing further gauging operations, while water levels at and between dams are thoroughly surveyed, could certainly contribute to a better assessment of the roughness coefficient.

These various improvements should be achieved taking into account more phenomena than dissolved transport only. As regards the Seine case at least, the flow and transport model will now be completed by biogeochemical parts. It is certainly interesting to look for a numerical accuracy as great as possible. However, in the meantime, it is without doubt more pragmactic to look for something which is reasonably accurate considering uncertainties plaguing various data required to model the fate of storm sewer overflows (for instance, what about the precise definition of pollutant fluxes entering the river ?).

There is finally another research subject we would like to suggest.
We are confident that the present model will prove to be a powerful tool for analyzing precisely the impact of storm sewer overflows (and other pollutant inputs, by the way). However, it is probably not adequate as a management tool. A management tool should ideally provide an unskilled user with immediate and straightforward answers concerning the change in the river conditions when different strategies are applied to reduce and/or treat pollutant inputs, considering the variability of the initial state of the receiving river, of the overflows, etc ... We fear that running hundreds of simulations corresponding to multiple scenarii with a full two-dimensional flow and transport model can be rather cumbersome, even if it is feasible ! Besides, in our opinion, to perform this task properly requires a minimum level of knowledge in hydraulics and numerical analysis.
We think that using for management purpose models with an intermediate level of complexity, such as stream tube model TULIPE, is an interesting compromise solution. TULIPE was able (Théry *et al.*, 1993a; Théry *et al.*, 1993b) to reproduce fairly dye tracing experiments, but its main shortcoming is the difficulty to relate its longitudinal diffusivity parameter to hydraulic conditions. This is certainly due to the fact that its basic hypothesis about the repartition of velocity across the river are not accurate enough. Relying on velocity fields forecast by the full two-dimensional model should help to correct errors linked to these hypothesis and consequently ease TULIPE calibration and enhance its predictive power.

11.7 Résumé français : "Application à un bief de Seine"

Comme nous l'avons indiqué au chapitre 2, le développement de notre modèle bidimensionnel plan d'hydraulique et transport de composés dissous, "ACHAB", s'est inscrit dans le cadre du programme PIREN-Seine. Cet outil est destiné à effectuer des analyses locales détaillées de la dilution de rejets ponctuels, notamment les déversements d'orages. Ces analyses devraient permettre une interprétation adéquate des différentes mesures effectuées par le PIREN-Seine. Ceci débouchera sur l'amélioration et la correction des prévisions issues des modèles de qualité monodimensionnels, lesquels apparaissent comme le seul choix pragmatique en matière d'outils d'aide à la gestion (la portion de Seine à modéliser s'étend de Montereau à Poses, respectivement 100 km à l'amont et 200 km à l'aval de Paris).

Nous présentons ci-dessous le premier exemple d'application de ACHAB à la Seine. La section 11.1 présente le bief étudié et la section 11.2 le traçage que nous interprétons. Les différentes étapes de cette interprétation constituent les sections 11.3 à 11.5.

Caractéristiques des bief et traçage étudiés (sections 11.1 et 11.2)

Le bief modélisé est situé en agglomération parisienne, en aval de la ville de Paris même. Il est compris entre l'île de la Grande Jatte et la pointe aval de l'île St-Denis (respectivement points kilométriques 22 et 33 dans le repère propre à la Seine). L'écoulement y est régulé par le fonctionnement des barrages de navigation de Suresnes (en amont, PK 17) et Chatou et Bougival (en aval, PK 45).

Le secteur inclut l'île St-Denis, dont le bras gauche n'a pas été pris en compte dans la modélisation. En effet, la bathymétrie y est inconnue, le chenal de la Seine n'étant relevé que dans les bras ouverts à la navigation des barges. Dans les portions ouvertes à la navigation les données sont par contre assez nombreuses et de bonne qualité. On a ainsi constitué (Simon, 1992) une base de données topographiques sur le bief modélisé comprenant en moyenne un profil en travers tous les 50 m (les points définissant les profils en travers étant espacés de 2 à 5 m). La Seine fait environ 150 m de largeur à l'amont de l'île, 80 à 100m de large dans le bras droit; la profondeur moyenne pour les conditions d'étiage (notamment lors du traçage) est de l'ordre de 5 m.

La Seine est traversée par des ponts. L'emprise de leurs piles sur la section mouillée est de 6 à 10 %, sauf en ce qui concerne le pont routier de St-Ouen (début du bras droit) dont l'emprise est de 19 %. Nous n'avons pas pris en compte les ponts dans notre représentation du bief.

Le bief a fait l'objet d'une expérience de traçage en Septembre 92, le traceur (rhodamine

B) étant injecté au droit du déversoir de Clichy (plus gros site de rejets pluviaux urbains dans l'agglomération). Les sections de mesure sont implantées au droit de ponts, afin de permettre un repérage facile des points de prélèvement. Elles sont situées respectivement à 1.2, 1.7, 2.7, 4.8, 8.1 km à l'aval de Clichy (ponts de Gennevilliers, SNCF, de St-Ouen, de St-Denis, d'Épinay) (cf figure 11.10). Au pont de Gennevilliers l'échantillonnage a été réalisé en surface et à 3m dessous, ce qui a permis de vérifier l'obtention de l'homogénéisation verticale du traceur (lequel est initialement injecté en surface). Aux ponts suivants, on échantillonne en surface uniquement. Il y a 4 à 7 points de prélèvement en travers. La fréquence des prélèvements, variable en fonction de l'étalement du nuage, va de 10 à 20 minutes.

Le traçage s'est déroulé pour un débit en Seine quasi-stationnaire de 170 m^3/s, 60 % du débit passant dans le bras droit de l'île. L'intégralité du nuage de traceur est passée dans le bras droit.

Interprétation hydraulique (sections 11.3 et 11.4)

Lors du traçage les débits ont été suivis avec fiabilité. Une estimation nous était fournie toutes les deux heures par les stations de mesure automatiques du Service de Navigation de la Seine (SNS). Ces mesures ont été corroborées par le jaugeage effectué en milieu de journée.
Par contre, les informations concernant la ligne d'eau sont pauvres : les cotes ne sont en effet relevées qu'aux barrages de navigation, quand il y a manoeuvre. On ne dispose ainsi que de la connaissance de la cote à trois instants dans la journée pour Suresnes et Chatou (le réglage de Bougival n'ayant pas été modifié). En tout état de cause, la pente de la ligne d'eau est faible.

Les données hydrauliques ont tout d'abord été interprétées à l'aide d'un modèle St-Venant monodimensionnel, PROSE, (Even & Poulin, 1993)) (cf sections 11.3 et H.1).
On a testé différentes hypothèses de reconstitution de la bathymétrie du bras gauche. Compte tenu du substrat en Seine, un même coefficient de Strickler a été utilisé sur l'ensemble du secteur (hormis ce bras gauche). PROSE est calé en fonction des cotes enregistrées aux barrages de navigation de Suresnes, Chatou et Bougival. La validité du calage est évaluée en comparant les temps de transit calculés et observés pour le centre de gravité du nuage de rhodamine. **Le coefficient de Strickler optimal est de 36.**
Compte tenu du faible battement des cotes aux barrages et des débits (fluctuation d'une dizaine de m^3/s autour de la moyenne) il y a peu de différences entre les lignes d'eau et vitesses calculées pour différentes combinaisons de débits amont et réglage de barrages. Nous avons donc estimé qu'il était **justifié de faire l'interprétation du traçage pour une situation d'écoulement permanent.** Cette situation correspond à l'état intermédiaire de réglage des barrages.
Le secteur modélisé avec ACHAB est plus restreint que celui modélisé avec PROSE, puiqu'il

débute environ 2 km avant Clichy (extrémité aval de l'île de la Grande Jatte) et se termine à la pointe aval de l'île St-Denis. Grâce à PROSE, nous avons pu reconstitué les niveaux d'eau intermédiaires entre les barrages : les cotes à la Grande Jatte et de part et d'autre de l'île sont respectivement de 23.674, 23.658 et 23.605 m. C'est en fonction de ces valeurs qu'on calera ACHAB.

Nous allons maintenant détailler les **conditions de mise en oeuvre d'ACHAB** (cf 11.4.1 et 11.4.2) :

- **Maillage** On utilise un maillage orthogonal, après rotation du bief dans un repère incliné de 70 degrés par rapport aux coordonnées cardinales Lambert (ainsi le domaine a une forme proche d'un L majuscule). Les pas d'espace utilisés sont en moyenne de 5 m en travers de la rivière, 25 m dans le sens longitudinal, sauf au niveau de la courbe (de part et d'autre du pont de l'île St-Denis) où les pas d'espace passent progressivement à 5 m dans les deux directions. Ce choix de taille de maillage a été guidé par des considérations physiques : c'est ce qui permet a priori de bien "capter" les variations bathymétriques en Seine. A postériori, après interprétation du traçage (cf sec 11.5), on a constaté qu'un maillage encore plus fin au voisinage des berges (mailles de 1 à 2 m) aurait éventuellement été souhaitable (meilleure description du gradient de vitesse à la paroi).
  Le maillage final comprend environ 13000 noeuds.
- **Conditions limite et initiales**
  - **Frontières ouvertes** On impose un débit amont et sa répartition en travers. Cette répartition est obtenue par des conditions d'équilibre entre la pente de la ligne d'énergie et la pente des fonds. La vitesse est alors proportionnelle à une fonction puissance de la hauteur d'eau. On peut escompter que d'éventuelles erreurs liées à cette hypothèse de répartition de débit se seront estompées quand débute la zone qui nous intéresse, à Clichy (2 km à l'aval de la frontière amont). Les cotes sont imposées à l'amorce du bras gauche ainsi qu'à l'aval du bras droit de l'île St-Denis.
  - **Frontières fermées** On y applique des conditions d'imperméabilité. Les niveaux d'eau à la paroi sont contraints (gradient normal à la paroi nul). Sur les frontières du domaine de calcul parallèles à un des axes de coordonnées, les vitesses normales sont annulées. Les vitesses aux angles saillants et tangentielles sont laissées libres.
  - **État initial** Le régime permanent recherché est obtenu à l'aide d'une simulation instationnaire. On part d'un plan d'eau au repos, avec un débit nul. Ce débit est progressivement augmenté à 170 m^3/s, puis stabilisé, et la simulation poursuivie jusqu'à ce qu'une solution stable soit atteinte. Ceci réclame de 20 à 30 minutes de calcul sur station SUN SPARC 10.
- **Hypothèses supplémentaires** Avant de procéder aux simulations avec ACHAB, on s'est livré à une exercice d'évaluation a priori des poids respectifs des différents termes intervenant dans les équations de St-Venant, ce grâce à l'analyse adimensionnelle (section 11.3.3). Il ressort que, **en Seine, les termes prépondérants sont les termes de propagation. Advection et frottement sont du même ordre de grandeur (100 fois plus petit que celui des termes**

de propagation). Les termes de dispersion horizontale de moment (évalués d'après les ordres de grandeur des coefficients de dispersion fournis par de précédents traçages) **seraient une dizaine à une vingtaine de fois plus petits que advection/frottement. En mettant en oeuvre ACHAB, on a donc purement et simplement négligé les termes de dispersion.** Le seul terme dissipatif des équations se retrouve donc être le frottement au fond.

Avec ACHAB, la ligne d'eau est reconstituée avec une précision suffisante pour un coefficient de Strickler K_s de 40 $\pm$ 10 % (section 11.4.3). Il semble délicat d'estimer le Strickler avec une précision supérieure : la pente de la ligne d'eau est faible, et l'incertitude sur sa mesure relativement élevée.
Les champs de vitesse et hauteurs obtenus pour les Strickler voisins de 40 (à $\pm$ 10 % près) sont trés proches. En fait, l'incertitude que l'on peut avoir sur le Strickler va se répercuter surtout sur le calage ultérieur du modèle de transport dissous.
Dans ACHAB, on exprime les coefficients de dispersion d'un composé dissous en fonction de la hauteur d'eau et de la vitesse de frottement au fond u^* : on cale les coefficients de proportionnalité entre diffusivités et produit $u^* h$. Compte tenu de la relation entre u^* et le Strickler, une incertitude de 20 % sur celui-ci entraîne une incertitude équivalente sur u^* et donc sur le coefficient de proportionalité !

Les figures 11.11 à 11.15 présentent le champ de vitesses obtenu. On notera que leur allure est tout à fait correcte. Les profils de vitesse sont extrêmement plats à l'amont de l'île (vitesse quasiment uniforme en travers). Ils revêtent une forme parabolique plus marquée dans le bras droit, plus étroit, où l'écoulement est d'ailleurs globalement plus rapide.

Interprétation du traçage (section 11.5)

Les conditions de mise en oeuvre du modèle de transport dissous sont les suivantes (cf section 11.5.1) :

- **Mode de résolution** Compte tenu de la faiblesse attendue des termes de dispersion, nous avons employé pour l'étape d'advection la méthode aux caractéristiques bicubiques de Rasch-Williamson, qui est bien adaptée aux cas d'advection dominante.
- **Condition limite amont** Dans le champ proche de l'injection, la répartition du traceur sur une verticale n'est pas homogène, ce qui rend difficile l'utilisation d'un modèle bidimensionnel plan. Nous avons donc décidé de ne simuler le transit du traceur qu'à partir du pont de Gennevilliers, où les mesures nous indiquent que l'homogénéisation est atteinte. Le profil de concentrations amont est obtenu par interpolation entre les points de mesure à Gennevilliers.
- **Formulation des diffusivités** Comme indiqué ci-dessus, on paramètre les diffusivités en fonction du produit hauteur d'eau par vitesse de frottement au fond.

En ce qui concerne la diffusivité en travers, le coefficient de proportionnalité α_y est compris, d'après la littérature, dans les gammes : (i) 0.1 à 0.3 pour des canaux rectilignes de laboratoire ou artificiels (par exemple, d'irrigation) (ii) 0.4 à 0.8 (moyenne recommandée de 0.6) pour des rivières légèrement irrégulières ou méandrées (iii) 1 à 4 pour canaux ou rivières présentant des courbes, méandres, accusés.

Pour la diffusivité longitudinale, le coefficient de proportionalité α_x devrait être aux alentours de 6. L'ordre de grandeur de cette diffusivité étant alors estimé, il apparaît qu'en Seine ses effets devraient être mineurs par rapport à ceux de la convection différentielle. **Nous avons donc décidé de négliger purement et simplement la diffusion longitudinale.** Ceci nous permettra de tester indirectement la cohérence des profils de vitesse calculés. S'ils sont corrects, l'étalement longitudinal du nuage de traceur devrait en effet être bien reproduit.

Nb : Nous n'avons pas essayé d'obtenir un ajustement optimal des diffusivités. Nous avons uniquement vérifié que leur gamme est cohérente avec ce que l'on recommande. Ceci justifie que nous ayons eu recours essentiellement à des comparaisons visuelles entre observations et simulations.

Passons en revue les sections de mesure les unes après les autres (section 11.5.2) :

1. **Pont SNCF** C'est le premier point possible de comparaison. On remarque d'après les mesures que le traceur reste trés concentré le long de sa rive d'injection, la rive droite. En effet, on n'observe que le passage de quelques traces au milieu de la rivière (concentrations de l'ordre de 2 μg/l au maximum, soit 40 fois plus faible qu'à 10 m de la rive droite).
   On note (figures 11.16 et 11.17) qu'un bon agrément est obtenu pour des valeurs de α_y comprises entre 0.23 (valeur moyenne pour canaux rectilignes) et 0.6 (valeur moyenne pour rivières régulières quasi-droites). Les concentrations maximales sont bien reproduites, ainsi que l'étalement dans le temps des pollutogrammes. Il y a cependant un léger déphasage (retard des pollutogrammes calculés).

2. **Pont routier de St-Ouen** La rhodamine reste encore concentrée dans la partie droite (les 2/3) du fleuve. La concordance est toujours correcte (figures 11.18 et 11.19) . Elle est cette fois-ci nettement meilleure pour $\alpha_y = 0.6$. L'étalement des pollutogrammes est toujours bien reproduit. On note cependant, cette fois-ci, une légère avance.

3. **Pont de St-Denis** Ce pont est situé au sein du large méandre décrit par la Seine dans la seconde moitié du bief Clichy-Epinay (la première moitié étant quasiment rectiligne). On remarque tout d'abord sur les mesures que l'homogénéisation en travers a considérablement progressée. Il n'y a plus qu'un ratio de 2 entre concentrations en berge droite et en rive opposée, alors que jusqu'ici on n'avait pas observé de présence de rhodamine en rive gauche. De même, il n'y a plus guère de différence entre rive droite (point A, fig 11.21) et zone centrale de la rivière (points B à D, fig 11.22).

On présente les résultats de simulation obtenus pour α_y constant à 0.4 ou 0.6, ainsi qu'en faisant passer α_y à 1 ou 2 (valeurs plus conformes à ce qui est recommandé en courbe) à partir du moment où le méandre s'amorce. Le deuxième type de simulation reproduit clairement une situation plus proche des observations. Avec α_y porté à 2, le modèle surestime un petit peu l'homogénéisation transverse, et donc les concentrations en rive gauche (fig 11.23). Aux autres points de mesure, concentrations et étalement longitudinal sont bien approximés. Le décalage temporel (avance) entre calculé et observé est faible en partie centrale. Il est d'autant plus prononcé que la diffusivité transverse est fixée à un niveau faible.

En fait, le comportement du nuage de traceur peut être analysé d'un autre point de vue, en faisant des "photos" du nuage sur l'ensemble du bief à différents instants (figures 11.24 à 11.34). On remarque que tant que le traceur reste dans la première moitié du bief, rectiligne, il reste "scotché" à la rive d'injection (fig 11.24 et 11.25). La situation évolue quand le nuage aborde le méandre (fig 11.26). Il tend alors à se décoller de la rive d'injection. Selon la valeur de la diffusivité transverse, les zones de plus forte concentration restent ensuite confinées dans la partie centrale, où le courant est le plus fort, ou gagnent l'ensemble de la rivière (comparer par exemple la figure 11.29, pour $\alpha_y = 0.6$, à la figure 11.28, pour $\alpha_y = 2$).

4. **Pont d'Épinay** Voyons la forme du nuage à l'extrémité aval, au pont d'Épinay. Avec un coefficient à 2, on voit qu'il prend vraiment la forme caractéristique prédite par la théorie de la dispersion (fig 11.33). Les zones de concentration maximale se sont également déplacées moins vite qu'avec une dispersion plus faible (fig 11.34).

   On retrouve ceci sur les pollutogrammes. Encore une fois, niveaux de concentration et étalement longitudinal sont bien approximés (fig 11.35 à 11.37). Les pollutogrammes calculés sont systématiquement en avance sur les observés. Cette avance est réduite, dans la zone centrale, si l'on prend α_y égal à 2. Le décalage est alors de 30 minutes, ce qui, sur un temps de transit de 10 heures depuis l'injection, représente une erreur de 5 %. Compte tenu des incertitudes sur les fluctuations de débit c'est tout à fait acceptable. Cependant, le décalage perdure aux rives, quelque soit la valeur prescrite pour la diffusivité.

Conclusions de l'application (section 11.6)

Finalement, quel bilan peut-on tirer de ce traçage et de l'exercice d'interprétation correspondant ?

- A toutes les sections on a une bonne approximation de l'étalement longitudinal, sans avoir eu besoin de rajouter explicitement de la diffusivité longitudinale. Ceci signifie que **la répartition transvese des vitesses, telle qu'elle est prédite par le modèle hydraulique, n'est certainement pas trop loin de la réalité.**
- **Cette bonne concordance est obtenue en fixant des coefficients de dispersion transverse qui sont conformes aux gammes proposées dans la littérature,** pour

des tronçons quasi-rectilignes d'une part, pour des courbes d'autre part.

- Le premier point négatif est que l'on a, dans le bras droit, une avance sytématique des pollutogrammes calculés par rapport aux mesurés. Ceci dit, cette erreur dans la partie centrale est de l'ordre de grandeur des incertitudes sur la mesure du débit. En ce qui concerne le voisinage des rives, peut-être réussirait-on à réduire l'étendue du problème en adoptant un maillage plus fin, ce qui devrait permettre d'approcher plus précisément les gradients de vitesse à la paroi.
- Le second point négatif est que nous n'avons pas une valeur optimale unique pour le coefficient de proportionnalité qui régit la dispersion transverse. Il faut être conscient qu'il n'y avait pas de miracle à attendre compte tenu de la géométrie trés différente des 2 moitiés du bief. Il est reconnu que le mélange transverse est plus fort dans les courbes, à cause de la force accrue des courants secondaires hélicoidaux. Ceux-ci ne peuvent être décrits que par des outils tri-dimensionnels. En bidimensionnel, l'augmentation de l'intensité du mélange se retrouvera implicitement dans le paramétrage de la diffusivité transverse. Ceci dit, choisir un α_y constant n'est pas le nec plus ultra. On a par exemple proposé des formulations où α_y est fonction du rayon de courbure ou des gradients de vitesse locaux (Leclerc, 1990) (également plus accentués en courbe), formulations dont le bien-fondé sur la Seine mérite d'être exploré !

On peut finalement énoncer que **cette première application en Seine est encourageante** ... même s'il sera bon de la confirmer sur d'autres secteurs de la Seine. **ACHAB produit en effet des résultats corrects quant à l'hydraulique et au transport dissous à l'issue d'un calage "minimal"**, puisqu'on ne joue que sur le Strickler et le paramétrage de la diffusivité transverse. De plus, le calage optimal est obtenu dans des gammes "recommandées par la littérature". **Ceci augure favorablement du caractère prédictif du modèle.**

Chapter 12

General conclusions

We have tried to provide intermediate conclusions at different stages of this report, which we do not intend to repeat here in too much detail. We shall at the contrary try to be brief (at least !).

In fact, in this work, our goal was to follow till the end a trail which begins with the understanding of the physical phenomena which control flow and transport development and concludes with a full scale application of a flow and transport model. In the meantime, we have been turning to numerical aspects : once physically based working equations are available how can we solve them so that numerical errors do not blur our estimation and interpretation of what occurs in the laboratory or in natural water bodies. Along this path we were led to make the following comments :

1. As regards transport modelling the more efficient algorithms in advection dominated problems belong to the category of backward characteristic methods. They are not only more accurate but, contrary to their poor reputation, more economical as far as multivariable problems are concerned. We have been performing our benchmark tests in the frame of finite difference methods. However, we consider its conclusions extend to finite element methods too. There, all the information needed for applying spatial interpolation are available; only the particle backtracking must be adapted ... which has already been done (Hervouet, 1986; Hervouet & Watrin, 1988).

2. Our work on the flow equations is much less comprehensive. The proposed models are based on a fractional steps approach, where advection, dispersion and the terms of propagation and friction are dealt with successively. Advection and dispersion are solved by the same algorithms selected for the transport model. As regards the computation of propagation and friction, we have essentially worked on the improvement of factorization methods previously suggested (Dan N'Guyen, 1988; Dan N'Guyen, 1993), following the

same methodology as when testing transport algorithms.

A more faithful finite difference aproximation of the equations, an eventual swap of dependent variables, an iterative solution of the propagation-friction terms, can indeed improve the model performance. The factorization method confirms it is both cheap and accurate.

However, work should go on ! We would like to suggest particularly two research directions. First, working in rectilinear coordinates can be cumbersome when dealing with meandering water bodies. Curvilinear coordinates would be more suited. Then, it can be a challenge to preserve the simplicity or even the mere possibility of applying factorization, as the change of coordinates introduces additional terms in the equations. Perhaps these terms could be dealt with explicitly, or eliminated iteratively, as factorization errors ? An alternative would be to apply nested grids techniques allowing local refinement of the discretization while still working with rectilinear coordinates.
Secondly, one aspect which is much less investigated (or published about?) is the treatment of boundary conditions, both from a numerical and a physical point of view. Once the core of an algorithm is behaving well, effort should be displaced towards the improvement of boundary treatment, boundary conditions being crucial in some cases, as illustrated in our last benchmark test.

3. The only full scale application we present is not intended, as previous hypothetical tests, to demonstrate the superiority of one algorithm over the other. It is just a contribution to the building opinion that, while useful enough when it is based on a good "feeling" of physics, to proceed with the rule of thumb is not always sufficient to deal with environmental problems and that it is worth investing in the development and implementation of more complex, physically based, models. Indeed, we note here that, since an hydraulic model able to describe the longitudinal and transverse gradients of velocity within the flow is applied, the need for calibration of the transport model for dissolved species is considerably reduced.

4. While dealing with the development or mere selection of transport or propagation solution algorithms, we often regreted there were few objective comparisons available. Benchmarks tests for transport are available through the convection-diffusion forum (Baptista *et al.*, 1988). As regards flow models, a few researchers have tried to work on the same set of cases (Weare, 1979; Benque *et al.*, 1982; Stelling *et al.*, 1986; Galland & Hervouet, 1988; Dan N'Guyen, 1988; Perec, 1991) : however they remain too isolated in our opinion.
As hardware becomes more and more poweful, the world of hydraulic modelling is becoming a jungle, where in some cases more emphasis seems to be put on the development of nice graphical user interface than on upgrading of the underlying algorithms. Yet, objective ranking of the available models would be useful, for three groups of people at least : beginners in the field of numerical hydraulic modelling, fellow researchers in environmental

sciences who are looking for an appropriate support to their biogeochemical formulations, non-specialist users and perhaps buyers (water agencies, technical services of cities or counties) who have practical problems at hand.

Finally, we should acknowledge that this started like the quest for the Holy Grail : desperately seeking the perfect algorithm(s). There were obviously many traps in this quest, as in the original. However, there is no such thing as the perfect model (besides, what's the point of finding it, overjoy can kill you ...). In fact, most efficient solutions are often somewhat tailored to the studied problem. What must guide one in the selection or the adjustment of an algorithm is the analysis of the dominant physical phenomena, and the intercomparison with field data collected as carefully as possible. Numerical modellers often get lost in mathematical details in front of their computers. Never forget the physics and field work !

La présente conclusion ne reprendra pas les conclusions intermédiaires fournies au cours de ce rapport. Nous allons au contraire tâcher d'être assez bref.

En résumé, l'objectif de ce travail était de suivre du début jusqu'à la fin le processus qui commence par la compréhension des phénomènes physiques en jeu en mécanique des fluides et s'achève par la mise en oeuvre sur un cas grandeur nature d'un modèle de transport et écoulement. Dans l'intervalle, nous avons été amenés à nous pencher de façon approfondie sur des aspects numériques : une fois que l'on dispose d'équations de travail raisonnables du point de vue physique, comment peut-on les résoudre afin que des erreurs numériques ne compromettent pas notre évaluation et notre interprétation de ce qui se passe, en laboratoire ou au sein d'un cours d'eau naturel ? Ce travail nous a inspiré les commentaires suivants :

1. En ce qui concerne la modélisation du transport de composés dissous au sein d'un écoulement, les algorithmes basés sur la méthode des caractéristiques s'avèrent incontestablement les plus performants quand le phénomène d'advection est prépondérant par rapport aux phénomènes dispersifs. Ces algorithmes se révèlent non seulement les plus précis et robustes mais également, contrairement à la mauvaise réputation qui leur est faite, les plus économiques dès lors qu'on s'intéresse à des problèmes multivariables.

   Nous avons pratiqué nos cas-tests dans le cadre de méthodes aux différences finies. Cependant, nous estimons que leurs conclusions peuvent s'étendre aux méthodes aux éléments finis. Dans le formalisme propre aux éléments finis, toutes les informations nécessaires à la pratique des interpolations spatiales sont disponibles. Il ne reste qu'à adapter l'algorithme de remontée de la caractéristique ... ce qui a déjà été fait (Hervouet, 1986; Hervouet & Watrin, 1988).

2. Notre travail sur la résolution des équations bidimensionnelles de St-Venant couvre un domaine beaucoup moins large. Les modèles proposés sont basés sur l'application d'une technique à pas fractionnaires, où l'on traite successivement les opérateurs d'advection, dispersion et propagation-friction. Advection et dispersion sont résolus par les techniques sélectionnées pour le calcul du transport dissous. Pour l'estimation des termes de propagation et friction, nous nous sommes concentrés sur le test et l'amélioration d'algorithmes de factorisation précédemment proposés (Dan N'Guyen, 1988; Dan N'Guyen, 1993), en suivant la même méthodologie que pour l'évaluation des algorithmes de transport.

   Une approximation aux différences finies plus précise des équations hydrauliques, le changement éventuel de variables de travail, l'introduction d'un mode de résolution itératif des termes de propagation et frottement, permettent bien, avec plus ou moins d'efficacité selon les situations considérées, d'améliorer les performances des algorithmes suggérés à ce jour. La méthode de factorisation confirme qu'elle est tout à la fois performante et efficace du point de vue des ressources informatiques mobilisées (temps calcul).

Néammoins, il reste encore beaucoup à faire. Deux pistes de recherche nous paraissent particulièrement importantes.

Tout d'abord, il peut se révéler pesant de travailler dans un système de coordonnées orthogonal quand on s'intéresse à des cours ou plan d'eau de géométrie irrégulière, ou qui méandrent. L'usage de coordonnées curvilignes serait plus adapté. Cependant, le passage au curviligne pose un challenge. Ce changement de coordonnées introduit en effet des termes croisés supplémentaires dans les équations. Dans ces conditions, sera-t-il encore possible de factoriser les opérateurs différentiels ? Peut-être peut-on songer à traiter explicitement, et éliminer itérativement, ces termes croisés, comme on l'a fait ici pour les erreurs de factorisation ? Une autre alternative serait de travailler avec des grilles de calcul imbriquées de différente finesse, ce qui permet de raffiner localement de façon souple la discrétisation spatiale tout en continuant à travailler dans un systéme de coordonnées rectilignes.

En second lieu, un aspect qui nous semble relativement peu développé est celui du traitement des conditions limite, tout à la fois du point de vue physique et numérique. Il n'est peut-être pas si difficile après tout de développer un modèle dont le "coeur" fonctionne correctement. Une fois que celà est acquis, l'effort devrait se porter sur l'amélioration de la prise en compte des conditions limite, celles-ci se révelant cruciales dans certains cas, comme en témoigne le dernier cas-test pratiqué.

3. La seule application grandeur nature que nous présentons ici n'est pas destinée, au contraire des cas-tests de référence, à démontrer la supériorité d'un algorithme sur l'autre. Elle constitue juste une contribution au parti qui soutient que, quoiqu'utile, la règle de trois n'est pas toujours suffisante en ce qui concerne les problèmes environnementaux et qu'il est "rentable" d'investir dans le développement de modèles plus sophistiqués, plus proches de la réalité physique. En effet, on constate ici comment, avec le support d'un modèle d'écoulement capable de décrire les gradients transverses et longitudinaux du champ de vitesses, les besoins de calage du modèle de transport se trouvent réduits au minimum.

4. Alors que nous travaillions sur le développement ou simplement la sélection d'algorithmes numériques appropriés quant à la solution de l'advection ou des termes de propagation, nous avons été souvent amenés à regretter que peu de comparaisons objectives de techniques numériques soient entreprises ...ou accessibles. Pour les modèles de transport, des problèmes de référence ont pourtant été proposés par le forum Convection-Diffusion (Baptista *et al.*, 1988). En ce qui concerne les modèles hydrauliques, on peut signaler que quelques chercheurs ont traité le même ensemble de cas-tests (Weare, 1979; Benque *et al.*, 1982; Stelling *et al.*, 1986; Galland & Hervouet, 1988; Dan N'Guyen, 1988; Perec, 1991) : ils restent malheureusement trop isolés à notre humble avis !

   Sous la poussée des progrès informatiques, le monde de la modélisation hydraulique tend à ressembler quelque peu à une jungle. De plus en plus de logiciels voient le jour et on peut

se demander dans certains cas si l'effort n'est pas essentiellement consacré au développement (certes utile) d'interfaces utilisateur de plus en plus conviviaux et "jolis" plutôt qu'à la mise à jour des techniques de résolution proprement dites. Il est pourtant utile de procéder, de temps à autre, à une consolidation des connaissances ! Une intercomparaison objective des modèles serait donc utile, ce pour trois groupes de gens au moins : les débutants dans le domaine de l'analyse numérique appliquée à l'hydraulique, nos collègues des sciences de l'environnement à la recherche d'un support adéquat pour y greffer leurs modélisations biogéochimiques , enfin la communauté des non-spécialistes (services techniques des collectivités locales, Agences de l'eau ...) qui se retrouvent confrontés à des problèmes pratiques d'analyse et prévention de la pollution.

Enfin, il faut reconnaître que, dans notre inexpérience, tout ceci a débuté quelque peu comme la quête du Saint-Graal : recherche algorithme(s) parfait(s) désespérément ! Il y eut bien sûr beaucoup d'embûches dans cette quête, comme dans l'originale. Son premier enseignement est sans doute que le modèle parfait n'existe pas (de plus, il pourrait être difficile de surmonter l'excès de bonheur que provoquerait sa découverte ...). En fait les solutions les plus efficaces sont souvent plus ou moins adaptées au problème considéré. Pour se diriger dans la sélection ou la mise au point d'une solution, c'est à l'analyse du problème d'un point de vue physique qu'il faut se raccrocher, ainsi qu'à la comparaison des simulations avec des données, de laboratoire ou de terrain, recueillies le plus soigneusement possible. Les numériciens ont sans doute trop tendance à se perdre dans des détails mathématiques en face de leurs écrans d'ordinateur. Il ne faut jamais perdre de vue la physique, ni la nécessité du travail de terrain !

Bibliography

Abbott, M.B. (1979). *Computational Hydraulics : Elements of the theory of free surface flows.* Pitman Publishing Limited.

Abbott, M.B., Mc Cowan, A. & Warren, I.R. (1981). Numerical modelling of free-surface flows that are two-dimensional in plan. In *Transport models for inland and coastal waters*, pages 222–283. Academic Press, H.G. Fischer editor.

Abraham, G., van Os, A.G. & Verboom, K.G. (1981). Mathematical modelling of flows and transport of conservative substances : requirements for predictive ability. In *Transport models for inland and coastal waters*, pages 1–31. Academic Press, H.G. Fischer editor.

Akima, H. (1970). A new method of interpolation and smooth curve fitting based on local procedures. *Journal of the Association for Computing Machinery*, 17(4):589–602.

Anderson, D.A., Tannehill, J.C. & Pletcher, R.H. (1984). *Computational Fluid Mechanics and Heat Transfer.* Mc Graw Hill, Series in computational methods in mechanics and thermal sciences.

Appere, P. (Août 1988). Modélisation numérique de la propagation d'une onde longue. Technical report, Projet de fin d'etudes de l'Ecole Nationale Supérieure d'Hydraulique de Grenoble mené au Laboratoire National d'Hydraulique (EDF), Groupe Hydraulique Maritime.

Arnold, R.J. (1987). An improved boundary condition for a tidal model of Bass Strait. In *Numerical Modelling : Application to Marine Systems*, pages 145–158. Elsevier Science Publishers B.V., Holland. J. Noye editor.

ASCE Task Committee on Turbulence Models in Hydraulic Computations (1988). Turbulence modeling of surface water flow and transport; Parts I to V. *Journal of Hydraulic Engineering*, 114(9):970–1073.

Baptista, A., Gresho, P. & Adams, E. (June 1988). Reference problems for the convection – diffusion forum. Technical report, VII International Conference on Computational Methods in Water Resources, Cambridge, Massachusets, USA.

Baptista, A.E.M., Adams, E.E. & Stolzenbach, K.D. (1984). Eulerian-lagrangian analysis of pollutant transport in shallow water. Technical report, No. 296, Ralph M. Parsons Laboratory, Aquatic Sciences and Environmental Engineering, Department of Civil Engineering, Massachusets Institute of Technology.

Beckers, J.M. (1991). Application of the GHER 3D general circulation model to the Western Mediterranean. *Journal of Marine Systems*, 1:315–332.

Beltaos, S. (1980a). Longitudinal dispersion in rivers. *Journal of the Hydraulics Division*, 106(1):151–172.

Beltaos, S. (1980b). Transverse mixing tests in natural streams. *Journal of the Hydraulics Division*, 106(10):1607–1625.

Benque, J.P., Cunge, J.A., Feuillet, J., Hauguel, A. & Holly, F.M. (1982). New method for tidal current computation. *Journal of the Waterway, Port, Coastal and Ocean Division*, 108(3):396–417.

Billen, G & Allardi, J. (June 1993). Le fonctionnement de l'écosystème : analyse des processus et modélisation (synthèse 1989-1992). Technical report, PIREN-SEINE Groupe 1, Paris.

Bills, P. & Noye, J. (1987). Open boundary conditions for tidal models. In *Numerical Modelling : Application to Marine Systems*, pages 159–194. Elsevier Science Publishers B.V., Holland, J. Noye editor.

Blumberg, A.F. & Kantha, L.H. (1985). Open boundary conditions for circulation models. *Journal of Hydraulic Engineering*, 111(2):237–255.

Booij, R. (1989). Depth-averaged $k - \epsilon$ modelling. In *Proceedings 23th Congress of the International Association on Hydraulic Research. Technical session : Turbulence in Hydraulics*, pages 199–206, Ottawa.

Book, D.L., Boris, J.P. & Hain, K. (1975). Flux-corrected transport II : Generalizations of the method. *Journal of Computational Physics*, 18:248–283.

Boris, J.P. & Book, D.L. (1973). Flux-corrected transport I : SHASTA, a fluid transport algorithm that works. *Journal of Computational Physics*, 11:38–69.

Boris, J.P. & Book, D.L. (1976). Flux-corrected transport III : Minimal-Error FCT algorithms. *Journal of Computational Physics*, 20:397–431.

Bott, A. (1989a). A positive definite advection scheme obtained by nonlinear renormalization of the advective fluxes. *Monthly Weather Review*, 117(5):1006–1015.

Bott, A. (1989b). Reply to discussion related to *a positive definite advection scheme obtained by nonlinear renormalization of the advective fluxes. Monthly Weather Review*, 117(11):2633–2636.

Brooks, A.N. & Hughes, T.J.R. (1982). Streamline Upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Computer methods in applied mechanics and engineering*, 32:199–259.

Bujon, G. (1983). Modélisation de la dispersion de substances solubles ou pseudo-solubles dans un cours d'eau. Application au cas de la Seine à l'amont de Paris. *La Houille Blanche*, (1).

Carlier, M. (1986). *Hydraulique Générale Appliquée*. Eyrolles, Coll. de la Direction des Etudes et Recherches d'Electricité de France.

Chassain, P. (1993). *Turbulence en mécanique des fluides : Analyse du phénomène dans une perspective de modélisation à l'usage de l'ingénieur*. Institut National Polytechnique de Toulouse, Cours ENSEEIHT, France.

Chen, C.J., Tien, C.J., Carlson, K.D. & Bernatz, R.A. (1993). Finite analytic method for two and three dimensional flows with complex geometry. In *Proceedings 5th International Symposium on Refined Flow Modelling and Turbulence Measurements*, pages 161–173, Paris.

Chen, Y. & Falconer, R.A. (1992). Advection-diffusion modelling using the modified QUICK scheme. *International Journal for numerical methods in fluids*, 15:1171–1196.

Chow, V.T. (1959). *Open Channel Hydraulics*. Mc Graw Hill, International Student Edition.

Crockett, C.P., Moys, G.D., Osborne, M.P. & Norreys, R.J. (1989). Modelling the river impact of intermittent urban discharges. In *International conference on Hydraulic and Environmental Modelling of coastal, estuarine and river waters*, pages 371–380, Bradford, UK.

Crowley, W.P. (1968). Numerical advection experiments. *Monthly Weather Review*, 96(1):1–11.

Cunge, J. A. (1988). Some examples of interaction of numerical and physical aspects of free surface flow modelling. In *Proceedings 7th International Conference on Computational Methods in Water Resources, Vol. 1*, pages 3–12, MIT, Cambridge, USA. Computational Mechanics Publication, Elsevier.

Cunge, J.A., Holly, F.M. & Verwey, A. (1980a). Chapter 2 : Mathematical formulation of physical processes. In *Practical aspects of computational river hydraulics*, pages 7–52. Pitman Advanced Publishing Program.

Cunge, J.A., Holly, F.M. & Verwey, A. (1980b). Chapter 8 : Transport of pollutants. In *Practical aspects of computational river hydraulics*, pages 312–349. Pitman Advanced Publishing Program.

Dan N'Guyen, K. (1988). *Modélisation numérique 2D et 3D de la circulation générale en milieux estuariens et côtiers : Application à l'estuaire de la Gironde.* Thèse de Doctorat, Université de Paris VI. Spécialité Mécanique et Energie.

Dan N'Guyen, K. (1993). *Modélisation numérique d'écoulements côtiers et Techniques de résolution des équations de Navier-Stokes en coordonnées généralisées.* Thèse de Doctorat, Université des Sciences et Technologies de Lille. Habilitation à diriger des recherches.

Daubert, A. & Graffe, O. (1967). Quelques aspects des écoulements presque horizontaux à deux dimensions en plan et non permanents. Application aux estuaires. *La Houille Blanche*, (847–860).

Day, T.J. (1975). Longitudinal dispersion in natural channels. *Water Resources Research*, 11(6):909–918.

Day, T.J. & Wood, I.R. (1976). Similarity of the mean motion of fluid particles dispersing in a natural channel. *Water Resources Research*, 12(4):655–666.

Demetracopoulos, A.C. & Stefan, H.G. (1983). Transverse mixing in wide and shallow river : case study. *Journal of Environmental Engineering*, 109(3):685–699.

Demuren, A.O. (1983). Three-dimensional numerical computation of flow and pollutant dispersion in meandering channels. In *Proceedings 20st Congress of the International Association on Hydraulic Research, Vol. III*, pages 29–36, Moscow.

Demuren, A.O. & Rodi, W. (1983). Side discharges into open channels : mathematical model. *Journal of Hydraulic Engineering*, 109(12):1707–1722.

Demuren, A.O. & Rodi, W. (1986). Calculation of flow and pollutant dispersion in meandering channels. *Journal of Fluid Mechanics*, 172:63–92.

Dervieux, A. & Palmerio, B. (1993). Advances in mesh adaptation for computational fluid dynamics. In *Proceedings 5th International Symposium on Refined Flow Modelling and Turbulence Measurements*, pages 151–160, Paris.

Ditmars, J.D., Adams, E.E., Bedford, K.W. & Ford, D.E. (1987). Performance evaluation of surface water transport and dispersion models. *Journal of Hydraulic Engineering*, 113(8):961–980.

Donea, J., Quarlapelle, L. & Selmin, V. (1987). An analysis of time discretization in the finite element solution of hyperbolic problems. *Journal of Computational Physics*, 70:463–499.

Dong, P., Walker, D.J. & Anastasiou, K. (1989). The modelling of wave breaking induced turbulent mixing using depth-averaged turbulence transport equations. In *Proceedings 23th*

Congress of the International Association on Hydraulic Research. Technical session : Turbulence in Hydraulics, pages 181–189, Ottawa.

Elder, J.W. (1959). The dispersion of marked fluid in turbulent shear flow. *Journal of Fluid Mechanics*, (5):544–560.

Even, S. & Poulin, M. (May 1993). Modélisation de l'écosystème Seine entre Montereau et Poses : modéle PROSE. Technical report, CIG, Ecole Nationale Supérieure des Mines de Paris, Fontainebleau, France.

Fairweather, G. & Navon, I.M. (1980). A linear ADI method for the shallow-water equations. *Journal of Computational Physics*, 37:1–18.

Falconer, R.A. & Liu, S. (1988). Modeling solute transport using QUICK scheme. *Journal of Environmental Engineering*, 114(1):3–19.

Fischer, H.B. (1967). The mechanics of dispersion in natural streams. *Journal of the Hydraulics Division*, 93(6):187–216.

Fischer, H.B. (1968). Dispersion predictions in natural streams. *Journal of the Sanitary Engineering Division*, 94(5):927–943.

Fischer, H.B. (1969). The effects of bends on dispersion in natural streams. *Water Resources Research*, 5(2):496–506.

Fischer, H.B. *et al.* (1979). *Mixing in inland and coastal waters.* Academic Press.

Fletcher, C.A.J. (Second Edition, 1991). *Computational techniques for fluid dynamics, Vol. I and II.* Springer series in computational physics, Springer-Verlag Ed., Berlin.

Fritsch, F.N. & Carlson, R.E. (1980). Monotone piecewise cubic interpolation. *SIAM Journal of Numerical Analysis*, 17(2):238–246.

Galland, J.C. & Hervouet, J.M. (1988). Code TELEMAC (système Ulysse). Résolution des équations de St-Venant bidimensionnelles. Dossier de validation. Technical report, No HE 43/87.40 & 42/88.11 Laboratoire National d'Hydraulique, Electricité de France.

Gaskell, P.H. & Lau, A.K.C. (1988). Curvature-compensated convective transport : SMART, a new boundedness-preserving transport algorithm. *International Journal for Numerical Methods in Fluids*, 8:617–641.

Giles, M. (1989). Numerical methods for unsteady turbomachinery flow. In *Numerical Methods for Flows in Turbomachinery.* Lecture Series 1989-06, von Karman Institute for Fluid Dynamics.

Goussebaile, J., Haugel, A. & Labadie, G. (1986). Introduction à la technique des élements finis. Développement en mécanique des fluides. Technical report, No HE 43/86.21 . Laboratoire National d'Hydraulique, Electricité de France.

Goutal, N. (1987). Résolution des équations de St-Venant en régime transcritique par une méthode d'éléments finis : application aux bancs découvrants. *Bulletin de la Direction des Etudes et Recherches. Série C Mathématique et Informatique. Electricité de France*, (4):5–120.

Goutal, N. & Hérard, G. (1990). Cours de simulation numérique du LNH. Méthodes de résolution des systèmes matriciels. Complément au tome 1. Technical report, No HE 41/90.21 A. Laboratoire National d'Hydraulique, Electricité de France.

Ha Minh, H. (1993). Modélisation de la turbulence : ses aspects physiques et son impact sur la simulation numérique des écoulements réels. In *Actes du 11^{eme} congrès français de Mécanique*, pages 19–60, Lille, FR.

Hackbusch, W. (1985). *Multigrid methods and applications.* Springer series in computational mathematics, Springer-Verlag Ed., Berlin.

Harden, T.O. & Shen, H. T. (1979). Numerical simulation of mixing in natural rivers. *Journal of the Hydraulics Division*. 105(4):393–408.

Hervouet, J.M. (1984). Application de la méthode des caractéristiques en formulation faible à la résolution des équations d'advection bidimensionnelles sur des maillages grilles. Technical report, No HE 41/84.11 Laboratoire National d'Hydraulique, Electricité de France.

Hervouet, J.M. (1986). CARAC, module de convection en éléments finis, par la méthode des caractéristiques. Technical report, No HE 41/86.21 Laboratoire National d'Hydraulique, Electricité de France.

Hervouet, J.M. (1991). Vectorisation et simplification des algorithmes en éléments finis. *Bulletin de la Direction des Etudes et Recherches. Série C Mathématique et Informatique. Electricité de France*, (1):1–37.

Hervouet, J.M. & Watrin, A. (1988). Code TELEMAC (système Ulysse). Résolution des équations de St-Venant bidimensionnelles. Théorie et mise en oeuvre informatique. Technical report, No HE 43/87.37 Laboratoire National d'Hydraulique, Electricité de France.

Holley, E.R. (1987). Transport of pollutants in rivers. In *Proceedings 22^{th} Congress of the International Association on Hydraulic Research. Technical session : Topics in fluvial hydraulics*, pages 19–41, Lausanne.

Holley, E.R. & Abraham, G. (1973). Field tests on transverse mixing in rivers. *Journal of the Hydraulics Division*, 99(12):2313–2331.

Holly, F., Jardin, P. & Antémi, E. (1990). Modélisation de la dispersion de pollutions accidentelles en rivière. Le programme POLDER. *La Houille Blanche*, (3/4):219–223.

Holly, F.M. (1975). Two-dimensional mass dispersion in rivers. Technical report, No 78 , Hydrology Papers, Colorado State University, Fort Collins, Colorado.

Holly, F.M. (1979). Le programme POLDER : simulation de la dispersion de polluants dans les rivières par modèle mathématique. Technical report, SOGREAH, Grenoble, France.

Holly, F.M. & Nerat, G. (1983). Field calibration of a stream-tube dispersion model. *Journal of Hydraulic Engineering*, 109(11):1455–1470.

Holly, F.M. & Preissmann, A. (1977). Accurate calculation of transport in two dimensions. *Journal of the Hydraulics Division*, 103(11):1259–1277.

Holly, F.M. & Usseglio-Polatera, J. M. (1984). Dispersion simulation in two dimensional tidal flow. *Journal of Hydraulic Engineering*, 110(7):905–926.

Hug, M., editor (1975). *Mécanique des Fluides Appliquée*. Eyrolles.

Hunt, J.C.R. (1993). Theoretical limitations of computational modelling of turbulent flows. In *Proceedings 5^{th} International Symposium on Refined Flow Modelling and Turbulence Measurements*, pages 13–38, Paris.

Hyman, J.M. (1983). Accurate monotonicity preserving cubic interpolation. *SIAM Journal of Scientific and Statistical Computing*, 4(4):645–654.

IAHR, editor (1993). *Refined Flow Modelling and Turbulence Measurements*, Paris. Proceedings of the 5^{th} International Symposium, Presses des Ponts et Chaussées.

Iliev, O.P., Makarov, M.M. & Vassilevski, P.S. (1992). Performance of certain iterative methods in solving implicit difference schemes for 2-D Navier-Stokes equations. *International Journal for numerical methods in engineering*, 33:1465–1479.

Józsa, J. (1989). 2-D particle model for predicting depth-integrated pollutant and surface oil slick transport in rivers. In *International conference on Hydraulic and Environmental Modelling of coastal, estuarine and river waters*, pages 332–340, Bradford, UK.

Kaasschieter, E.F. (1986). The solution of non-symmetric linear systems by bi-conjugate gradients or conjugate gradients squared. Technical report, No 86-21. Delft University of Technology, Department of Mathematics and Informatics.

Keller, R.J. & Rodi, W. (1988). Prediction of flow characteristics in main channel/flood plain flows. *Journal of Hydraulic Research*, 26(4):425–441.

Krishnappan, B.G. & Lau, Y.L. (1986). Turbulence modeling of flood plain flows. *Journal of Hydraulic Engineering*, 112(4):251–266.

Labadie, G. (1986). Méthodes de résolution des systèmes linéaires (tome 1). Technical report, No HE 43/86.33 . Laboratoire National d'Hydraulique, Electricité de France.

Lassale, F. (1992). Etude de la modélisation du transport de polluant en rivière. Master's thesis, Institut des Sciences de l'Ingénieur de Montpellier - Cergrene ENPC.

Lau, Y.L. & Krishnappan, B.G. (1977). Transverse dispersion in rectangular channels. *Journal of the Hydraulics Division*, 103(10):1173–1189.

Lau, Y.L. & Krishnappan, B.G. (1981). Modeling transverse mixing in natural streams. *Journal of the Hydraulics Division*, 107(2):209–226.

Lebosse, A. (1987). Calcul d'une tâche thermique dans la Loire à St-Laurent des Eaux. In *Proceedings 22th Congress of the International Association on Hydraulic Research. Technical session : Topics in fluvial hydraulics*, pages 275–280, Lausanne.

Leclerc, M. (1990). A finite element model of estuarian and river flows with moving boundaries. *Advance Water Resources*, 13(4):158–168.

Leclerc, M. & Boudreault, P. (1993). Méthodologie d'analyse détaillée de la contamination par tronçon du fleuve St-Laurent apr modélisation numérique : le cas du lac St-Pierre. *Revue des Sciences de l'Eau*, 6(4):427–452.

Leendertse, J.J. (1970). *A water quality simulation model for well-mixed estuaries and coastal seas. Vol. 1 : Principles of computation.* Rand Corporation Memorandum, RM-6230-RC.

Leonard, B.P. (1979). A stable and accurate convective modelling procedure based on quadratic upstream interpolation. *Computer methods in applied mechanics and engineering*, 19:59–98.

Leonard, B.P. (1981). A survey of finite differences with upwinding for numerical modelling of the incompressible convective diffusion equation. In *Recent Advances in Numerical Methods in Fluids*, pages 63–111. Pineridge Press, C. Taylor editor.

Leonard, B.P. (1988). Simple high-accuracy resolution program for convective modelling of discontinuities. *International Journal for Numerical Methods in Fluids*, 8:1291–1318.

Leonard, B.P. & Niknafs, H.S. (1991). Sharp monotonic resolution of discontinuities without clipping of narrow extrema. *Computers & Fluids*, 19(1):141–154.

Lesieur, M. (1987). *Turbulence in Fluids.* Mechanics of Fluids and Transport Processes, M. Nijhoff Publishers, Dordrecht.

Lesieur, M. (1993). Advance and state of the art on large-eddy simulations. In *Proceedings 5^{th} International Symposium on Refined Flow Modelling and Turbulence Measurements*, pages 3–11, Paris.

Lesieur, M. (1994). *La turbulence.* Presses Universitaires de Grenoble, Collection Grenoble Sciences, France.

Li, C.W. (1990). Advection simulation by minimax-characteristics method. *Journal of Hydraulic Engineering*, 116(9):1138–1144.

Liren, Y. & Shu-nong, Z. (1989). A new depth-averaged $k - w$ turbulent closure model and its application. In *Proceedings 23^{th} Congress of the International Association on Hydraulic Research. Technical session : Turbulence in Hydraulics*, pages 171–180, Ottawa.

Liu, H. & Cheng, A.H. (1980). Modified fickian model for predicting dispersion. *Journal of the Hydraulics Division*, 106(6):1021–1040.

Lyn, D.A. (1993). Turbulence measurements in open channel flows over artificial bed forms. *Journal of Hydraulic Engineering*, 119(3):306–326.

Marivoet, J.L. & Van Craenenbroeck, W. (1986). Longitudinal dispersion in ship-canals. *Journal of Hydraulic Research*, 24(2):123–132.

Mary, D. (1982). *Modélisation numérique bidimensionnelle des écoulements en rivière. Application à l'étude de rejets thermiques.* Thèse de Doctorat, Ecole nationale des Ponts et Chaussées.

Masbernat, L., Line, A. & Mocke, G. (1987). Etude de l'hydrodynamique du Bassin de Thau par modélisation mathématique. Technical report, Institut de Mécanique des Fluides de Toulouse - BCEOM - Service Maritime et de Navigation du Languedoc - Roussillon.

Miller, C.T. & Rabideau, A.J. (1993). Development of Split-Operator, Petrov-Galerkin methods to simulate transport and diffusion problems. *Water Resources Research*, 29(7):2227–2240.

Mouchel, J.M. (1989). Modélisation de la qualité des eaux de l'Arc; Application du modèle RIVOLI aux campagnes de Juillet et Septembre 1987; Validation en Juillet 1988. Technical report, Research contract for Agence de Bassin Rhône-Méditerranée-Corse, CERGRENE (ENPC).

Mouchel, J.M. (1990). Un modèle lagrangien de qualité : application à l'Arc en aval d'Aix-en-Provence. *La Houille Blanche*, (3/4):181–186.

Mouchel, J.M. & Simon, L. (1993). Impact of wet weather discharges in the River Seine : major water quality parameters. In *Proceedings 6th International Conference on Urban Storm Drainage, J. Marsalek and H.C. Torno Editors*, pages 200–205, Niagara Falls, Canada.

Mouchel, J.M. *et al.* (1993). Impact des surverses d'orage sur la qualité des eaux de la Seine dans l'agglomération parisienne : Rapport de Synthèse 1989-1992. Technical report, Programme de Recherches CNRS PIREN-SEINE.

Moulin, V., Caruso, A., Daubert, O., Pot, G. & Thomas, B. (1993). Improvement of finite element algorithms implemented in CFD code N3S for turbulent incompressible and dilatable flows. In *Proceedings 5th International Symposium on Refined Flow Modelling and Turbulence Measurements*, pages 223–230, Paris.

Naot, D. & Nezu, I. (1993). Towards the calculations of secondary currents in compound channel flows; shallow and deep flood plains. In *Proceedings 5th International Symposium on Refined Flow Modelling and Turbulence Measurements*, pages 687–693, Paris.

Naot, D., Nezu, I. & Nagakawa, H. (1993). Hydrodynamic behavior of rectangular compound open channel flows. *Journal of Hydraulic Engineering*, 119(3):390–408.

Naot, D. & Rodi, W. (1982). Calculation of secondary currents in channel flow. *Journal of the Hydraulics Division*, 108(8):948–968.

Nezu, I. & Nagakawa, H. (1993). Three-dimensional structures of coherent vortices generated behind dunes in turbulent free-surface flows. In *Proceedings 5th International Symposium on Refined Flow Modelling and Turbulence Measurements*, pages 603–612, Paris.

Nezu, I., Nagakawa, H. & Tominaga, A. (1993a). Turbulent structures and bursting phenomena over roughness discontinuity in open channel flows. In *Proceedings 5th International Symposium on Refined Flow Modelling and Turbulence Measurements*, pages 629–636, Paris.

Nezu, I. & Rodi, W. (1986). Open-channel flow measurements with a lased Doppler anemometer. *Journal of Hydraulic Engineering*, 112(5):335–355.

Nezu, I., Tominaga, A. & Nagakawa, H. (1993b). Field measurements of secondary currents in straight rivers. *Journal of Hydraulic Engineering*, 119(5):598–615.

Nihoul, J.C.J., Deleersnijder, E. & Djenidi, S. (1989a). Modelling the general circulation of shelf seas by 3D $k - \epsilon$ models. *Earth Science Review*, 26:163–189.

Nihoul, J.C.J., Deleersnijder, E. & Djenidi, S. (1989b). Modelling the general circulation of shelf seas by 3d $k - \epsilon$ models. *Earth Science Review*, 26:163–189.

Nihoul, J.C.J. & Djenidi, S. (1987). Perspective in three-dimensional modelling of the marine system. In *Three-dimensional models of marine and estuarine dynamics*, pages 1–33. Elsevier Oceanography Series 45, J. Nihoul and B. Jamart editors.

Nokes, R.I & Wood, I.R. (1987). Lateral turbulent dispersion in open channel flow. In *Proceedings 22th Congress of the International Association on Hydraulic Research. Technical session : Topics in fluvial hydraulics*, pages 233–238, Lausanne.

Noorishad, J., Tsang, C.F., Perrochet, P. & Musy, A. (1992). A perspective on the numerical solution of convection-dominated transport problems : a price to pay for the easy way out. *Water Resources Research*, 28(2):551–561.

Nordin, C.F. & Troutman, B.M. (1980). Longitudinal dispersion in rivers : the persistence of skewness in observed data. *Water Resources Research*, 16(1):123–128.

Oey, L.Y., Mellor, G. & Hires, R (1985a). A three-dimensional simulation of the Hudson-Rarity estuary. Part I: Description of the model and model simulations. *Journal of Physical Oceanography*, 15:1676–1692.

Oey, L.Y., Mellor, G. & Hires, R (1985b). A three-dimensional simulation of the Hudson-Rarity estuary. Part II: Comparison with observation. *Journal of Physical Oceanography*, 15:1693–1710.

Oey, L.Y., Mellor, G. & Hires, R (1985c). A three-dimensional simulation of the Hudson-Rarity estuary. Part III: Salt flux analyses. *Journal of Physical Oceanography*, 15:1711–1720.

Orlanski, I. (1976). A simple boundary condition for unbounded hyperbolic flows. *Journal of Computational Physics*, 21:251–269.

Ouillon, S. (1993). *Modélisation mathématique de l'hydrodynamique à surface libre et du transport en suspension de sédiments non-cohésifs. Aide à l'interprétation d'images spatiales.* Thèse de Doctorat, Institut National Polytechnique de Toulouse.

Pavlovic, R.N. & Rodi, W. (1985). Depth-averaged numerical predictions of velocity and concentration fields in meandering channels. In *Proceedings 21st Congress of the International Association on Hydraulic Research*, pages 121–125, Melbourne.

Perec, G. (1991). Experimental demonstration of the tomatopic organization in the Soprano (Cantatrix Sopranica L). In *Cantatrix Sopranica L. et autres écrits scientifiques, Editions Seuil*, pages 12–33, Paris.

Peters, A. (1988). Vectorized programming issues for FE models. In *Proceedings 7th International Conference on Computational Methods in Water Resources, Vol. 1*, pages 13–22, MIT, Cambridge, USA. Computational Mechanics Publication, Elsevier.

Phillips, R.E. & Schmidt, F.W. (1984). Multigrid techniques for the numerical solution of the diffusion equation. *Numerical Heat Transfer*, 7:251–268.

Phillips, R.E. & Schmidt, F.W. (1985). Multigrid techniques for the solution of the passive scalar advection-diffusion equation. *Numerical Heat Transfer*, 8:25–43.

Pochat, R. (1989). *Cours d'Hydraulique Générale.* Ecole nationale du Génie Rural, des Eaux et des Forêts.

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1989). *Numerical Recipes. The Art of Scientific Computing (Fortran Version).* Cambridge University Press.

Prinos, P. (1993). Three-dimensional flow in compound open channels. In *Proceedings 5^{th} International Symposium on Refined Flow Modelling and Turbulence Measurements*, pages 669–677, Paris.

Rasch, P.J. & Williamson, D.L. (1990). On shape preserving interpolation and semi – lagrangian transport. *SIAM Journal of Scientific and Statistical Computing*, 11(4):656–687.

Rastogi, A.K. & Rodi, W. (1978). Predictions of heat and mass transfer in open channels. *Journal of the Hydraulics Division*, 104(3):397–420.

Raviart, P.A. (1981). *Les méthodes d'éléments finis en mécanique des fluides.* Eyrolles, Coll. de la Direction des Etudes et Recherches d'Electricité de France.

Richtmyer, R.D. & Morton, K.W. (Second Edition, 1967). *Difference methods for initial value problems.* Interscience Publishers, J. Wiley and Sons.

Rigaudière, P. (1992). Simulation de transfert de pollution sur le Cher; Rapport final. Technical report, Research contract for DIREN CENTRE, CEMAGREF - CERGRENE (ENPC).

Roache, P.J. (Revised Edition 1985). *Computational Fluid Dynamics.* Hermosa Publishers.

Rodi, W. (1980). Turbulence models and their application in hydraulics. Technical report, International Association for Hydraulic Research. Delft.

Rodi, W., Pavlovic, R.N. & Srivatsa, S.K. (1981). Prediction of flow and pollutant spreading in rivers. In *Transport models for inland and coastal waters*, pages 63–111. Academic Press, H.G. Fischer editor.

Rosello-Tournoud, M.G. (1991). *Analyse du comportement d'un écosystème lagunaire à diverses échelles de temps et d'espace : application à l'étang de Thau.* Thèse de Doctorat, Université des Sciences et Techniques du Languedoc.

Ruger, M. & Sommerfeld, M. (1993). A finite volume multigrid method for calculating iron-pentacarbonyl decomposition in turbulent flows. In *Proceedings 5th International Symposium on Refined Flow Modelling and Turbulence Measurements*, pages 175–182, Paris.

Sabol, G.V. & Nordin, C.F. (1978). Dispersion in rivers as related to storage zones. *Journal of the Hydraulics Division*, 104(5):695–708.

Schohl, G.A. & Holly, F.M. (1991). Cubic-spline interpolation in lagrangian advection computation. *Journal of Hydraulic Engineering*, 117(2):248–253.

Shiono, K. & Muto, Y. (1993). Secondary flow structure for in-bank and over-bank flows in trapezoidal meandering compound channel. In *Proceedings 5th International Symposium on Refined Flow Modelling and Turbulence Measurements*, pages 645–652, Paris.

Shokin, Y.I. & Kompaniets, L.A. (1987). A catalogue of the extra-boundary conditions for the difference schemes approximating the hyperbolic equations. *Computers & Fluids*, 15(2):119–136.

Silveira Neto, A. (1991). *Simulation numérique des grandes échelles d'un écoulement décollé à l'aval d'une marche*. Thèse de Doctorat, Institut National Polytechnique de Grenoble.

Simon, L. (Dec. 1992). Dispersion des rejets en seine : 1^{ere} phase d'expoitation des traçages du 7/7 et 8/9/1992. Technical report, No IV/92/09, PIREN-SEINE program, and CERGRENE (Ecole Nationale des Ponts et Chaussées).

Simon, L. (Jun. 1990a). Numerical resolution of advection-dispersion equation. Technical report, CERGRENE (Ecole Nationale des Ponts et Chaussées).

Simon, L. (Nov. 1990b). Tests monodimensionnels - Schémas de Bott. Technical report, CERGRENE (Ecole Nationale des Ponts et Chaussées).

Simon, L., Maldiney, M.A. & Mouchel, J.M. (1994). Transfer of combined sewer overflows in the Seine River. *Water Sciences and Technology*, 29(1-2):209–217.

Smolarkiewicz, P.K. (1983). A simple positive definite advection scheme with small implicit diffusion. *Monthly Weather Review*, 111(3):479–486.

Smolarkiewicz, P.K. (1984). A fully multidimensional positive definite advection transport algorithm with small implicit diffusion. *Journal of Computational Physics*, 54:325–362.

Smolarkiewicz, P.K. (1989). Comment on *a positive definite advection scheme obtained by nonlinear renormalization of the advective fluxes*. *Monthly Weather Review*, 117(11):2626–2632.

Smolarkiewicz, P.K. & Clark, T.L. (1986). The multidimensional positive definite advection transport algorithm : Further development and applications. *Journal of Computational Physics*, 67:396–438.

Smolarkiewicz, P.K. & Grabowski, W.W. (1990). The multidimensional positive definite advection transport algorithm : Nonoscillatory option. *Journal of Computational Physics*, 86:355–375.

Sonneveld, P. (1984). CGS, a fast Lanczos-type solver for nonsymmetric linear systems. Technical report, No 84-16. Delft University of Technology, Department of Mathematics and Informatics.

Stelling, G.S. & Wang, L.X. (1984). Experiments and computation on unsteady separating flow in an expanding flume. Technical report, No 84-2 . Delft University of Technology, Department of Civil Engineering, Laboratory of Fluid Mechanics.

Stelling, G.S., Wiersma, A.K. & Willemse, J.B.T.M. (1986). Practical aspects of accurate tidal computations. *Journal of Hydraulic Engineering*, 112(9):802–817.

Szymkiewicz, R. (1993). Oscillation-free solution of shallow water equations for nonstaggered grid. *Journal of Hydraulic Engineering*, 119(10):1118–1137.

Takacs, L.L. (1985). A two-step scheme for the advection equation with minimized dissipation and dispersion errors. *Monthly Weather Review*, 113(6):1050–1065.

Tassin, B. (1986). *Contribution à la modélisation écologique du lac Léman : modèles physiques et biogéochimiques du lac.* Thèse de Doctorat, Ecole nationale des Ponts et Chaussées.

Théry, L., Simon, L. & Poulin, M. (Nov. 1993b). Simulation numérique du transport de substances dissoutes en rivière à l'aide d'un modèle à tubes de courant; Rapport Final. Technical report, Research contract for Compagnie Générale des Eaux, CERGRENE (ENPC) - CIG (ENSMP).

Théry, L., Simon, L. & Poulin, M. (Oct. 1993a). Simulation numérique du transport de substances dissoutes en rivière à l'aide d'un modèle à tubes de courant. 2^{nd} Rapport Intermédiaire. Technical report, Research contract for Compagnie Générale des Eaux, CERGRENE (ENPC) - CIG (ENSMP).

Tremback, C.J., Powell, J., Cotton, W.R. & Pielke, R.A. (1987). The forward-in-time upstream advection scheme : extension to higher orders. *Monthly Weather Review*, 115(2):540–555.

Uan, M. (1979). étude d'un modle simplifié d'évolution des courants secondaires dans les coudes des canaux et rivières. Technical report, No E43/79.47 Laboratoire National d'Hydraulique, Electricité de France.

Usseglio-Polatera, J. M. (April 1988). Validation of ARGOS-2D modelling system. Technical report, CEFRHYG, Centre de Formation, de Recherche et d'Essais Hydraulique de Grenoble, FRANCE.

Usseglio-Polatera, J. M. & Chenin-Mordojovich, M.I. (1988). Fractional steps and process splitting methods for industrial codes. In *Proceedings 7th International Conference on Computational Methods in Water Resources, Vol. 2*, pages 167–172, MIT, Cambridge, USA. Computational Mechanics Publication, Elsevier.

Valentine, E.M. & Wood, I.R. (1977). Longitudinal dispersion with dead zones. *Journal of the Hydraulics Division*, 103(9):975–990.

Valentine, E.M. & Wood, I.R. (1979). Experiments in longitudinal dispersion with dead zones. *Journal of the Hydraulics Division*, 105(8):999–1016.

Van Leer, B. (1977). Towards the ultimate conservative difference scheme IV : a new approach to numerical convection. *Journal of Computational Physics*, 23:276–299.

Vatvani, D.K. & Montazeri, M. (1989). Performance of some high accurate semi-lagrangian numerical schemes for the scalar advection equation. Technical report, Waterloopkundig Laboratorium, Delft, Pays-Bas.

Vinçon-Leite, B. (1991). *Contribution de la modélisation mathématique à l'étude de la qualité de l'eau dans les lacs sub-alpins: le lac du Bourget (Savoie).* Thèse de Doctorat, Ecole nationale des Ponts et Chaussées.

Wang, J.D., Cofer-Shabica, S.V. & Fatt, J.C. (1988). Finite element characteristic advection model. *Journal of Hydraulic Engineering*, 114(9):1098–1114.

Weare, T.J. (1979). Errors arising from irregular boundaries in ADI solutions of the shallow-water equations. *International Journal for numerical methods in engineering*, 14:921–931.

Webel, G. & Schatzmann, M. (1984). Transverse mixing in open channel flow. *Journal of Hydraulic Engineering*, 110(4):423–435.

Weiyan, Tan (1992). *Shallow water hydrodynamics.* Elsevier Oceanography Series 55, Elsevier, Amsterdam.

Wilders, P., van Stijn, T.L., Stelling, G.S. & Fokkema, G.A. (1988). A fully implicit splitting method for accurate tidal computations. *International Journal for numerical methods in engineering*, 26:2707–2721.

Williamson, D.L. & Rasch, P.J. (1989). Two – dimensional semi – lagrangian transport with shape preserving interpolation. *Monthly Weather Review*, 117(1):102–129.

Yanenko, N.N. (1968). *Méthode à pas fractionnaires. Résolution de problèmes polydimensionnels de physique mathématique.* Collection Intersciences, Librairie Armand Colin.

Yee, H.C. (1987). Construction of explicit and implicit symmetric TVD schemes and their applications. *Journal of Computational Physics*, 68:151–179.

Yotsukura, N. & Cobb, E.D. (1972). Transverse diffusion of solutes in natural streams. Technical report, U.S. Geological survey professional paper 582-C.

Yotsukura, N., Fischer, H.B. & Sayre, W. W. (1970). Measurement of mixing characteristics of the Missouri River between Sioux City, Iowa, and Plattsmouth, Nebraska. Technical report, U.S. Geological survey water-supply paper 1899-G.

Yotsukura, N. & Sayre, W. W. (1976). Transverse mixing in natural channels. *Water Resources Research*, 12(4):695–704.

Zalesak, S.T. (1979). Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of Computational Physics*, 31:355–362.

Mémoire présenté pour l'obtention du titre de
Docteur de l'Ecole Nationale des Ponts et Chaussées

Spécialité Sciences et Techniques de l'Environnement

Contribution à la modélisation numérique du transport de polluants en rivière.

par

Laure SIMON

ANNEXES

Thèse soutenue le 13 Janvier 1995

Jury

M. Bruce BECK	Directeur de Thèse
M. Kim DAN N'GUYEN	Rapporteur
M. Laszlo SOMLYODY	Rapporteur
M. Jean CUNGE	Examinateur
M. Ghislain de MARSILY	Examinateur
M. Rémy POCHAT	Examinateur

11

Appendix A

Turbulence modelling

A.1 Fluid equations

All symbols are indicated at the end of this appendix

As detailed in Chapter 1, the Reynolds equations, which govern the mean flow and the three-dimensional scalar transport, read (in tensor notations) :

$$\frac{\partial U_i}{\partial x_i} = 0 \tag{A.1}$$

$$\frac{\partial U_i}{\partial t} + U_j \frac{\partial U_i}{\partial x_j} = \underbrace{\frac{-1}{\rho}\frac{\partial P}{\partial x_i}}_{\text{Pressure}} - \underbrace{\frac{\partial \overline{u_i u_j}}{\partial x_j}}_{\text{Turbulence}} + \underbrace{g_i \frac{\rho - \rho_r}{\rho_r}}_{\text{Buoyancy}} \tag{A.2}$$

$$\frac{\partial \phi}{\partial t} + U_j \frac{\partial \phi}{\partial x_j} = - \underbrace{\frac{\partial \overline{\varphi u_j}}{\partial x_j}}_{\text{Turbulence}} \tag{A.3}$$

The task of turbulence models is to represent terms such as $\overline{u_i u_j}$ and $\overline{\varphi u_j}$ in a way which *closes* the equations by relating these correlations to the mean-flow averaged quantities. Turbulence models do not describe the detail of the turbulent motion but only its average effects on mean flow.

The usual steps in turbulence modelling are the following. The local state of turbulence, and thus the turbulence correlations, are assumed to be described only by a few parameters. Then, the model must on one hand bridge parameters and correlations, i.e. postulate a relationship between them, on the other hand determine the distribution of the parameters over the flow field.

Based on dimensional analysis, the velocity and length scale of the turbulent motion play a major part in turbulence parametrization. It is also worth noting that, no matter how complex and "physically-based" a turbulence model is, it always contain some amount of empirical information, in the form of empirical constants or functions.

The depth-averaged version of Reynolds equations (in the usual $x-y-z$ coordinate system) is :

$$\frac{\partial h}{\partial t} + \frac{\partial h\overline{U}}{\partial x} + \frac{\partial h\overline{V}}{\partial y} = 0 \tag{A.4}$$

$$\begin{aligned}\frac{\partial \overline{U}}{\partial t} + \overline{U}\frac{\partial \overline{U}}{\partial x} + \overline{V}\frac{\partial \overline{U}}{\partial y} &= -g\frac{\partial \zeta}{\partial x} \\ &+ \underbrace{\frac{1}{\rho h}\frac{\partial (h\bar{\tau}_{xx})}{\partial x} + \frac{1}{\rho h}\frac{\partial (h\bar{\tau}_{xy})}{\partial y}}_{\text{turbulence}} + \frac{\tau_{sx}-\tau_{bx}}{\rho h} \\ &+ \underbrace{\frac{1}{\rho h}\frac{\partial}{\partial x}\int_{z_b}^{\zeta}\rho\left(U-\overline{U}\right)^2 dz + \frac{1}{\rho h}\frac{\partial}{\partial y}\int_{z_b}^{\zeta}\rho\left(U-\overline{U}\right)\left(V-\overline{V}\right)dz}_{\text{dispersion}}\end{aligned} \tag{A.5}$$

$$\begin{aligned}\frac{\partial \overline{V}}{\partial t} + \overline{U}\frac{\partial \overline{V}}{\partial x} + \overline{V}\frac{\partial \overline{V}}{\partial y} &= -g\frac{\partial \zeta}{\partial y} \\ &+ \frac{1}{\rho h}\frac{\partial (h\bar{\tau}_{xy})}{\partial x} + \frac{1}{\rho h}\frac{\partial (h\bar{\tau}_{yy})}{\partial y} + \frac{\tau_{sy}-\tau_{by}}{\rho h} \\ &+ \frac{1}{\rho h}\frac{\partial}{\partial x}\int_{z_b}^{\zeta}\rho\left(U-\overline{U}\right)\left(V-\overline{V}\right)dz + \frac{1}{\rho h}\frac{\partial}{\partial y}\int_{z_b}^{\zeta}\rho\left(V-\overline{V}\right)^2 dz\end{aligned} \tag{A.6}$$

$$\begin{aligned}\frac{\partial \overline{\phi}}{\partial t} + \overline{U}\frac{\partial \overline{\phi}}{\partial x} + \overline{V}\frac{\partial \overline{\phi}}{\partial y} &= \frac{q_s-q_b}{\rho h} + \frac{1}{\rho h}\frac{\partial (h\bar{J}_x)}{\partial x} + \frac{1}{\rho h}\frac{\partial (h\bar{J}_y)}{\partial y} \\ &+ \frac{1}{\rho h}\frac{\partial}{\partial x}\int_{z_b}^{\zeta}\rho\left(U-\overline{U}\right)\left(\phi-\overline{\phi}\right)dz + \frac{1}{\rho h}\frac{\partial}{\partial y}\int_{z_b}^{\zeta}\rho\left(V-\overline{V}\right)\left(\phi-\overline{\phi}\right)dz\end{aligned} \tag{A.7}$$

The closure of these equations involves the modelisation of the depth-averaged turbulent stresses and fluxes (such as $\bar{\tau}_{xx}$, $\bar{\tau}_{xy}$ or $\bar{J}_x$), of the dispersion terms (which result from vertical non-uniformities but NOT from turbulence) and of the free surface and bottom stresses and fluxes (τ_b, τ_s, q_b, q_s).

A.2 Nature of turbulence

Different models will be introduced in the next section. They correspond to a growing complexity in the description of turbulence motion, which tries to encompass more and more relevant features of this phenomena. Therefore, before looking at these models, it may be useful to recall the basic nature of turbulence (Hug, 1975; Rodi, 1980).

The turbulence motion can be thought as a tangle of eddies which stretch each other. The largest eddies are determined by the boundary conditions of the flow. They interact with the mean flow (as the scale of both are similar), extracting kinetic energy from the mean motion and feeding it into the large scale turbulent motion. Due to vortex stretching, the energy is passed on to smaller and smaller eddies until viscous forces become active and dissipate the energy. This process is called energy cascade.

It is important to note that viscosity does not determine the amount of energy dissipated (which is in fact controlled by the amount of energy extracted by the large scale turbulent motion) but only the scale at which dissipation takes place (i.e. the smallest eddies size).

Because of its interaction with the mean flow, the large scale motion has often preferred directions. It can therefore be strongly anisotropic : both the intensity of turbulent fluctuations and their length scale are direction dependent. During the cascade process, the direction sensitivity decreases and can even vanish when large-scale and small-scale motions are sufficiently apart (high Reynolds numbers) : the small-scale motion becomes isotropic *(local isotropy concept).*

It is mainly the large-scale turbulent motion that transports momentum, heat and mass and that contributes to the turbulent stresses and fluxes. Therefore, it is the large-scale motion that need to be simulated. Any quantity introduced to parametrize the turbulence (e.g. velocity or length scale) should be related to the large scale motion.

A.3 Examples of turbulence-closure models

A.3.1 Eddy-viscosity concept

Many models rely on the **eddy-viscosity diffusivity concept**, first introduced by Boussinesq. It assumes that, in analogy to viscous stresses in laminar flow, turbulent stresses are proportional to the mean-velocity gradients. Similarly, the turbulent heat or mass transport, as its molecular counterpart, is assumed to be proportional to the gradient of the transported quantity :

$$-\overline{u_i u_j} = \nu_t \left(\frac{\partial U_i}{\partial x_j} + \frac{\partial U_j}{\partial x_i} \right) - \frac{2}{3} k \, \delta_{ij} \tag{A.8}$$

$$-\overline{\varphi u_j} = \Gamma_t \frac{\partial \phi}{\partial x_j} \tag{A.9}$$

In the above equations, we have been introducing a new variable k, which is the kinetic energy per unit mass contained in the turbulent motion :

$$k = \frac{1}{2} \overline{u_i^2} \tag{A.10}$$

δ_{ij} stands here for the Kronecker symbol (nul when $i \neq j$, unity if $i = j$).

In contrast to their molecular counterparts, the turbulent eddy viscosity ν_t and eddy diffusivity Γ_t are no fluid property but depend strongly on the state of turbulence, and vary considerably over the flow field. The Reynolds analogy between heat or mass transport and momentum transport suggests that ν_t and Γ_t should be closely related. Indeed, it has been observed experimentally that their ratio varies but little across the flow field and from flow to flow (Rodi, 1980). Consequently, most models set Γ_t to

$$\Gamma_t = \frac{\nu_t}{\sigma_t} \tag{A.11}$$

with σ_t (termed the turbulence Prandtl or Schmidt number) as a constant.

On dimensional grounds, the eddy viscosity is assumed to be proportional to a velocity scale $\widehat{V}$ and a length scale L of turbulent motion :

$$\nu_t \propto \widehat{V} L \tag{A.12}$$

If relations A.8 and A.12 are trusted, the problem of turbulence modelling shifts to the problem of assessing the distribution of $\widehat{V}$ and L, or of some combinations of these variables. Albeit it as been criticized as physically unsound, the eddy viscosity concept has often worked well in practice. It is precisely due to the fact that $\widehat{V}$ and L can be approximated reasonably well in many flows, by the way of more or less complex formulations.

A.3.2 Mixing length model

The simplest models for ν_t relate it directly to the local mean-velocity distribution. The most popular model of this category is known as the Prandtl mixing length model. Considering shear layers with only one significant mean velocity gradient $\frac{\partial U}{\partial z}$ (z being the direction orthogonal to the wall), Prandtl postulated that $\widehat{V}$ is equal to the product of this gradient times a "mixing length" l_m. l_m also accounts for the length scale L. Thus, equation A.12 reduces to :

$$\nu_t = l_m^2 \left| \frac{\partial U}{\partial z} \right| \tag{A.13}$$

l_m need to be specified.

- In *free shear flows*, such as mixing layers, jets or wakes, l_m can be assumed constant across the layer and proportional to the local layer width δ. However, the proportionality factor depends on the type of the free flow considered (Rodi, 1980).

- In *wall boundary layers*, l_m is often described by a ramp function : in the near-wall region $l_m = \kappa z$ where z is the distance to the wall and κ is usually equal to the Von Karman constant ($\simeq 0.4$); in the outer region ($z \geq \lambda\delta$) l_m is proportional to the layer width.
- Empirical adaptations of the above mixing-length formula have been developped for duct flows, areas close to the walls where the viscous effects are no longer negligible, stratified fluid layers (Rodi, 1980) ...

Equations A.11 and A.13 may be combined in order to assess the diffusivity. The turbulent Prandtl number varies between 0.9 and 0.5 according to the shear layer type. It is affected by buoyancy (stratification) effects (Rodi, 1980).

From a physical point of view, the mixing-length model suffers from two major shortcomings.

1. As it relates ν_t only to *local instantaneous* mean flow characteristics, it does not and cannot account for any transport or history effect (namely, influence of turbulence generated elsewhere or at previous times). Consequently, it does not apply to flows where these effects are important (e.g. recirculating flows, tidal flows ...).
2. It predicts that ν_t and Γ_t vanish whenever the velocity gradient is zero. Let us consider for instance flow in a straight, rectangular open channel : the velocity gradient is nul at the centre line, where eq. A.13 would therefore forecast a nul diffusivity. Consequently, a mixing-length model would not allow the transfer of heat or mass from one wall of the channel to the opposite one. Such outcome is in contradiction with experiments and in situ observations.

From a practical point of view, l_m can be approximated both satisfactorily and economically (i.e. with simple formulas) only for some simple shear layer flows. Consequently, the use of mixing length model should be restricted to these simple situations.

A.3.3 $k - \epsilon$ model

One important step in the development of turbulence models was to give up the direct link between the fluctuating velocity scale and the mean velocity gradients and to determine this scale from a transport equation. If the velocity fluctuations are to be characterized by one scale only, the natural and most meaningful candidate is the kinetic energy k (cf eq. A.10) which provides a direct measure of the intensity of the fluctuating velocities. k is mainly contained in the large scale fluctuations (cf section A.2) and therefore $\sqrt{k}$ is a velocity scale appropriate for the large scale turbulent motion. When this scale is used in A.12, it results in :

$$\nu_t = C'_\mu \sqrt{k}\, L \tag{A.14}$$

where C'_μ is an empirical constant. Relation A.14 was independently introduced by Kolmogorov and Prandtl, with the accompanying equation governing k, which forms the core of the so-called "one-equation models". The exact equation for k is derived from the Navier-Stokes equations and contains, as these, unknown correlations which must be modelled (as explained for instance in (Rodi, 1980)) in order to achieve closure. The resulting form of the $k-$equation is (cf (ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988)) :

$$\frac{\partial k}{\partial t} + \underbrace{U_i \frac{\partial k}{\partial x_i}}_{\text{advection}} = \underbrace{\frac{\partial}{\partial x_i}\left(\frac{\nu_t}{\sigma_k}\frac{\partial k}{\partial x_i}\right)}_{\text{turbulent diffusion}} + \underbrace{\nu_t\left(\frac{\partial U_i}{\partial x_j} + \frac{\partial U_j}{\partial x_i}\right)\frac{\partial U_i}{\partial x_j}}_{\text{P = Production}} - \underbrace{C_\epsilon \frac{k^{3/2}}{L}}_{\epsilon = \text{dissipation}} + \underbrace{G}_{\text{buoyant production/destruction}} \tag{A.15}$$

Eq. A.15 is valid for high Reynolds numbers and is not applicable to the viscous sub-layer near walls. The production term P is the product of the turbulent stresses by the mean flow velocity gradients : it represents the transfer of energy from the mean motion to the turbulent motion. ϵ accounts for the dissipation of energy into heat by viscous action. The derivation of the buoyancy term G may be found in (Rodi, 1980). Stable stratification hinder vertical exchanges. In this case, G is negative reflecting the fact that the turbulence is damped. At the contrary, G is positive in case of unstable stratification.

In one-equation models, the length scale L, which appears both in A.14 and A.15, must be specified empirically in order to complete the determination of ν_t. This is done with the help of formulas (Rodi, 1980; Beckers, 1991) which bear frequently strong resemblance with those governing the mixing length and display unfortunately the same lack of universality. While satisfying applications have been reported, e.g. in the field of marine hydrodynamics (Nihoul & Djenidi, 1987; Nihoul *et al.*, 1989a; Beckers, 1991), as regards industrial flows, the scope of application of one-equation models did not appear to encompass much more cases than dealt with the simpler mixing-length model nor did they bring outstanding improvements (Rodi, 1980). Consequently, for calculating general (industrial, fluvial ...) flows, the trend has been to move on to two-equations models which determine also the length scale from a transport equation.

The length-scale determining equation need not have L itself as a dependent variable : any combination of the form $Z = k^m L^n$ will suffice as k is known already from solving the $k-$equation. The most popular combination is undoubtedly $k^{3/2}/L$ which physically represents the dissipation rate ϵ (cf eq. A.15). This choice has been criticized on physical grounds : indeed, dissipation occurs within the small-scale turbulent eddies while we are interested in a length scale associated with the large-scale, energy-containing eddies. However, the amount of energy dissipated at small scales is controlled by the energy fed from the large scale motion through the

energy cascade (cf sec. A.2). Hence, ϵ may nevertheless be considered a parameter characterizing the large-scale motion. Besides, the ϵ-equation, as presented further down, keeps relatively simple, especially as regards the prescription of boundary conditions.

By introducing k and ϵ instead of $\sqrt{k}$ and L, and relying on dimensional analysis, the equation A.14 determining ν_t is modified into :

$$\nu_t = C_\mu \frac{k^2}{\epsilon} \tag{A.16}$$

where C_μ is an empirical constant. The $\epsilon-$equation is semi-empirical as the $k-$equation : indeed, the exact equation derived from Navier-Stokes includes complex correlations whose behaviour is little known and which are modelled under somewhat drastic assumptions. It will be sufficient here to note that the $\epsilon-$equation accounts, besides advective and diffusive processes, both for the "generation of vorticity" through the energy cascade *(emergence of smaller eddies $\Rightarrow$ reduction of the effective length scale)* and for its viscous destruction *(disappearance of small eddies $\Rightarrow$ increase of the length scale).*
The result reads (ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988) :

$$\frac{\partial \epsilon}{\partial t} + \underbrace{U_i \frac{\partial \epsilon}{\partial x_i}}_{\text{advection}} = \underbrace{\frac{\partial}{\partial x_i}\left(\frac{\nu_t}{\sigma_\epsilon}\frac{\partial \epsilon}{\partial x_i}\right)}_{\text{diffusion}} + \underbrace{C_{1\epsilon}\frac{\epsilon}{k}(P + C_{3\epsilon}G) - C_{2\epsilon}\frac{\epsilon^2}{k}}_{\text{generation-destruction}} \tag{A.17}$$

where P is the production term defined in eq. A.15. Equations A.15 and A.17 constitute the famous $k-\epsilon$ model. A standard set of values for the empirical constants was determined both with the help of direct measurements in some specific situations and by computer optimization when applying the $k-\epsilon$ model to a number of well-documented laboratory shear flows (Rodi, 1980; ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988; Ouillon, 1993) :

$$C_\mu = 0.09 \quad C_{1\epsilon} = 1.44 \quad C_{2\epsilon} = 1.92 \quad \sigma_k = 1 \quad \sigma_\epsilon = 1.3 \tag{A.18}$$

The value of the constant $C_{3\epsilon}$ is not universal : it varies according to the state of stratification. A sensitivity study showed that the calculations are most sensitive to the values of $C_{1\epsilon}$ and $C_{2\epsilon}$ (Rodi, 1980).

Depth-averaged $k-\epsilon$ model The $k-\epsilon$ model was adapted by Rodi et al. (Rastogi & Rodi, 1978) for use in depth-averaged situations where the depth-averaged turbulent stresses $\bar{\tau}_{xx}$, $\bar{\tau}_{xy}$, $\bar{\tau}_{yy}$ (cf eq. A.5 and A.6) and mass fluxes $\bar{J}_x$, $\bar{J}_y$ (cf eq.A.7) need to be determined.

This model applies the eddy viscosity/diffusivity concept to the depth-averaged stresses or fluxes : it relates them to the horizontal gradients of the depth-averaged velocities and scalar concentrations. In analogy with the original $k-\epsilon$ model, (Rastogi & Rodi, 1978) assume that

the local *depth-averaged* state of turbulence can be characterized by two parameters, namely the turbulence energy $\tilde{k}$ and the dissipation rate $\tilde{\epsilon}$. The eddy viscosity $\tilde{\nu_t}$ is deduced from $\tilde{k}$ and $\tilde{\epsilon}$ by a relation strictly analogous to eq. A.16. The equations governing $\tilde{k}$ and $\tilde{\epsilon}$ read :

$$\frac{\partial \tilde{k}}{\partial t} + \overline{U}\frac{\partial \tilde{k}}{\partial x} + \overline{V}\frac{\partial \tilde{k}}{\partial y} = \frac{\partial}{\partial x}\left(\frac{\tilde{\nu_t}}{\sigma_k}\frac{\partial \tilde{k}}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\tilde{\nu_t}}{\sigma_k}\frac{\partial \tilde{k}}{\partial y}\right) + P_h + P_{kv} - \tilde{\epsilon} \qquad \text{(A.19)}$$

$$\frac{\partial \tilde{\epsilon}}{\partial t} + \overline{U}\frac{\partial \tilde{\epsilon}}{\partial x} + \overline{V}\frac{\partial \tilde{\epsilon}}{\partial y} = \frac{\partial}{\partial x}\left(\frac{\tilde{\nu_t}}{\sigma_\epsilon}\frac{\partial \tilde{\epsilon}}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\tilde{\nu_t}}{\sigma_\epsilon}\frac{\partial \tilde{\epsilon}}{\partial y}\right) + C_{1\epsilon}\frac{\tilde{\epsilon}}{\tilde{k}}P_h + P_{\epsilon v} - C_{2\epsilon}\frac{\tilde{\epsilon}^2}{\tilde{k}} \qquad \text{(A.20)}$$

(nb: the overbars or tilde denote depth-averaged quantities).

Eq. A.19 and A.20 can be considered as depth-averaged forms of the original $k-\epsilon$ equations, respectively eq. A.15 and A.17 . The production term P_h is similar to the production term P in eq. A.15 : it corresponds to the interaction of depth-averaged turbulent stresses with horizontal mean velocities gradients,

$$P_h = \tilde{\nu_t}\left(\frac{\partial \overline{U_i}}{\partial x_j} + \frac{\partial \overline{U_j}}{\partial x_i}\right)\frac{\partial \overline{U_i}}{\partial x_j} = \tilde{\nu_t}\left[2\left(\frac{\partial \overline{U}}{\partial x}\right)^2 + 2\left(\frac{\partial \overline{V}}{\partial y}\right)^2 + \left(\frac{\partial \overline{U}}{\partial y} + \frac{\partial \overline{V}}{\partial x}\right)^2\right] \qquad \text{(A.21)}$$

The depth-averaging causes the terms related to buoyancy effects to disappear. On the contrary, there appear two supplementary source terms P_{kv} and $P_{\epsilon v}$ which are supposed to account for anything originating from non-uniformities of vertical profiles. (Rastogi & Rodi, 1978) suggest that the main contribution to these terms comes from the near-bottom region where we find significant vertical velocity gradients (the streamwise velocities rise sharply from a zero value to near their depth-averaged mean) interacting with large shear stresses at the bottom-flow interface. Assuming the usual log-law and linear profiles respectively for the vertical distribution of velocity and shear stress, they propose :

$$P_{kv} = C_k\frac{U_*^3}{h} \qquad P_{\epsilon v} = C_\epsilon\frac{U_*^4}{h^2} \qquad \text{(A.22)}$$

where h is the local water depth and U_* the friction velocity so that the bed shear stress τ_b is equal to ρU_*^2. U_* is related to the mean velocities by the usual quadratic friction law $U_* = \sqrt{C_f\left(\overline{U}^2 + \overline{V}^2\right)}$ where C_f is a coefficient controlled by the bed roughness.

The empirical constants C_k and C_ϵ were determined by (Rastogi & Rodi, 1978) when considering the equilibrium reached in the center portion of a wide channel at normal steady-state uniform flow : in that case, every temporal and spatial gradients in eq. A.19 and A.20 are negligible so that they reduce to

$$P_{kv} - \tilde{\epsilon} = 0 \qquad \text{and} \qquad P_{\epsilon v} - C_{2\epsilon}\frac{\tilde{\epsilon}^2}{\tilde{k}} = 0 \qquad \text{(A.23)}$$

By combining A.22 and A.23 under the assumption of normal flow, we obtain that :

$$C_k = \frac{1}{\sqrt{C_f}} \qquad C_\epsilon = C_{\epsilon\Gamma} \frac{C_{2\epsilon}}{C_f^{3/4}} \sqrt{C_\mu} \tag{A.24}$$

The constants $C_{1\epsilon}$, $C_{2\epsilon}$, C_μ are the same than for the standard 3D $k-\epsilon$ model (cf A.18). The coefficient $C_{\epsilon\Gamma}$ was found to depend on the dimensionless diffusivity $e^* = \tilde{\Gamma}_t/U_* h = \tilde{\nu}_t/U_* h\sigma_t$ so that further empirical input is needed to determine C_ϵ.

For applications to laboratory flumes (Rastogi & Rodi, 1978; Rodi, 1980; Pavlovic & Rodi, 1985; Keller & Rodi, 1988), a satisfactory value for e^* appears to be 0.15, which corresponds to an average of dye-tracing laboratory experiments in wide flumes (Fischer *et al.*, 1979) : this leads to $C_{\epsilon\Gamma} \simeq 3.6$.
Application to natural channels requires a different tuning (Rodi *et al.*, 1981; ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988). Indeed, this depth-averaged $k-\epsilon$ model has been applied in relation to a mean flow depth-averaged model (eq. A.4 to A.7) where the dispersion terms were simply dropped. Thus, the actual depth-averaged "turbulent stresses and fluxes" must account implicitly for dispersion effects. These are more marked in natural rivers than in flumes : better agreement with field data was obtained when setting e^* to a typical value for slightly irregular channels, namely $e^* = 0.6$, for which $C_{\epsilon\Gamma} \simeq 2$ (Rodi *et al.*, 1981; ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988).
Finally, relations A.22 become

$$P_{kv} = \frac{1}{C_f^{1/2}} \frac{U_*^3}{h} \qquad P_{\epsilon v} = \frac{\alpha}{C_f^{3/4}} \frac{U_*^4}{h^2} \quad \text{with } \alpha \simeq 2.074 \text{ for flumes, } \alpha \simeq 1.152 \text{ otherwise} \tag{A.25}$$

Other modifications of the empirical constants, e.g. of C_k (Booij, 1989), have occasionally been suggested as well as slightly different formulations of the depth-averaged turbulence equations (e.g. (Dong *et al.*, 1989; Liren & Shu-nong, 1989)) but, to our knowledge, none has been as thoroughly applied and tested than Rodi's model.

Comments Two-equations models are the simplest available ones that may be applied with reasonable hope of success to complex flows where the turbulence length scale cannot be specified empirically in a straightforward way. The $k-\epsilon$ model is probably the most widely tested and successfully used among turbulence models. It has been applied to many different flow situations, yet perhaps more in mechanical and aeronautical engineering than in water surface problems. The same set of empirical constants proved suitable for a number of different flows. In particular, the standard model performs well for shear-layer flows and confined recirculating flows. According to (Rastogi & Rodi, 1978; Rodi, 1980; Rodi *et al.*, 1981; ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988), its depth-averaged version appears

promising, notably for near-field calculations involving the interaction of turbulence generated both at the river bed and by the discharge itself. However, there exist cases where the model had to be tuned differently (e.g. modification of C_μ for weak shear flows (Rodi, 1980), or of σ_k in canals, as reported in (Ouillon, 1993)) or where it displays systematic failures (e.g. underprediction of the separation regions in unconfined recirculating flows). In those cases, the blame was mostly put on the present form of the $\epsilon-$ equation.

One obvious drawback of two-equation models is that they assume an isotropic eddy viscosity/diffusivity (cf eq. A.8) and that they relate the turbulent stresses and fluxes to an unique velocity scale. In certain situations, the fluctuations, and hence the stresses and their scales, do develop differently for different directions and components (cf sec. A.2) : for instance, in shallow water bodies, the horizontal motion has an intensity and length scale greater than the vertical motion. Whenever this anisotropy is important (e.g. in turbulence driven secondary motions, cf sec. A.4.2), the only possible move is to drop the eddy viscosity concept and to gó for models which describe directly the behaviour of the individual stresses and fluxes.

A.3.4 Turbulent stresses/flux equation models

Models which employ individual transport equations for the turbulent stresses $\overline{u_i u_j}$ (and the accompanying turbulent mass fluxes $\overline{\varphi u_j}$) theoretically allow for different rates of development of the various stresses.

Exact equations for $\overline{u_i u_j}$ can be derived from the Navier-Stokes equations. Their somewhat simplified expression (obtained by neglecting the buoyancy term and the molecular diffusion term) reads (Rodi, 1980; Ouillon, 1993) :

$$\begin{aligned}
\frac{\partial \overline{u_i u_j}}{\partial t} + \underbrace{U_l \frac{\partial \overline{u_i u_j}}{\partial x_l}}_{\text{advection}} &= - \underbrace{\frac{\partial \overline{u_i u_j u_l}}{\partial x_l}}_{\text{diffusive transport}} \\
&- \underbrace{\left(\overline{u_i u_l} \frac{\partial U_j}{\partial x_l} + \overline{u_j u_l} \frac{\partial U_i}{\partial x_l} \right)}_{\text{correlation with mean flow}} \\
&- \underbrace{\frac{1}{\rho} \left(\overline{u_i \frac{\partial p}{\partial x_j}} + \overline{u_j \frac{\partial p}{\partial x_i}} \right)}_{\Pi_{ij}=\text{correlation with fluctuating pressure}} \\
&- \underbrace{2\nu \overline{\frac{\partial u_i}{\partial x_l} \frac{\partial u_j}{\partial x_l}}}_{\epsilon_{ij}=\text{viscous dissipation}}
\end{aligned} \tag{A.26}$$

These equations include high-order correlation terms which need to be modelled. The model

for the diffusive term yields a diffusive tensor (as in the k and ϵ equations). The model for the dissipation term relies on the local isotropy concept (cf section A.2) and allows to express it as a simple function of the energy dissipation rate ($\epsilon_{ij} = \frac{2}{3}\epsilon\,\delta_{ij}$). The main difference between different turbulent stress equation models generally lies in the approximation of the Π_{ij} factor (termed the pressure-strain correlation). It is possible to eliminate the fluctuating pressure p from eq. A.26 : by doing so, it appears that three processes contribute to Π_{ij}, the first related to the interaction between fluctuating velocities, the second arising from interaction between mean strain and fluctuating velocities, the third due to buoyancy forces. Each contribution is usually modelled separately (Rodi, 1980; ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988). It eventually includes corrections related to wall or free surface effects.

The equations for the scalar fluxes $\overline{u_j\varphi}$ are similar to A.26 (Rodi, 1980). The equation for the scalar fluctuations intensity $\overline{\varphi^2}$ is simpler and analogous to the equation which governs the velocity fluctuations intensity (namely the kinetic energy k) except that it does not include buoyancy terms (Rodi, 1980).

The application of a turbulent stress/flux equations model may prove burdensome. Indeed, in general flows, there are 6 components of the stresses $\overline{u_iu_j}$ and 3 components for the fluxes $\overline{\varphi u_j}$. Considering two additional equations - for the dissipation rate ϵ and the scalar fluctuations intensity $\overline{\varphi^2}$ - this represents a set of eleven partial differential equations. Suggestions were therefore made to simplify them. The principle of these simplifications lies in the approximation of the gradients of the dependent variables, which appear in the advective and diffusive terms : by eliminating these terms, the equations reduce to algebraic expressions (total difference equations) and are consequently much easier and more economical to solve.

Two successive reviews (Rodi, 1980; ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988) give an account of the (full or algebraic) stress/flux equations models which have been so far applied in connection with surface water models.

A.3.5 Other methods

More advanced turbulence modelling strategies (e.g. Large Eddy Simulation (Lesieur, 1993)) are being developped and used, but mostly in the field of industrial applications. Some examples may be found for instance in (IAHR, 1993).

While the methods presented above aim at computing *statistics* of the turbulent motion (mainly second-order moments), methods as Large Eddy Simulation (LES) involve the *direct simulation* of some part of the turbulence spectrum. This approach has formal similarities with

the averaging of the Navier-Stokes equations which leads to the Reynolds mean flow equations, except that the "averaging" is not temporal but spatial : the turbulent velocities are thought as the sum of components which wavelength is either superior or inferior to some length Δx. The averaging (generally termed "filtering") introduces correlations which depend on turbulent fluctuations occurring at a scale smaller than the computational mesh size Δx : a *subgrid-scale* model representing their action upon the explicitly resolved scales must then be introduced. While the Reynolds equations involve a mean flow which varies rather smoothly and slowly in space and time, the LES equations involve a "mean flow" which may be extremely chaotic if Δx is small enough. Repeated direct simulations of the turbulence motion may eventually be used to yield statistics relative to the state of turbulence.

It is worth noting that such higher level simulations of turbulent flows involve a greater interaction between the task of closing the equations and the task of solving them by appropriate numerical techniques (Dervieux & Palmerio, 1993; Chen *et al.*, 1993; Ruger & Sommerfeld, 1993; Moulin *et al.*, 1993).

A.4 Turbulence processes in fluvial hydraulics

The choice of an adequate turbulence model is obviously problem dependent.

Ideally, a prior analysis of the main flow features should always precede the choice of a modelling tool. An interesting attempt at providing a framework for classifying turbulent flows, and consequently how they ought to be modelled, can be found in a paper of Hunt (Hunt, 1993). The proposed classification relies on the examination of the type of boundaries (closed or open), of the boundaries conditions (e.g. state of turbulence and existence of preferred directions in the entering flow), of the initial conditions in the flow domain (are the initial state of turbulence and boundary conditions in a state close to equilibrium or not ?) and of the general eddy structure. We shall give hereafter an example of the kind of theoretical analysis conducted by Hunt.

Hunt introduces for instance a "relaxation time" T_L for the turbulence inside the studied domain. This is a measure of the time overwhich some disturbance to the turbulence structure decays. On dimensional grounds, T_L can be related to the ratio of the eddy length scale by the average intensity of fluctuating velocities. By comparing the relaxation time to the average residence time or travel time of a fluid particle within the domain (T_D , which is controlled by the size of the domain in the main direction of the flow and by the average *mean flow* velocity) it is possible to understand to what extent the inside turbulence depends on the imposed entering flow :

- If T_L is much bigger than T_D, the turbulent features of the entering flow must be described

in full detail and every history and transport effect should be explicitly accounted for.

- If T_L and T_D are of the same order (which is the case in most unconfined turbulent flows (Hunt, 1993)), the inside turbulence will display some sensitivity to initial and boundary conditions but far less than in the first case, so that the representation of turbulence effects with the help of some low-order statistical features (e.g. k and ϵ) should prove sufficient.

- Lastly, there are cases when $T_L \ll T_D$. In such cases, past and "upstream" turbulence do not affect the local normal turbulence and it sounds justified to relate it only to local characteristics of the mean flow, as in the mixing length formulation of Prandtl. Examples of such flows are indeed boundary layers near rigid walls. The turbulent motion is confined in the direction normal to the wall, but not in the main direction of the flow, parallel to the wall : the difference of scales between the "normal" and "parallel" motion implies that the relaxation time characterizing the normal turbulent motion is much smaller than the travel time of fluid particle along the wall.

Albeit appealing from an intellectual point of view, the approach advocated by Hunt is certainly more easily put in practice in the field of industrial fluids, mechanical or aeronautical engineering, than for environmental flows, where experiments are a lot more difficult to implement, control, monitor and repeat. Thus, the main source of information concerning the performance and relevance of turbulence models is the literature.

A thorough attempt at exposing the current status of solutions brought by turbulence modelling to the simulation of surface water flows has recently been achieved in the frame of an ASCE review (cf (ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988), part 5). From this review, it stems that application of advanced turbulence modelling to typical environmental problems is quite sparce with respect to its application to hydraulic engineering problems. As could be expected, there is no general closure model which performs well in any situation. Generally, the models applied belong to the $k - \epsilon$ model family, which are the least costly to integrate in an hydraulic code. It is interesting to note that, according to the case studies, very little or none empirical coefficient adjustment with respect to the values published in literature was performed. This could result in savings as regards the cost of data collection needed to tune more simpler formulations.

As regards **fluvial hydraulics and mass transport**, the series of studies achieved by Rodi and co-workers is particularly enlightening in that it helps to figure out what is the limit of applicability of depth-averaged models and what is the right level of turbulence modelling for some practical problems.

Rodi et al. studied more especifically the near-field of discharges and the influence of secondary motion in the cross-sectional plane upon the development of flow and transport in the streamwise direction.

A.4.1 Near-field calculation of discharges

Here the near-field is defined as the region where the discharge influences the river flow field. Its extent depends on the river and discharge features (flow ratios, discharge geometry and buoyancy, etc ..) : in wide and shallow rivers, its order of magnitude is usually a few hundred meters to one kilometer long.

- In 1978, Rastogi and Rodi (Rastogi & Rodi, 1978) sum up the results of their investigation about the fate of steady coaxial slot-discharges. They compare the outcome of a 2D $k-\epsilon$ depth-averaged (DA) model and of a 3D $k-\epsilon$ model with respect to available experimental laboratory data. As mentionned above (cf A.3.3), the 2D model neglects the dispersion terms, which are calculated explicitly by the 3D model, so that the justification for their neglect can be examined directly. The effect of buoyancy in the 3D model is accounted for by the inclusion of a buoyancy term in the mean flow equations (see eq. A.2) but the accompanying $k-\epsilon$ model neglects the direct impact of buoyancy on turbulence structure.

  The performances of 2D and 3D calculations depend on the densimetric Froude number

  $$F = U_d / \sqrt{gh|\rho_d - \rho_r|/\rho_r}$$

  where ρ_r and ρ_d denote respectively the river and discharge density, h is the water depth and U_d the discharge velocity.

  In smooth channels, the lower limits of applicability of the 3D and 2D models appear to be respectively $F \simeq 5$ and $F \simeq 10$. When the channel is fairly rough, however, the bottom stresses increase the vertical mixing, which counteracts the secondary motion and stratification induced by buoyancy : 2D and 3D predictions agree even at $F \simeq 5$. At high Froude numbers, both 3D and 2D calculations perform well for various velocity ratios and flume roughnesses.

  The same 3D $k-epsilon$ model is extended in order to deal with the case of side isothermal discharges (Demuren & Rodi, 1983). A special emphasis is put on the description of the subsequent wall and open boundaries conditions for the turbulence model. Besides, by using different numerical algorithms to solve the equations, the authors highlight how numerical inacurracies can blunder the model outcomes and interfere with conceptual modelling. The sumulations conducted with the high-order QUICK scheme (Leonard, 1979) agree fairly well with the flume experiments for both smooth and rough bed conditions.

- In 1981, Rodi et al. ((Rodi *et al.*, 1981), see further calculations and comments in (ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988)) report the application of a chain of 2D-DA models to the simulation of full scale side discharges under steady-state flow conditions. The first case study deals with sewage effluent into the River Neckar, the second one with a thermal effluent into the River Rhine.

  - Both reaches downstream the discharges are fairly straight. However, in the river Neckar, there is a slight bend upstream the discharge which might induce some residual secondary motion increasing the pollutant spreading and, in the river Rhine, there are groynes on the bank opposite to the discharge which might similarly introduce some specific circulation patterns in the cross section.
  - Both rivers have an important width to depth ratio (about 20 for the river Neckar, more than 50 for the river Rhine). Contrary to the river Rhine (average velocity varying between 1 and 2 m/s according to flow), the river Neckar is slow flowing (average velocity 0.15 m/s).
  - In the example related to river Neckar, the sewage flow amounts to 10 % of the river flow and the ratio of discharge velocity to average river velocity is 5. Measurements (depth-averaged velocity and concentration profiles) and subsequent comparisons have been made 410, 910, 1410 m downstream the discharge and, for velocities only, 10 m downstream and 290 m upstream.
  - Four different discharge conditions have been investigated in the Rhine river. The flow discharge was but a few percents (0.8 to 2.4) of the river Rhine flow. The excess temperature of the discharge ranges between 5.4 and 8.3 degrees. Most of the measurements consisted in temperature depth-averaged profiles, a few dealt with velocity profiles. In two cases, monitoring took place 100,200,300, 700 and 800 m downstream the discharge. In the other cases, only the near-field (within 120 m of the discharge) was surveyed.

  The model-chain components are respectively : a full 2D $k-\epsilon$ elliptic model, a 2D $k-\epsilon$ parabolic model in which the equations for the mean flow are simplified by neglecting both transverse velocities and longitudinal transport of momentum and mass, a 2D diffusion model where turbulence is accounted for by a diffusion tensor. The $k-\epsilon$ model is tuned differently than in flumes in order for the turbulent stresses and fluxes to account for stronger dispersion effects (cf A.3.3).

  The application of the models chain to the River Neckar is satisfying. The case of the river Rhine is more thoroughly discussed. The results are correct too, except for the highest flow rate : in that case, the discharge was completely submerged and this seems to induce 3D effects that cannot be properly accounted for. In the other cases, with discharges located at or close to the surface, the depth-averaged $k-\epsilon$ model appears to be accurate enough,

even in the immediate vicinity of the outfall. Finally, the results in the far-field (700 or 800 m downstream the outfall) of the models chain would match the results of a simple diffusion model (the last component of the chain) applied overall the studied reach (ASCE Task Committee on Turbulence Models in Hydraulic Computations, 1988).

Other satisfying applications of depth-averaged models to the simulation of thermal discharges in rivers have been reported. For instance, (Lebosse, 1987) investigated the fate of a cooling discharge from a nuclear power plant in the river Loire. The studied reach was 8 km long, beginning 200 m upstream the outfall. The water depth ranged from 0.5 to 1.4 m while the river width was about 200 m. Contrarily to the cases studied by Rodi, the discharge (60 $m^3.s^{-1}$) represented this time a significant contribution to the river flow (110 $m^3.s^{-1}$). The turbulence and mean-flow model is analogous to Rodi's one, except that no simplifications are introduced in the mean-flow equations A.4 to A.7. Simulations are consistent with the observed thermal field (measured by thermography) except very locally, at the vicinity of an island.

A tentative conclusion based on the above literature review is that *for the near-field and intermediate-field calculation of discharges into moving streams, a 2D $k-\epsilon$ depth-averaged model will generally prove adequate enough.*

A.4.2 Secondary motion in open channel flows

Secondary motion has been observed in any kind of cross section (Chow, 1959; Hug, 1975). As it is not present in laminar flows, its development has been ascribed to the inequality between turbulent correlations induced by vertical and transverse velocity fluctuations within the cross section. The influence of turbulence-driven secondary motion is twofold :

- *It influences the primary flow.*

  A first example of its influence is the usual depression of longitudinal velocity at the free surface (i.e. the maximum velocity point generally lies below the surface, at a distance of 0.05 to 0.25 of the total depth (Chow, 1959)). This can be traced back to the existence of the secondary motion which, in straight reaches, near the free surface, is directed away from the banks and transports fluid with low longitudinal momentum towards the central portion of the channel. This secondary motion effect is most perceptible as the width to depth ratio (B/h) of the channel decreases. Ven Te Chow (Chow, 1959) advocates to neglect it for straight and very wide ($B/h \geq 10$) open channels.

  The secondary motion organization may be mono or multi-cellular (Naot & Rodi, 1982; Shiono & Muto, 1993; Nezu *et al.*, 1993b). Secondary motion is partly controlled by cross

section irregularities and roughness (Nezu *et al.*, 1993a) and medium-scale bed features (e.g. dunes, cf (Lyn, 1993; Nezu & Nagakawa, 1993)).

Secondary motion is also increased by bends. Measurements (references given in (Demuren & Rodi, 1986)) of the development of longitudinal velocity in curved channels have shown that the velocity maximum usually occurs near the inner bank at the inlet to the bend. Within the bend, the secondary motion (termed *spiral motion*) transports high momentum fluid of the upper layers to the outer bank and low momentum fluid of the near-bottom layers towards the inner bank, eventually leading to a shift of the velocity maximum to the outer bank at the outlet of the bend. This secondary motion is observed to decay but slowly downstream the bend. This can eventually lead to very complex flow patterns in meandering channels : the sense of the spiral motion is reversed in subsequent bends so that the residual motion from the previous bend counteracts the setting up of the spiral in the next bend ...

- Consequently, *it influences heat and mass transfer and hence the distribution of pollutants.*

  Viewed globally, this influence increases the mixing but the actual process is not that of turbulent mixing but of *convective transport* by the secondary motion. In depth-averaged models, these effects are represented by the dispersion terms. Already noticeable in mildly irregular channels, they are especially marked in meandering ones.

  An other aspect of secondary motion, namely the development of turbulent structures linked to bed discontinuities, should be correctly appraised in order to achieve a better understanding of erosion and sediment transport phenomena.

Calculation of secondary currents is of course possible only within the frame of a 3D hydraulic model. Rodi and co-workers tested different possible closures.

In a paper of 1982 (Naot & Rodi, 1982), they show that a simplified algebraic-stress model (ASM) seems sufficient enough to predict the secondary motion. This ASM includes an elaborate model only for the turbulent stresses in the cross-sectional plane. In contrast, the stresses which contribute to the longitudinal transport of momentum are dealt with by refering to the simpler eddy-viscosity concept. A particular emphasis is put on the representation of the influence of the free surface on the vertical turbulent motion (the free surface basically dampens the turbulence and imposes geometrical restrictions to the eddy scale). This model is favorabily compared to a few available experimental data relative either to velocity contours, eddy-viscosity or fluctuations distributions, and to the outcome of a full stress model. It is then employed to study qualitatively the evolution of flow patterns according to various width to depth ratios.

A somewhat simpler closure model is applied to predict velocity and concentration fields in two experimental meandering channels (Demuren, 1983; Demuren & Rodi, 1986). The hydraulic

and turbulence closure models are written in cylindrical polar coordinates. Turbulent transport of momentum and mass in the streamwise direction is neglected. The closure model is a simple $k - \epsilon$ model with an empirical correction in the transverse direction for the eddy-viscosity and diffusivity, which are assumed to depend on the streamlines curvature and on the transverse gradient of longitudinal velocity.

The experiments concern both a wide, shallow ($B/h = 20$) and smooth channel with one meander and a narrow ($B/h = 5$), seven meanders, channel with smooth then rough bed. For the first channel, vertical profiles of both longitudinal and transverse velocities were available at several cross sections. For the second channel, the data included vertical profiles of longitudinal velocity and transverse profiles of depth-averaged longitudinal velocity and dye concentration. The agreement between measurements and computations in the wide channel is fairly good. The agreement concerning vertical profiles is less satisfactorily in the narrow channel, especially for the rough bed case. However, the tranverse depth-averaged profiles are approximated fairly well.

Outcomes of the two-dimensional $k - \epsilon$ depth-averaged model (cf A.3.3) have also been compared to these experiments (Pavlovic & Rodi, 1985). While no modification nor different calibration (than proposed in (Rastogi & Rodi, 1978; Rodi, 1980)) was necessary to reproduce satisfactorily the velocity profiles, it appeared that a term representing transverse dispersion needed to be included explicitly in the scalar transport equation. This dispersion term is, as usual (cf appendix B.2), modelled by a diffusion-like operator. The results of 3D modelling have been used to build an empirical formula (eq. B.22 in B.3.2) which relates the corresponding dispersion coefficient to the average and friction velocities, width to bend radius ratio and longitudinal location within the bend (Pavlovic & Rodi, 1985).

A similarly detailed experimental and modelling work is reported by (Mary, 1982) : it deals with the fate of an heated discharge into an S-shaped flume with ratio $B/h \simeq 7$, the discharge being distributed all over the water column. The study aimed at the validation of a depth-averaged two-dimensional St-Venant model, i.e. relying on a diffusion operator to represent both turbulence and dispersion terms. No refined turbulence model was used because the turbulent terms were deemed to be negligible, in the momentum equations with respect to the pressure and advective terms, in the heat equation with respect to the dispersive contributions. As in (Pavlovic & Rodi, 1985), the corresponding thermal dispersion coefficients are calculated from analytical expressions relating them to the vertical distribution of velocity. Yet, this time, the velocity distribution is not computed by a 3D model but is assumed to obey experimental formula previously established (Mary, 1982) (sec. 2.5 and appendix A.2). First, different trials were conducted in order to identify a suitable range of experimental conditions (i.e. devoid from the development of a marked and lasting stratification). One experiment is then carefully dis-

cussed. Measurements took place in 12 cross-sections (respectively 7 and 12 locations across the section for longitudinal velocity and temperature, 3 points at different depths of the water column at each location). A reasonable agreement between measures and simulations is observed, without any tuning of the dispersion coefficients.

The above investigations are limited to open channels with rectangular cross section. However, more complicated cross-section shapes have been studied, especially in the objective of determining the flow features in compound channels made of a main channel and a flood plain. The subject is rather controversial. The modelling of compound flows is often tackled with the help of mono-dimensional or two-dimensional St Venant models which ignore the explicit influence of turbulence and secondary motions on the flow development and on the setting up of exchanges (of mass and momentum) between main channel and flood plain. Such effects are implicitly lumped with the bed resistance coefficient when the model is calibrated against existing data. Rodi (Keller & Rodi, 1988) casts doubt on the portability of such models for forecasting very different flood conditions.

Developments of more sophisticated models rely heavily on laboratory data, as monitoring the velocity field on a true flood plain is still a challenge ! The experiments have been dealing mainly with straight flumes. Till now, the results are relatively inconclusive. While some researchers claim that algebraic stress turbulence models (ASM) are indeed the right tool to simulate compound channel behavior (cf (Krishnappan & Lau, 1986; Naot *et al.*, 1993; Naot & Nezu, 1993)), some others (Prinos, 1993) note that complex ASM do not perform much better than simpler ASM or even standard $k-\epsilon$ model as regards the prediction of longitudinal (streamwise) bed stresses and (cross-section or depth-averaged) longitudinal flow velocity (which are determinant for erosion process for instance). Some (Keller & Rodi, 1988) even state that three-dimensional models go far beyond what is called for in practical applications and that $k-\epsilon$ depth-averaged models provide a reasonable assessment of velocities and stresses, apart from quite limited areas of the cross-section (i.e. around the steep step between main channel and flood plain).

Notations

- t time
- x, y, z coordinates (the directions x, y, z correspond respectively to the x_1, x_2, x_3 directions in tensor notations)
- U, V, W (also denoted U_1, U_2, U_3 in tensor notations) temporal mean velocities in x, y, z directions
- u, v, w (also denoted u_1, u_2, u_3 in tensor notations) turbulent fluctuating deviations from these mean values
- $\overline{U}$, $\overline{V}$ depth-averaged velocities in x and y directions
- P temporal mean of static pressure
- p fluctuating pressure
- ϕ temporal mean of scalar (concentration or temperature), φ the associated turbulent fluctuation
- $\overline{\phi}$ depth-averaged concentration or temperature
- g magnitude of gravitational acceleration
- ρ fluid density (the subscript r denotes a reference value, e.g. in a river with respect to a discharge)
- z_b bed elevation, with respect to a reference level
- h water depth
- $\zeta = h + z_b$ free surface elevation, with respect to a reference level
- q_s heat or mass flux (source or sink term) through surface
- τ_b bottom shear stress
- U_* bed friction velocity, defined by $\tau_b = \rho U_*^2$
- C_f friction coefficient. When the friction is expressed by a quadratic law, τ_b modulus is computed by the relationship $\tau_b = \rho C_f \left(\overline{U}^2 + \overline{V}^2\right)$. C_f depends on the bed roughness.
- τ_s surface shear stress
- $\bar{\tau}_{xx}$ ($\bar{\tau}_{xy}$, $\bar{\tau}_{yy}$) depth-averaged value of the turbulent stress $-\rho\overline{u.u}$ (respectively $-\rho\overline{u.v}$ and $-\rho\overline{v.v}$).

- $\bar{J}_x$ ($\bar{J}_y$) depth-averaged value of the turbulent heat or mass fluxes ($-\rho\overline{\varphi.u}$ and $-\rho\overline{\varphi.v}$), respectively in the $x-$ and $y-$ directions
- ν and λ respectively kinematic molecular viscosity and molecular diffusivity
- ν_t and Γ_t respectively turbulent eddy viscosity and eddy diffusivity
- σ_t turbulent Prandtl or Schmidt number
- $\widehat{V}$ turbulence velocity scale
- L turbulence length scale
- l_m Prandtl mixing length
- κ Von Karman constant
- k kinetic turbulent energy per unit mass
- ϵ dissipation rate of k
- C_μ, $C_{1\epsilon}$, $C_{2\epsilon}$, $C_{3\epsilon}$. σ_k, σ_ϵ empirical constants related to the $k-\epsilon$ model
- C_k, C_ϵ, $C_{\Gamma\epsilon}$ empirical constants in the depth-averaged $k-\epsilon$ model
- B river width

Appendix B

Dispersion processes

This appendix deals only with dispersion modelling in the field of fluvial hydraulics, for non-stratified flows. As stated above, in this context, flows are mostly two-dimensional in plan and their vertical component may often be discarded. Besides, bed friction appears to play a dominant part in the generation of turbulence and secondary motions. As we shall see later on, this feature is reflected in the various empirical formula which aim at assessing dispersive parameters.

Moreover, we focus on the intermediate and far-field downstream discharges, i.e. areas where the spreading is mainly controlled by the receiving water body hydraulics and not by the discharge intrinsic dynamics.

An introduction to dispersion modelling in other situations may be found for instance in (Fischer *et al.*, 1979) (mixing in unconfined water bodies like oceans, mixing in reservoirs, dilution of jets and plumes in coastal waters) or (Tassin, 1986; Vinçon-Leite, 1991) (vertical mixing in lakes) ...

B.1 Dispersion : an advective process

From a fundamental point of view, the primary mechanism which governs the transport of pollutants in water bodies is advection (cf eq. 2.3). Advection is responsible for moving species downstream a river and differences in advection with respect to either time or space are responsible for most of the spreading of a pollutant. As stated above (cf. section 2.1.1), the resolution of the exact equations for scalar transport is not feasible for practical engineering problems. Thus, various parts of the advective transport have traditionally been approximated.

- When time-averaged representations are used, the net advection associated with the turbulent temporal fluctuations of velocity and concentration (cf eq. 2.7) is usually represented as turbulent diffusion (cf. section 2.2).
- When depth-averaged representations are used (cf. for instance eq. 2.12), there appear in the equations governing the mean flow quantities so-called "dispersion terms" which represent the spreading effects due to the variation over the depth of concentration and either longitudinal or transverse velocities.
- In 1D cross-section averaged representations, the "dispersion terms" account for the effect of any variation over the cross-section of concentration and streamwise velocity.

In practical situations, the available data frequently do not allow to distinguish the process of turbulent mixing from dispersion effects, the latter proving quite bigger in some specific situations (e.g. river bends). This is particularly true as regards the survey and simulation of large water bodies. In such case, a closure model which accounts for both kind of process has to be introduced.

In the present appendix, we shall first examine on which grounds the dispersion terms have been modelled by a diffusion operator and we shall give an overview of the empirical formulas commonly in use to express the associated diffusion coefficients, named dispersion coefficients.

B.2 Theoretical justification for the use of dispersion coefficients

Taylor was the first one, in 1954, to justify the introduction of a one-dimensional dispersion coefficient for representing the mixing at asymptotically large times. He did so by studying laminar flow in a pipe. Fischer (Fischer *et al.*, 1979) showed that Taylor's analysis can apply to a wider range of flows. We recall hereafter the outlines of his demonstration.

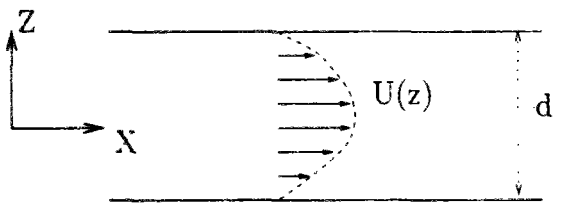


Figure B.1: Two-dimensional flow between parallel plates

Let us consider the two-dimensional flow sketched in figure B.1. All the streamlines are parallel to the walls which guide the flow. We assume the flow to be uniform, steady and laminar.

Let U be the flow velocity in the $x-$direction, $\overline{U}$ its average over the distance d between the walls, and u the deviation from this mean. The flow transports a species which concentration C is a function of x, z and t. Similarly, we introduce $\overline{C}$, mean of C over the distance d and c deviation from this mean. As the flow is laminar and unidirectional, the equation governing C is :

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} = \lambda\left[\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial z^2}\right] \tag{B.1}$$

where λ stands for the molecular diffusivity. Eq. B.1 is modified by developing C as the sum of $\overline{C}$ and c and by shifting to a coordinate system whose origin moves at the mean flow velocity :

$$\xi = x - \overline{U}t \qquad\qquad \tau = t$$

The resulting equation is :

$$\frac{\partial(\overline{C}+c)}{\partial \tau} + u\frac{\partial(\overline{C}+c)}{\partial \xi} = \lambda\left[\frac{\partial^2(\overline{C}+c)}{\partial \xi^2} + \frac{\partial^2 c}{\partial z^2}\right] \tag{B.2}$$

In the streamwise direction, the effect of molecular diffusivity can be neglected with respect to the effect of advection, so that eq. B.2 becomes :

$$\frac{\partial(\overline{C}+c)}{\partial \tau} + u\frac{\partial(\overline{C}+c)}{\partial \xi} = \lambda\frac{\partial^2 c}{\partial z^2} \tag{B.3}$$

Averaging eq. B.2 over the distance d yields the following equality

$$\frac{\partial \overline{C}}{\partial \tau} + \overline{u\frac{\partial c}{\partial \xi}} = 0$$

which can be substracted from eq. B.3, so that we finally obtain

$$\frac{\partial c}{\partial \tau} + u\frac{\partial \overline{C}}{\partial \xi} + u\frac{\partial c}{\partial \xi} - \overline{u\frac{\partial c}{\partial \xi}} = \lambda\frac{\partial^2 c}{\partial z^2} \tag{B.4}$$

The analysis may go on provided we are sufficiently far away from the species source, so that $\overline{C}$ and c are smoothly varying functions and that c is much smaller than $\overline{C}$. In that case, we can assume that the third and fourth terms in the left-hand side of eq. B.4 nearly balance each other and are anyway much smaller than the second term. Thus, the distribution of the deviations c over a cross-section is governed by a diffusion equation

$$\frac{\partial c}{\partial \tau} - \lambda\frac{\partial^2 c}{\partial z^2} = -u\frac{\partial \overline{C}}{\partial \xi} \tag{B.5}$$

in which term $-u\frac{\partial \overline{C}}{\partial \xi}$ acts as an external source. If C is so smoothly varying that $\frac{\partial \overline{C}}{\partial \xi}$ can be deemed constant over significant time intervals, c reaches a steady-state equilibrium

$$\lambda\frac{\partial^2 c}{\partial z^2} = u\frac{\partial \overline{C}}{\partial \xi} \tag{B.6}$$

whose solution reads :

$$c(z) = \frac{1}{\lambda}\frac{\partial \overline{C}}{\partial \xi}\int_0^z \int_0^z u\,dz\,dz + c(0) \tag{B.7}$$

For an entire cross section, the rate of mass transport in the streamwise direction due to the deviation from the mean flow is :

$$\frac{\partial M}{\partial t} = \int_0^d uC\,dz = \int_0^d uc\,dz$$

From eq. B.7 it stems that

$$\frac{\partial M}{\partial t} = \frac{1}{\lambda}\frac{\partial \overline{C}}{\partial \xi}\int_0^d u \int_0^z \int_0^z u\,dz\,dz\,dz \tag{B.8}$$

It results that **the mass transport in the streamwise direction is proportional to the mean concentration gradient in the streamwise direction.** This is similar to what we observ for molecular diffusion, so that, in analogy to the law of Fick, we can introduce a dispersion coefficient K defined by :

$$\frac{\partial M}{\partial t} = -d\,K\frac{\partial \overline{C}}{\partial \xi}$$

K appears to be a rather complicated function of the flow velocity profile :

$$K = \frac{-1}{\lambda d}\int_0^d u \int_0^z \int_0^z u\,dz\,dz\,dz \tag{B.9}$$

If we write the conservation of mass for a small "slice" of the flow, we shall obtain the following equation for $\overline{C}$ in the moving coordinate system

$$\frac{\partial \overline{C}}{\partial \tau} = K\frac{\partial^2 \overline{C}}{\partial \xi^2} \tag{B.10}$$

which gives way, in a fixed coordinate system, to the well known one-dimensional advection – dispersion equation :

$$\frac{\partial \overline{C}}{\partial t} + \overline{U}\frac{\partial \overline{C}}{\partial x} = K\frac{\partial^2 \overline{C}}{\partial x^2} \tag{B.11}$$

The above demonstration is based on a few **strong** assumptions :

1. The analysis is restricted to two-dimensional flows with only one significant flow direction.
2. Besides, the flow is steady and uniform.
3. We consider a region sufficiently far away from the source so that the concentration distribution is a smoothly varying function. It is tantamount to assume we consider the pollutant cloud some (large) time after its injection.
4. Velocity and concentration deviations from their cross-sectional mean are supposed to be small.

After sufficiently long time, the solution of eq. B.11 is a normally distributed cloud moving at the mean speed $\overline{U}$ and whose standard deviation grows linearly with time. According to Fischer (Fischer *et al.*, 1979), this "sufficiently" long time is approximately equal to d^2/λ and eq. B.11 applies only for $t \geq 0.4d^2/\lambda$.

Some extensions of this analysis can apply reasonably to unidirectional turbulent flow or to two-dimensional laminar shear flow.

- As regards the unidirectional turbulent flow, the extension is relatively straightforward provided we adopt the Boussinesq eddy diffusivity concept to model the turbulent mass transport. In that case, the only significant difference is that in eq. B.1 the molecular diffusivity λ is replaced by a turbulent diffusivity Γ, which can be spatially variable. Thus, eq. B.6 and B.9 become respectively

$$u\frac{\partial \overline{C}}{\partial \xi} = \Gamma(z)\frac{\partial^2 c}{\partial z^2} \tag{B.12}$$

$$K = \frac{-1}{d}\int_0^d u \int_0^z \frac{1}{\Gamma(z)} \int_0^z u\, dz\, dz\, dz \tag{B.13}$$

- Similarly, in a skewed shear flow such as sketched in figure B.2, where the velocity deviations depend only on the coordinate z, the mean concentration $\overline{C}(x,y)$ obeys a two-dimensional diffusion equation

$$\frac{\partial \overline{C}}{\partial t} + \overline{U}\frac{\partial \overline{C}}{\partial x} + \overline{V}\frac{\partial \overline{C}}{\partial y} = K_{xx}\frac{\partial^2 \overline{C}}{\partial x^2} + K_{xy}\frac{\partial^2 \overline{C}}{\partial x \partial y} + K_{yx}\frac{\partial^2 \overline{C}}{\partial y \partial x} + K_{yy}\frac{\partial^2 \overline{C}}{\partial y^2}$$

  While K_{xx} and K_{yy} are the same as if there were only velocity profiles in those directions, K_{xy} and K_{yx} are related to the interaction between the U and V profiles. They mean that a velocity gradient in the $x-$ direction may induce transport in the $y-$direction and vice versa.

The assumptions which allow to introduce dispersion coefficients do not hold for general flows (unsteady, fully three-dimensional and non-uniform flows) and for any kind of pollutant source. For instance, in bends, the secondary motion can be so important that it is not safe to assume that the velocity deviations are small with respect to the depth-averaged velocities. Despite that, the modelling with the help of a diffusion operator of pollutant spreading due to velocity deviations from the mean flow has been widely extended to any kind of situation. The dispersion coefficients which play a part in this modelling result partly from theoretical or dimensional analysis and mostly from experiments.

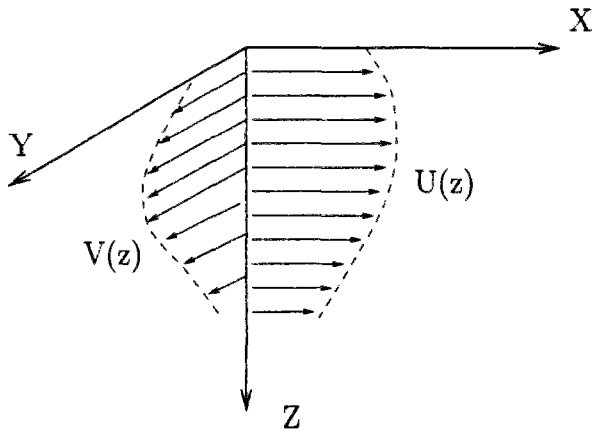


Figure B.2: Two-dimensional skewed flow

B.3 Usual formulae for dispersion coefficients in rivers

B.3.1 Elder analysis of wide open channels

Elder (Elder, 1959) considered the idealized case of a wide, uniform and straight open channel with a constant slope and depth. The channel under study is considered wide enough so that side-walls effects have no influence on the flow pattern. As the channel is uniform and the flow constant, we can expect the turbulence to be stationary and homogenous. As the walls are so far apart, the important length scale should be the depth.

Let ABCD be a small volume of water within this flow, with unit width and unit length (AD=BC=1). As the flow is uniform, the forces which act on this volume balance each other. Let us consider the projection of these forces parallel to the bed. We obtain :

$$\tau = \rho g I (d - z)$$

where τ denotes the shear stress, ρ the water density, g the gravitational acceleration, I the bed slope and d the depth. Thus, the shear stress varies linearly with depth, being null at the free surface, and reaching a maximum value $\tau_0 = \rho g d I$ at the bottom. This shear stress corresponds to the sum of the turbulent and viscous stresses, the latter being negligible except for the viscous sublayer near the bed.

At the time of Elders'analysis, the velocity profile in such kind of flow was supposed to be well approximated by the Von Karman logarithmic distribution :

$$u = \frac{U_*}{\kappa} \left(1 + \ln \frac{z}{d}\right) \tag{B.14}$$

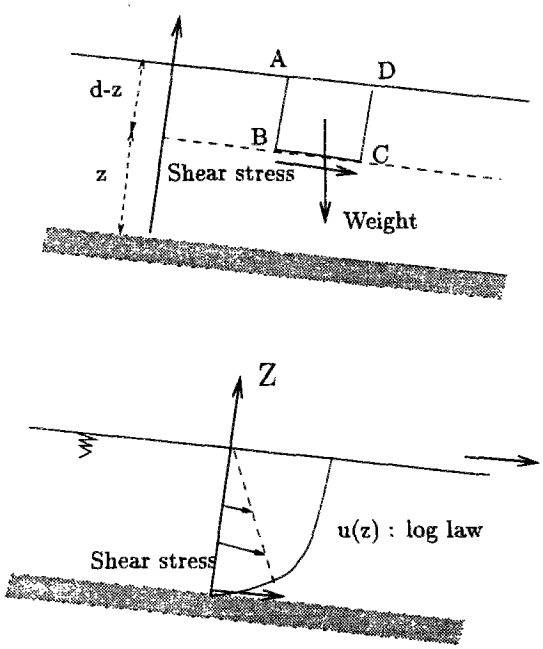


Figure B.3: Schematic flow in infinitely wide open channel

where U_* is the bottom friction (or shear) velocity defined by $\tau_0 = \rho U_*^2$ and κ the Von Karman constant.

According to the eddy-viscosity concept, the turbulent shear stress within the flow is related to the mean velocity gradient by

$$\tau = \rho \nu_t \frac{\partial U}{\partial z} \tag{B.15}$$

so that we can deduce the repartition of the eddy viscosity over the depth :

$$\nu_t = \kappa U_* d \frac{z}{d}\left(1 - \frac{z}{d}\right) \tag{B.16}$$

ν_t distribution is parabolic and, as $\kappa \simeq 0.4$, ν_t average value, denoted K_z, is

$$K_z = \overline{\nu_t} = 0.067 U_* d \tag{B.17}$$

Experiments in straight open channels have confirmed that the eddy viscosity defined by eq. B.15 has indeed a near-parabolic distribution, but this distribution is not universal and depends on the Reynolds number (Nezu & Rodi, 1986). In fact, the velocity distribution follows the universal log-law only close to the wall (for $z/d \leq 0.2$) and then deviates from it due to free surface effects. By analogy to boundary-layer and closed-channel analysis, this deviation has been modelled by a "wake function" :

$$\omega(z) = 2\Pi \frac{U_*}{\kappa} \sin^2\left(\frac{\pi}{2}\frac{z}{d}\right)$$

where the parameter Π is dependent on the Reynolds number $R_* = dU_*/\nu$ (ν denotes the kinematic molecular viscosity, $\nu = 10^{-6}$). (Nezu & Rodi, 1986) suggested that Π is null or negligible for $R_* \leq 500$, then increases with R_* and assumes a constant value of about 0.2 for $R_* \geq 2000$.

Elder pursued his analysis by applying formula B.13, so that he concluded that the longitudinal dispersion coefficient due to velocity vertical non-uniformities is (Elder, 1959) :

$$K_x = 5.93\, U_* d \tag{B.18}$$

This theoretical derivation was reasonably confirmed by Elder's own experiments.

B.3.2 Transverse dispersion

In wide channels there is no typical transverse velocity profile analogous to the log-law vertical velocity profile. Thus, it was not possible to establish theoretically a transverse analogy of eq. B.16. Consequently, the evaluation of the transverse dispersion coefficient relies only on experimental dye-tracing studies.

Straight channels The first experiments about transverse dispersion were performed by Elder (Elder, 1959). He found that the dispersion was not isotropic : the transverse dispersion coefficient K_y was approximately three times bigger than the vertical dispersion coefficient :

$$K_y \simeq 0.23\, U_* d \tag{B.19}$$

In wide channels the motion is less confined in the transverse than in the vertical direction so that this result is not really surprising.

The order of magnitude suggested by Elder for the adimensional coefficient $\alpha_y = K_y/U_* d$ was confirmed by numerous other laboratory flume studies (see review in (Fischer *et al.*, 1979; Yotsukura & Cobb, 1972; Lassale, 1992)). The experimental flumes were typically rectangular channels with rather smooth sides. Provided it is much wider than deep, this kind of channel is the closest possible approximation to an idealized infinitely wide open channel. Albeit these experiments concerned a rather large set of friction factor ($f = 8(U_*/U)^2$, ratio of shear velocity to mean velocity) and width to depth ratio ($5 \leq B/h \leq 60$), it turns out that these combinations led frequently to relatively moderate values of the Reynolds number ($R_* \leq 1000$). In summary, α_y was found to lie in the range 0.1 - 0.26 .

These findings were further completed by field experiments on lined irrigation canals (Yotsukura & Cobb, 1972). These canals were relatively small ($B \simeq 20 - 25$ m) but kept a large aspect ratio ($B/h \geq 30$). They were essentially straight. The α_y values varied from 0.22 to 0.3,

the larger value occurring for the canal with the larger amount of depth variation. According to a review in (Lau & Krishnappan, 1981), similar values were observed in straight natural reaches.

In more recent experiments (Webel & Schatzmann, 1984; Nokes & Wood, 1987) on straight flumes, researchers have tried to investigate systematically the influence of such variables as the width and the friction factor on α_y. (Webel & Schatzmann, 1984) state that α_y is insensitive to the aspect ratio B/h as soon as it is superior to 8 . (Nokes & Wood, 1987) made a selection among all the previously published experiments about lateral dispersion in rectangular flumes, including their own : by analysing the hydraulics conditions of the experiments, they discarded all flows likely to present strong three-dimensional, surface (too shallow flows) or wave effects (Froude number greater than 0.8), so as to sort out the cases which best approximate the flow conditions in an idealized infinitely wide open channel. They finally obtain a set of forty experiments. They conclude that most of the α_y values lie within 10 % of a value of 0.135 , except for some flows corresponding to very smooth beds (friction factor less than 0.055). For these flows, α_y is generally slightly more important but varies somehow randomly.

Irregular channels Natural channels can be much more complicated than experimental flumes or man-made canals. Their depth may vary irregularly, the channel is likely to curve and there may be large sidewalls irregularities. All this can contribute to the setting up of much more significant transverse motion and consequently can lead to a bigger rate of transverse spreading.

Indications about the influence of side-walls irregularities were given by experiments of Holley and Abraham (Fischer *et al.*, 1979; Holley, 1987) who worked on a rectangular channel with groins on its sides : α_y was found to vary from 0.35 to 0.5 . Similarly, Beltaos (Beltaos, 1980b) reported a value of $\alpha_y = 0.75$ for a straight river with irregular cross-sections. Holly & Nerat (Holly & Nerat, 1983) noted a considerable scatter in evaluating the longitudinal variation of α_y ($0.5 \leq \alpha_y \leq 2.5$) in a mountain river, rather sloping and strongly influenced by dikes and gravel bars.

The values of α_y observed on gently meandering channels (e.g. laboratory model of the Ijssel River (Fischer *et al.*, 1979; Holley, 1987), Missouri (Yotsukura *et al.*, 1970) or Potomac (Fischer *et al.*, 1979) rivers, Athabasca river (Beltaos, 1980b), Seine river (Théry *et al.*, 1993a)) all fall in the range $0.4 - 0.8$, the average being approximately 0.6 .

Higher values of α_y have been observed in sharply curving channels, both in laboratory flumes ((Fischer, 1969), $0.51 \leq \alpha_y \leq 2.4$; (Holley, 1987), $1.3 \leq \alpha_y \leq 2.4$) and in natural streams :

- Missouri river downstream of Cooper nuclear station (Fischer *et al.*, 1979), $\alpha_y \simeq 3.4$;

- Upper Missouri river downstream of Monticello nuclear power plant (Demetracopoulos & Stefan, 1983), $0.24 \leq \alpha_y \leq 4.65$ and $\overline{\alpha_y} \simeq 2$;
- Beaver river (Beltaos, 1980b), $\alpha_y \simeq 1$;
- Seine river upstream Paris (Thery *et al.*, 1993b), $2.5 \leq \alpha_y \leq 3.5$).

All the α_y values reported above have been computed either directly from the measured moments of dye-cloud distributions (Yotsukura *et al.*, 1970; Webel & Schatzmann, 1984; Nokes & Wood, 1987; Demetracopoulos & Stefan, 1983), either by trial and error, while calibrating two-dimensional stream-tube models (Fischer, 1969; Yotsukura *et al.*, 1970; Yotsukura & Cobb, 1972; Beltaos, 1980b; Holly & Nerat, 1983; Thery *et al.*, 1993a; Thery *et al.*, 1993b). These stream-tube models (for a summary of their principles, see (Cunge *et al.*, 1980b)) allow to describe some features of the differential convection over a cross section. Based either on consideration of the river bathymetry, either on field data, they approximate the transverse profile of the streamwise velocity. However, they simply neglect the existence of lateral velocities, which good description requires at least a depth-averaged two-dimensional model and, in some areas like bends, preferably a three-dimensional model.

If no truly two-dimensional analysis of the flow is provided, the lateral dispersion coefficient K_y deduced from the application of stream-tube models or directly from measurements is but a "catch-all" parameter describing the combined effects on lateral spreading of turbulent mixing, vertical non-uniformities, secondary flow. Thus, the scatter of α_y values observed in natural rivers is no wonder.

Empirical formulas for sharply meandering channels Fischer (Fischer, 1969) developed a theoretical expression for the additional transverse mixing $\Delta\alpha_y$ due to secondary motion in an infinitely wide channel, based on empirical formulas describing the distribution of the radial velocity component over the flow :

$$\Delta\alpha_y = 25 \left(\frac{Ud}{U_* R_c}\right)^2 \tag{B.20}$$

where R_c is the radius of curvature of the bend. He advised that this expression should be applied to practical situations only after an initial period for the bend. Indeed, if the bend is too short or, more exactly, if the flow time through the bend is too short, it is possible that there is no significant vertical mixing. Thus, the concentration differences induced by the helical motion have no consequence all over the water column. They amount only to net transverse displacements which can be easily reversed if the flow passes into an opposite, similar bend.

Yotsukura & Sayre and Sayre & Caro-Cordero (quoted in (Fischer *et al.*, 1979; Holley, 1987)) argued that, in natural bends, the width of the river should play some part, as it imposes

restrictions on the length scale of the secondary motion in its most significant direction, namely the transverse direction. They proposed to use instead of formula B.20 the following expression

$$\alpha_y = 0.4 \left(\frac{UB}{U_* R_c}\right)^2 \tag{B.21}$$

which provided better fits both with Fischer experiments and field tests.
Other researchers (Pavlovic & Rodi, 1985) reached a similar expression. (Pavlovic & Rodi, 1985) applied a depth-averaged flow and $k-\epsilon$ model to the interpretation of dye-tracing experiments in meandering flumes. Besides turbulent diffusion, a lateral dispersion term was considered in the scalar transport equation. The corresponding dispersion coefficient was evaluated from the outcome of a three-dimensional $k-\epsilon$ and mean-flow model (described in (Demuren, 1983; Demuren & Rodi, 1986)). Finally, the dimensionless diffusivity reads :

$$\alpha_y = \left(\frac{U}{U_*}\right)^{0.3} \left(\frac{B}{R_c}\right) f(x) \tag{B.22}$$

$f(x)$ being a sinusoidal function of the longitudinal location x within the bend. Indeed, (Pavlovic & Rodi, 1985), who performed a more detailed analysis than Yotsukura et al, noted a strong variation of α_y in the streamwise direction. The experimental flume was made of alternate bends separated by short straight reaches : α_y was observed to reach a maximum at the outlet of bends, then to decay steadily and reach a minimum at the next bend inlet.

In conclusion, although there is no definite theoretical basis for eq. B.21, no better correlation has been proposed for river bends. If formula B.21 is trusted, it is possible to suggest (Holley, 1987) some kind of classification scheme for meandering rivers. Eq. B.21 could be used to evaluate if the secondary circulation in a given bend is strong enough to induce significant additional transverse spreading. Holley (Holley, 1987) suggested that the additional mixing be judged significant if α_y was 10 % bigger than a typical value for straight channels, which he considered to be 0.4 . This led to the conclusion that secondary motion was significant for bends which satisfy

$$\frac{UB}{U_* R_c} \geq 0.3 \tag{B.23}$$

As mentionned above, the bend should also be long enough for vertical mixing to develop. Holley (Holley, 1987) suggested to compare the typical flow time in the bend (L_b/U where L_b denotes the bend length) and a typical time for achieving vertical mixing d/Γ_z where Γ_z is the vertical diffusivity. Albeit the vertical profiles of velocity within a bend are probably not adequately described by a log-law profile, precisely because of the secondary motion, Holley used the Elder formula B.17 to assess Γ_z. By doing so, he obtained the following supplementary criteria for eq. B.21 to be relevant :

$$\frac{U_* L_b}{Ud} \geq 15 \tag{B.24}$$

B.3.3 Longitudinal dispersion

Longitudinal dispersion in depth-averaged 2D models We must distinguish between two kinds of longitudinal dispersion coefficients, those which enter depth-averaged models and account only for vertical non-uniformities and those which enter one-dimensional models and represent the effect of deviations with respect to the cross-sectional means of velocity and concentration.

As regards the first kind, no other formula than the Elder's one (eq. B.18) has been suggested. As reported in (Rodi, 1980), when the turbulence is mainly bed generated, as in channel flow, the depth-mean diffusivities are indeed reasonably well correlated with the friction velocity U_* and the depth d, so that for far-field calculation of discharges, empirical formula such as $K \propto U_* d$ are used with reasonable success. However, the proportionality constant is river or reach dependent, and can also be variable according to the river discharge rate. Formulas as eq. B.18 or B.19 can be applied to obtain an order of magnitude of the dispersion coefficients but these should be calibrated more accurately with the help of specific field experiments.

As regards dispersion in ocean and large lakes, when the cloud of pollutant keeps significantly smaller than the large scale circulations within the water body and is not affected by boundaries effects, it has been suggested (Fischer *et al.*, 1979; Rodi, 1980) to use empirical relationship such as $K = \beta L^{4/3}$ where β is a dissipation parameter and L a typical length-scale of the circulations.

Fischer's analysis for 1D models When applied to natural streams in one-dimensional models, formula B.18 was found to typically underestimate the dispersion. A rigorous explanation of this fact was given by Fischer (Fischer, 1967; Fischer, 1968), who showed that longitudinal dispersion in natural streams is primarily governed by *transverse* variations of the longitudinal velocity rather than by vertical variations, as considered in Elder's analysis. He proposed a formula analogous to eq. B.13 in order to compute the longitudinal dispersion coefficient :

$$K_x = \frac{-1}{A} \int_0^B u' d \int_0^y \frac{1}{\Gamma_y d} \int_0^y u' d \, dy \, dy \, dy \tag{B.25}$$

where y denotes the transverse coordinate, A the wet section, Γ_y the transverse dispersion coefficient and d the local depth. Γ_y and d depends on y. u' represents the following quantity : let $\overline{U}$ be the mean value of the velocity over the cross-section, and $u^z(y)$ the depth-averaged value of the velocity at location y; then, $u'(y) = u^z(y) - \overline{U}$.
Fischer (Fischer *et al.*, 1979) advises that formula B.25 provides only an estimate of the longitudinal dispersion coefficient in real streams, as its derivation is based on the assumption of uniform flow in a constant cross-section. Besides, a monodimensional representation of the pollutant transport process makes sense only when the pollutant has achieved a reasonable homogeneity across the wet section. Fischer (Fischer *et al.*, 1979) suggested that such was the case

at a distance x of the (side) discharge which satisfies :

$$\frac{x\,\overline{\Gamma_y}}{\overline{U}\,B^2} \geq 0.4 \tag{B.26}$$

Fischer simplified eq. B.25 by introducing the adimensional variables : $\epsilon_t = \Gamma_y/\overline{\Gamma_y}$, $d' = d/\overline{d}$, $u'' = u'/\sqrt{\overline{u'^2}}$ where $\overline{d}$ and $\overline{\Gamma_y}$ denote respectively the average depth and transverse diffusivity over the cross-section, and $\overline{u'^2}$ is the average intensity of the velocity deviations u' (i.e. the second-order moment of u'). By doing so, he showed (Fischer *et al.*, 1979) that eq. B.25 reduces to :

$$K_x = I_u \frac{\overline{u'^2}B^2}{\overline{\Gamma_y}} \tag{B.27}$$

where I_u is an adimensional integral,

$$I_u = -\int_0^1 u''d' \int_0^y \frac{1}{\epsilon_t\, d'} \int_0^y u''d'\, dy\, dy\, dy$$

The exact computation of I_u requires to know the detail of the flow pattern within the cross-section. However, based on laboratory experiments, Fischer suggests that the range of values for I_u in natural streams is narrow enough as is the range of the ratio $\overline{u'^2}/\overline{U}^2$. By assuming furthermore that $\overline{\Gamma_y}/U_*\overline{d}$ takes a mean value of 0.6 (a typical value in gently meandering channels, see above), Fischer (Fischer *et al.*, 1979) finally proposed :

$$K_x = 0.011 \frac{\overline{U}^2\,B^2}{U_*\,\overline{d}} \tag{B.28}$$

Fischer (Fischer *et al.*, 1979) observs that, albeit all the approximations it is based on, eq. B.28 has been found to agree with the observations within a factor of four or so, which is not bad when one bears in mind that Elder's formula B.18 was found to disagree with a factor of 100 with respect to some field experiments. Such remark has been confirmed by other researchers (Marivoet & Van Craenenbroeck, 1986; Mouchel, 1989; Rigaudière, 1992) .

A review of alternative formula for predicting K_x is given in (Lassale, 1992). At the time being, no proposed formula has been found to apply in any case without a good amount of careful calibration exercise !

The shortcomings of the mono-dimensional dispersion theory Should the equation B.11 apply, the pollutant distribution should take after some time a regular, gaussian-type shape. However, since long, field experiments have been pointing to the fact that, in real cases, the distributions generally exhibit a lasting skewness and display persistent tails of concentration (Day, 1975; Day & Wood, 1976; Beltaos, 1980a; Nordin & Troutman, 1980).

Several explanations have been proposed. Some researchers state that, due to the irregularity of natural rivers, the distance after which the one-dimensional analysis can apply is much longer

than suggested by eq. B.26 (Holley, 1987) and propose to adapt the one-dimensional model in order to account more correctly for the initial stages of the pollutant cloud development (Beltaos, 1980a; Liu & Cheng, 1980; Holley, 1987). Some others suggest that the discrepancy between theoretical curves and observed ones is mainly related to the presence of storage (dead zones) located on the river bed (behind roughness singularities) or on its banks (Valentine & Wood, 1977; Valentine & Wood, 1979; Sabol & Nordin, 1978; Nordin & Troutman, 1980). Thus, they advocate (Valentine & Wood, 1979; Sabol & Nordin, 1978; Nordin & Troutman, 1980; Holley, 1987) to include in eq. B.11 a term which accounts for the exchange of matter between the main flow and the storage zones, these stagnant zones being considered as perfectly mixed and being characterized by a long residence time. Experiments for determining the exchange rate between main flow and dead zones have even been conducted (Valentine & Wood, 1977; Valentine & Wood, 1979).

B.3.4 Conclusions

A sound justification of the representation of dispersion effects with the help of a diffusion operator and related diffusivities (the dispersion coefficients) was in fact restricted to unidimensional situations, at asymptotically large times of the cloud development. However, this representation has been extended to varied, multi-dimensional situations, and unsteady flows.

Numerous empirical formulas have been proposed in order to derive the dispersion coefficients. As could be expected, due to the difficulty of conducting field experiments and to the lack of strong physical basis for developing these formulas, none has been found to perform well under any circumstances and none can be considered as universal. Thus, the determination of appropriate dispersion coefficients within a given water body still requires the acquisition of specific data and model tuning.

However, it can be noted that, for application in depth-averaged calculations in river, the transverse and longitudinal dispersion coefficients have been found to be reasonably well correlated with the product $U_* d$, U_* being the shear velocity at the river bed and d the water depth. The proportionality coefficient is unfortunately site-specific. As regards the transverse dispersion, it appears that the proportionality constant is significantly affected by river irregularities and bends, which are known to increase the secondary motion within the cross section.

For the longitudinal dispersion coefficient used in monodimensional models, the scatter of values mentionned in the literature and the number of proposed empirical formulas are still

wider. Fischer demonstrated that this coefficient is essentially dependent on the tranverse profile of streamwise velocities and is not governed primarily by the product U_*d. His formula B.28, which takes into account the river width and average velocity, has been frequently used with reasonable success as a basis for field calibration. Besides, the 1D advection-dispersion equation has been shown to frequently ignore some typical features of the pollutant distributions, such as persistent skewness and long tails of concentration. This has been often ascribed to the interaction with the mean flow of storage zones at the river bed or banks, which cannot be described properly in a solely monodimensional representation.

Appendix C

Consistency, stability and convergence of FDM

C.1 Definitions

- **Consistency**
  A finite difference numerical representation is said to be consistent if the truncation error vanishes as the finite-difference mesh size (both in space and time) approaches zero. It means that the limit of the difference representation is indeed the continuous differential equation. Consistency implies more than merely a good limit behaviour of each finite difference approximation to individual derivatives.

- **Stability**
  Stability is a concept applicable to "marching" problems, i.e. involving time t. Marching problems can be related, either to the simulation of really transient flows, either to the asymptotic calculation of steady flows with a scheme devised for unsteady situations.

  As explained in Chapter 3, the application of a finite difference representation results in replacing the PDE by a system of algebraic equations. Stability is the tendency for any spontaneous perturbations (such as round-off errors, noise in initial or boundary conditions, ...) in the solution of this system of equations to decay.

- **Convergence**
  A convergent finite-difference scheme is defined mathematically as one in which all values of the finite difference solution approach the parent continuum differential equation solution as the mesh size approaches zero.

  Lax's theorem states that : given a properly posed initial value problem governed by a

system of *linear* PDE, a finite difference representation of this system is convergent if and only if it is both consistent and stable.

In practical situations, the Lax theorem is frequently assumed to hold even for *non - linear* PDE.

The study of stability and convergence of finite difference representations is of course essential. Hereafter we introduce the usual methods available to achieve this kind of analysis.

C.2 Consistency study

Consistency is studied by considering the truncation error (TE) which is usually a polynomial expression of the space and time steps. Notation TE$= O(\Delta x^n, \Delta t^m)$ means that the lower-order terms of the truncation error are n^{th} and m^{th} order ones respectively for the space and time step. Obviously, as soon as $n \geq 1$ and $m \geq 1$, the truncation error vanishes as the mesh sizes decrease and the related scheme proves to be consistent.

Let us consider an example, the one-dimensional heat (or mass) diffusion equation :

$$\frac{\partial F}{\partial t} = D \frac{\partial^2 F}{\partial x^2} \tag{C.1}$$

where we assume the diffusion coefficient D to be uniform and constant. The equation is solved on an uniformly spaced computational grid. To derive the FDE, we adopt a forward differencing of the time derivative (cf eq. C.2) and an explicit centred differencing of the second-order space derivative (cf eq. C.3) :

$$\frac{\partial f}{\partial t}(x_i, t_n) \simeq \frac{f_i^{n+1} - f_i^n}{\Delta t} \tag{C.2}$$

$$\frac{\partial^2 f}{\partial x^2}(x_i, t_n) \simeq \frac{f_{i+1}^n - 2f_i^n + f_{i-1}^n}{\Delta x^2} \tag{C.3}$$

The FDE reads :

$$\frac{f_i^{n+1} - f_i^n}{\Delta t} - D \frac{f_{i+1}^n - 2f_i^n + f_{i-1}^n}{\Delta x^2} = 0 \tag{C.4}$$

By developping the Taylor series corresponding to each derivative in the PDE, we check that the FDE is second-order accurate in space and first-order accurate in time. Indeed :

$$\text{T.E.} \simeq \frac{\Delta t}{2} \frac{\partial^2 f}{\partial t^2} - D \frac{\Delta x^2}{12} \frac{\partial^4 f}{\partial x^4} = O\left(\Delta x^2, \Delta t\right)$$

Considering that if $n > m$, δ^n tends to zero quicker than δ^m as δ decreases, it can be tempting to prefer high-order finite difference approximations of the derivatives as they result in

higher-order truncation error. However, the formal superior accuracy of high-order schemes can be misleading. Indeed, for practical reasons, the numerical grid employed in most calculations has to remain somewhat coarse. Similarly, large time steps are often desirable.

Deriving higher-order spatial approximations usually implies to involve more neighbours of a node. If the space step is too big, it may just become meaningless, from a physical viewpoint, to introduce an explicit influence of these farther off neighbours by the way of our numerical algorithm.

Besides, higher-order methods generally correspond to more complex and costly schemes. For instance, in order to increase the temporal order of a scheme, it becomes necessary to use the value of the dependent variables at several time levels, which must then be stored. Similarly, as more neighbours of each node are used, the algebraic linear systems produced by the finite difference or finite element representation of the PDE are less easy to handle (e.g. pentadiagonal matrices instead of tridiagonal ones).

Finally, it can become more troublesome to express boundary conditions in a manner which keeps consistent with the method used at the inner points, and does not lower too badly the overall order and accuracy of the scheme.

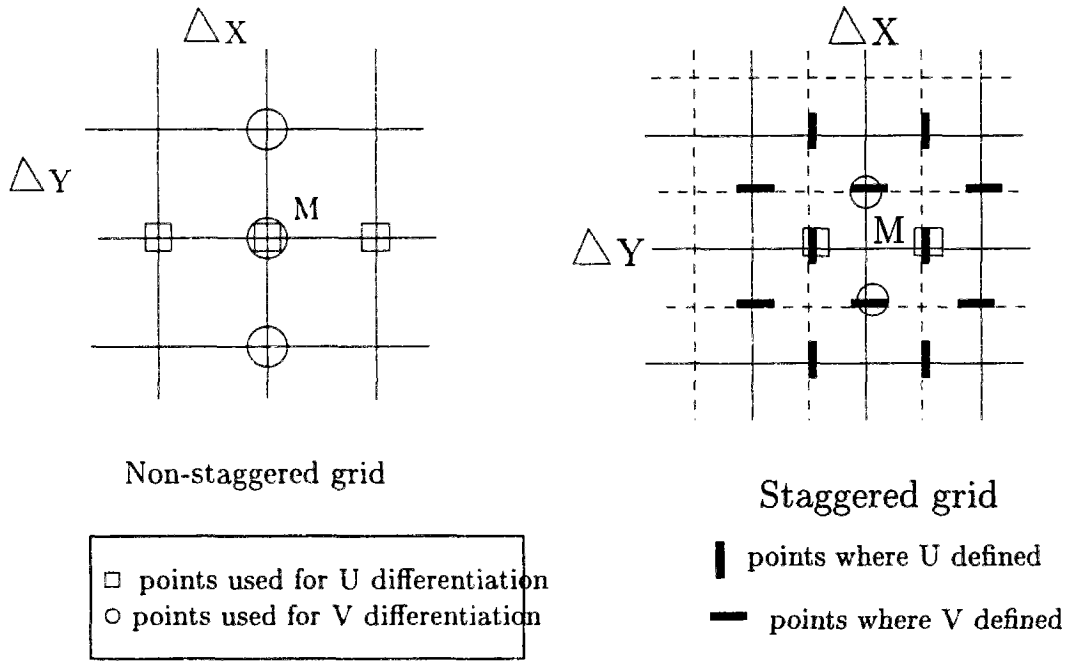


Figure C.1: Two-dimensional plain and staggered grids

When the problem under study deals with several dependent variables, as in the flow equa-

tions, one way to improve the approximation without significantly increasing the number of nodes involved is to use staggered grids (i.e. different variables are defined on different grid points). For instance, let us consider the two grids as sketched in figure C.1 and the problem of deriving a finite difference representation of the depth-averaged flow continuity equation :

$$\frac{\partial h}{\partial t} + \frac{\partial Uh}{\partial x} + \frac{\partial Vh}{\partial y} = \frac{\partial h}{\partial t} + U\frac{\partial h}{\partial x} + V\frac{\partial h}{\partial y} + h\left(\frac{\partial U}{\partial x} + \frac{\partial V}{\partial y}\right) = 0$$

Let us consider second-order, centred differences approximations of the first-order spatial derivatives of the velocities at node M, e.g

$$\frac{\partial U}{\partial x}(x_i, y_j, t_n) \simeq \frac{U^n_{i+1,j} - U^n_{i-1,j}}{2\Delta x} \tag{C.5}$$

The reader would easily check that, on the first grid, the dominant term of the error made in evaluating $\frac{\partial U}{\partial x}$ (resp. $\frac{\partial V}{\partial y}$) is $\frac{1}{6}\Delta x^2 \frac{\partial^3 U}{\partial x^3}$(M) (resp. $\frac{1}{6}\Delta y^2 \frac{\partial^3 V}{\partial y^3}$(M)). On the second, staggered, grid, errors are four times smaller.

C.3 Stability study

Different useful methods are available in order to study the stability of a FDE (Richtmyer & Morton, 1967; Roache, 1985). Hereafter, we recall only the two most common techniques.

C.3.1 Von Neumann method

In this method, we consider a Fourier series expansion of the solution of the finite difference representation (FDE) of a PDE. Then, the decay or amplification of each mode is considered separately to determine stability or instability.

Von Neumann analysis is best illustrated when taking a concrete example. Let us consider the one-dimensional transport equation

$$\frac{\partial F}{\partial t} + U\frac{\partial F}{\partial x} = 0 \tag{C.6}$$

where we assume the flow velocity U to be constant and uniform. The equation is solved on an uniformly spaced computational grid. Let f be the solution of a finite-difference representation of C.6. We assume it can be written as :

$$f^n_i = f(x_0 + i\Delta x, t_0 + n\Delta t) = \sum_k \mathcal{A}_k\,(t_0 + n\Delta t)\,\exp j\omega_k.i\Delta x \tag{C.7}$$

j is such that $j^2 = -1$. The ω_k denote the wave numbers (related wavelength $\wedge_k = 2\pi/\omega_k$). To shorten the above expressions, we introduce the phase angle $\theta_k = \omega_k \Delta x$ and denote $\mathcal{A}^n_k = \mathcal{A}_k\,(t_0 + n\Delta t)$.

Eq. C.6 and consequently its finite difference representations are linear. Thus, the principle of superposition applies : f being solution of the FDE implies that each elementary Fourier component $\varphi_i^n = \mathcal{A}_k^n \exp j\theta_k i$ also satisfies the FDE. Developping the FDE at each node, and eliminating the common factor $\exp j\theta_k i$, leads for each k to an equation of the kind :

$$\mathcal{A}_k^{n+1} = \mathcal{G}_{\text{num}} \mathcal{A}_k^n \tag{C.8}$$

$\mathcal{G}_{\text{num}}$ is termed the **amplification factor** of the FDE. It is usually a function of the space and time step, of the flow (U) and of the phase angle. If the solutions (and consequently the errors) are to remain bounded, we must have

$$|\mathcal{G}_{\text{num}}| \leq 1 \quad \text{for all } \theta \tag{C.9}$$

Inequation C.9 expresses the stability criterion of the FDE. It generally results on inequalities that must be fulfilled by the space and time steps. For instance let us assume we adopt a forward differencing of the time derivative (eq. C.2) and an explicit upwind differencing of the space derivative, that is to say :

$$\frac{\partial f}{\partial x}(x_i, t_n) \simeq \frac{f_i^n - f_{i-1}^n}{\Delta x} \quad \text{if } U \geq 0$$

Consequently, the FDE reads :

$$\frac{f_i^{n+1} - f_i^n}{\Delta t} + U \frac{f_i^n - f_{i-1}^n}{\Delta x} = 0$$

This yields :

$$\mathcal{G}_{\text{num}} = 1 - C_r (1 - \exp -j\theta) = [\, 1 - C_r(1 - \cos\theta) \,] - jC_r \sin\theta \tag{C.10}$$

where C_r is the **Courant number** defined as $C_r = U\Delta t/\Delta x$.

$$|\mathcal{G}_{\text{num}}|^2 = 1 - 4C_r (1 - C_r) \sin^2 \frac{\theta}{2} \tag{C.11}$$

Stability for each possible value of the wavelength (so of θ) is achieved if and only if $C_r \leq 1$. This criterion defines, else a maximum allowable time step if the space step is fixed, else a minimum space step if the time step has been chosen. Should U be unsteady, time and space steps would have to be adjusted so that they satisfy the criterion for the most critical value of U throughout the computation.

For more complicated PDE and related FDE, the analytical expression of G_{num} can be much more complex, so that its boundedness is studied numerically.

The PDE may lend itself to some decomposition into different operators, which correspond to different physical processes (e.g. advection and diffusion) or apply to different coordinate directions, and can be written for instance :

$$\frac{\partial F}{\partial t} + \mathcal{L}_1 F + \mathcal{L}_2 F = 0 \tag{C.12}$$

It can be tempting to study individually the stability of each operator approximation. Let G^i_{num}, $i = 1, 2$ be the amplification factor corresponding to the discretisation and differentiation of each individual equation :

$$\frac{\partial F}{\partial t} + \mathcal{L}_i F = 0 \qquad i = 1, 2 \tag{C.13}$$

Let us write each G^i_{num} as $G^i_{\text{num}} = 1 + \Delta t\, K^i$. If, when dealing with the global equation C.12, the $\mathcal{L}_i$ operators are treated simultaneously, the overall amplification factor appears to be :

$$G_{\text{num}} = 1 + \Delta t \left(K^1 + K^2\right) \tag{C.14}$$

Consequently, the stability of each finite difference representation (i.e. $|G^i_{\text{num}}| \leq 1$ for $i = 1, 2$) does not warrant the stability of the overall FDE. In fact, the stability criteria for the global FDE may be significantly more stringent than the criteria for each operator (e.g (Leonard, 1979) for the case of explicit schemes applied to the scalar advection-diffusion equation).

On the other hand, if the $\mathcal{L}_i$ operators are treated in *successive* steps, assuming that

$$\left(\frac{\partial}{\partial t} + \mathcal{L}_1\right) \circ \left(\frac{\partial}{\partial t} + \mathcal{L}_2\right) F \simeq \left(\frac{\partial}{\partial t} + \mathcal{L}_1 + \mathcal{L}_2\right) F \tag{C.15}$$

we have $G_{\text{num}} = G^1_{\text{num}}.G^2_{\text{num}}$ and overall stability is ensured by the stability of each step. In conclusion, caution must be exercised when combining the stability study of different operators.

Von Neumann method can also be applied to a *system* of linear PDE. In such cases, we consider $\vec{F}$, the vector of the dependent variables, and its Fourier decomposition. The analysis then leads to an amplification matrix $\underline{G}_{\text{num}}$ in place of the amplification factor. The dimension of this square matrix is the number of variables. It can be demonstrated that *the Fourier modes remain bounded if and only if the matrix norm* $\|\underline{G}_{\text{num}}\|$ *is inferior to 1. This is tantamount to the fact that all eigenvalues of* $\underline{G}_{\text{num}}$ *have a modulus less than or equal to one. Such condition is also expressed as : the spectral radius of* $\underline{G}_{\text{num}}$ *is less than 1.*

C.3.2 Matrix method

In this technique, we express the set of equations governing the error propagation in matrix form and examine the eigenvalues of the associated matrix.

Let f^* be the "ideal", error-free, solution of a FDE, f the numerical solution of the system of algebraic equations associated to the FDE, and $\xi = f^* - f$ the corresponding error. It can be shown that, if the algebraic equations produced by discretisation are linear, the error terms ξ satisfy the same set of equations than f^*.

Let us look back to the example of the one-dimensional diffusion equation C.1. Relation C.4 yields :

$$f_i^{n+1} = df_{i-1}^n + (1-2d)\, f_i^n + df_{i+1}^n \tag{C.16}$$

where $d = D\Delta t/\Delta x^2$. Eq. C.16 governs the exact solution f^*, as well as the numerical solution f and the error ξ .

Let us assume first that the boundary conditions are of the Dirichlet type, that is to say, that values of the function are imposed at each boundary (corresponding respectively to nodes 1 and $\mathcal{N}$). Consequently,

$$\forall n, \quad \xi_1^n = \xi_{\mathcal{N}}^n = 0$$

Thus, the ξ_j obey the following set of equations :

$$\begin{aligned}
\xi_2^{n+1} &= (1-2d)\,\xi_2^n + d\xi_3^n \\
\xi_i^{n+1} &= d\xi_{i-1}^n + (1-2d)\,\xi_i^n + d\xi_{i+1}^n \quad \forall i,\ 3 \le i \le \mathcal{N}-2 \\
\xi_{\mathcal{N}-1}^{n+1} &= d\xi_{\mathcal{N}-2}^n + (1-2d)\,\xi_{\mathcal{N}-1}^n
\end{aligned}$$

This can be written in matrix form as :

$$\underline{\xi}^{n+1} = \underline{G}\,\underline{\xi}^n \quad \text{for } n = 0,\ 1,\ 2 \ldots \tag{C.17}$$

where $\underline{G}$ is a square matrix and $\underline{\xi}$ a vector, both of dimension $\mathcal{N}-2$:

$$\underline{G} = \begin{bmatrix}
(1-2d) & d & & & & & \\
d & (1-2d) & d & & & & \\
 & d & (1-2d) & d & & & \\
 & & \cdot & \cdot & \cdot & & \\
 & & & \cdot & \cdot & \cdot & \\
 & & & & d & 1-2d &
\end{bmatrix}, \quad \underline{\xi}^n = \begin{bmatrix} \xi_2^n \\ \cdot \\ \cdot \\ \xi_i^n \\ \cdot \\ \xi_{\mathcal{N}-1}^n \end{bmatrix}$$

Introducing norms over the vectorial space $\Re^{\mathcal{N}-2}$, it can be demonstrated that *the errors remain bounded if and only if the spectral radius (i.e.* $\max_j |\lambda_j|$*) of* $\underline{G}$ *is less than unity.* (nb : a similar criterion was derived in case of multivariable Von Neumann analysis). Subsequent calculations would indicate for this scheme that stability is achieved for $d \le 1/2$.

Changes in boundary conditions can readily be investigated through the matrix method. Let us consider that we replace the Dirichlet boundary conditions by Neumann ones, i.e. of the type

$$\frac{\partial f}{\partial x} = g \quad \text{for } x = x_1,\ x_{\mathcal{N}}$$

This boundary condition could be discretised in several ways. One method is to introduce a fictitious point f_0 such that :

$$\frac{\partial f}{\partial x}_{|1} = \frac{f_2 - f_0}{2\Delta x} = g_1 \quad \text{and} \quad f_0 = f_2 - 2\Delta x\, g_1 \tag{C.18}$$

Thus, first order derivatives estimates at the boundary are second-order accurate. Combining C.18 with the "normal" algebraic equation $f_1^{n+1} = df_0^n + (1-2d)\, f_1^n + df_2^n$, we obtain a modified equation at node 1 :

$$f_1^{n+1} = (1-2d)\, f_1^n + 2df_2^n - 2d\Delta x\, g_1$$

Both the "exact" and computed solution of the FDE obey this equation, so that, assuming there is no noise on the boundary condition g, the error ξ eventually satisfies :

$$\xi_1^{n+1} = (1-2d)\, \xi_1^n + 2d\xi_2^n$$

An analogous relation may be derived for node $\mathcal{N}$, introducing a fictitious node $\mathcal{N}+1$. ξ verifies once again a matrix equation of the form

$$\underline{\xi'}^{n+1} = \underline{G'}\, \underline{\xi'}^n$$

where $\underline{G'}$ and $\underline{\xi'}$ denote this time respectively a square matrix and a vector of dimension $\mathcal{N}$:

$$\underline{G'} = \begin{bmatrix} (1-2d) & 2d & & & & & \\ d & (1-2d) & d & & & & \\ & d & (1-2d) & d & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & 2d & 1-2d \end{bmatrix}, \quad \underline{\xi'}^n = \begin{bmatrix} \xi_1^n \\ \cdot \\ \cdot \\ \xi_i^n \\ \cdot \\ \xi_{\mathcal{N}}^n \end{bmatrix}$$

The matrix method can also be applied to implicit schemes. However, these schemes result in matrix relations such as

$$\underline{A}\, \underline{\xi}^{n+1} = \underline{B}\, \underline{\xi}^n$$

so that the amplification matrix $\underline{G}$ is more complicated ($\underline{G} = \underline{A}^{-1}\, \underline{B}$). In this case, as in case of variable coefficients (e.g. space-dependent diffusion D), the eigenvalues are not straightforward to compute. However, some algorithms (Fletcher, 1991) allow to solve this problem. Finally, in case of time-dependent coefficients, the stability criterion could evolve in time.

C.3.3 Extension of Von Neumann method

Strictly speaking, Von Neumann analysis applies only to linear initial values problems with constant coefficients. Besides, while the matrix method includes naturally the evaluation of the influence of boundary conditions treatment on stability, the Von Neumann method deals only with interior points.

Practical problems typically involve variable coefficients, nonlinearities and complicated types of boundaries conditions. In which case, the method can only be applied locally and with the non-linearities temporarily **"frozen"**. As explained in section 3.2.3, non-linear systems such as

$$\frac{\partial \vec{F}}{\partial t} + \frac{\partial \vec{H}}{\partial \vec{x}} = 0 \quad \text{with } \vec{H} \text{ function of } \vec{F} \tag{C.19}$$

are locally (i.e. at each computational node and for each time step) approximated by linear systems such as described by

$$\frac{\partial \vec{F}}{\partial t} + J_{\mathrm{M}} \frac{\partial \vec{F}}{\partial \vec{x}} = 0 \tag{C.20}$$

where J_{M} is the Jacobian matrix :

$$J_{\mathrm{M}} = \frac{\partial \vec{H}}{\partial \vec{F}}_{|\mathrm{M}}$$

Then, the Von Neumann analysis can proceed on eq. C.20. However, for this situation, it provides *necessary*, but not always *sufficient* requirements for stability of the finite difference representation of the original system of equations eq. C.19.

Without considering splitting problems (cf section 3.6), it can be said that the boundary conditions which raise problems are those which involve the solution space derivatives (Neumann or mixed conditions). The best we can do with the Von Neumann method is to apply it separately to the algorithms used at the boundaries : it can then give heuristic information concerning the propagation of boundaries errors.

Time, space or process splitting (cf section 3.6) has the effect that it replaces the partial differential operator under study by a product of simpler operators. The stability of each operator is generally studied separately, as their combined behaviour is most frequently too difficult to assess. Then, it becomes even more delicate to assess the reach of the stability requirements : on one hand, stability of each operator is a sufficient, but not necessary, condition for overall stability; on the other hand, if each individual stability requirement is a condition necessary but not sufficient to ensure the stability of the related operator, where does it lead us ? In practical situations, modelers safely allow for a margin of security on any stability conditions. Unfortunately, this does not work always !

C.4 Analysis of dissipation and dispersion errors

Going back over linear initial value problems with constant coefficients, we shall see how, apart from providing us with stability criteria, the Von Neumann analysis can help us to gain more insight into the behaviour of the finite difference representation.

Let us consider again eq. C.6 : it is straightforward to check that exact solutions satisfy $F(x, t + \Delta t) = F(x - U\Delta t, t)$. Let us assume that the elementary Fourier component $\varphi = \mathcal{A}_k(t) \exp j\omega_k x$ is an exact solution of C.6. In that case, we should have :

$$\varphi_i^{n+1} = \mathcal{A}^{n+1} e^{j\omega_k i\Delta x} = \mathcal{A}^{n+1} e^{j\theta_k i} = \mathcal{A}^n e^{j\omega_k(i\Delta x - U\Delta t)} = \mathcal{A}^n e^{j\theta_k(i - C_r)}$$

Consequently, the **exact** amplification factor (i.e. amplification factor for the exact PDE) is :

$$\mathcal{G}_{\text{ex}} = \exp -jC_r\theta = \cos(C_r\theta) - j\sin(C_r\theta) \tag{C.21}$$

By comparing $\mathcal{G}_{\text{num}}$ to $\mathcal{G}_{\text{ex}}$, we can assess how well the FDE respects the features of the solutions.

For instance, eq. C.21 tells us that for the exact PDE there is no damping neither growth of the solution modulus. On the other hand, eq. C.11 indicates that, apart from the special case when $C_r = 1$, the FDE tends to damp most components (i.e. $|\mathcal{G}_{\text{num}}| < 1$), except those which satisfy $\omega = 2n\pi/\Delta x$ (n integer), namely $\wedge = \Delta x/n$. Unfortunately, given to the space discretisation, the Fourier series components of the numerical solution have wavelength which are *proportional* to Δx ($\wedge = n\Delta x$) and are no fraction of it ! The damping is maximum for components whose wavelength is $\wedge = 2\Delta x$. For a given Fourier component, attenuation raises as C_r approaches 0.5 .
The introduction of damping by the FDE is referred to as a numerical **dissipation phenomenon**.

We can also compare the exact and numerical phase shift $[\Delta\theta]$ during one time step.

$$[\Delta\theta]_{\text{ex}} = -C_r\theta \tag{C.22}$$

$$\sin[\Delta\theta]_{\text{num}} = -\frac{C_r \sin\theta}{\sqrt{1 - 4C_r(1 - C_r)\sin^2\frac{\theta}{2}}} \tag{C.23}$$

For a Fourier component such that $\theta \ll 1$ (it means $\wedge = n\Delta x$ with n big), C.23 can be developed so that, assuming that $\sin[\Delta\theta] \simeq \theta$, we obtain :

$$[\Delta\theta]_{\text{num}} - [\Delta\theta]_{\text{ex}} = C_r \frac{\theta^3}{6}[\,1 - 3C_r(1 - C_r)\,] \tag{C.24}$$

Each Fourier component should propagate with the same celerity, equal to the flow velocity. However, from eq. C.24, we can deduce that the Fourier component corresponding to θ propagates with a celerity U_{num} so that

$$\frac{U_{\text{num}}}{U} = 1 - \frac{\theta^2}{6}[\,1 - 3C_r(1 - C_r)\,] \quad (\leq 1)$$

All Fourier components travel at different speeds. Thus, different components will spread apart, or disperse, as the numerical solution proceeds : the phenomenon is frequently referred to as **dispersion error.** In the above example, all Fourier components travel slower than they should : the scheme is said to have a *lagging phase error.* If the scheme was exhibiting the opposite behaviour, we would say it has a *leading phase error.*

It is not sufficient to rank a scheme only according to its dissipative properties or, at the contrary, solely with respect to its dispersive ones. In fact, both kind of errors combine either to worsen, either to improve the scheme behavior. For instance, small damping may be useful if it helps smoothing out quickly the wavelengths which exhibit the largest celerity error.

In summary, the Von Neumann analysis provides useful informations concerning the errors generated by a FDE. However, it is by no means a base sufficient to conclude upon the overall accuracy of a scheme. Indeed, the Von Neumann analysis deals with elementary Fourier components. Yet, most real functions, even the simplest ones, have rather complex Fourier series expansions. Consequently, individual components errors mix in a way which is difficult to forecast : some errors can add, the others can counteract; dissipation and dispersion errors combine, the latter being eventually dominant in case of transient problems ... Besides, as mentionned earlier, a complete assessment of the numerical and exact amplification factors is possible only in the simplest cases.

Appendix D

Solving systems produced by FDM, FVM and FEM

When applied to flow equations, FDM, FEM and FVM typically produce system of equations that can be written, for each time step

$$\underline{A}(\vec{f})\,\vec{f} = B \tag{D.1}$$

where $\vec{f}$ is the vector of $\mathcal{N}$ unknown nodal values. The regular matrix $\underline{A}$ contains the algebraic coefficients arising from discretisation and, in general (non-linear PDE), may depend on the solution $\vec{f}$ itself. $\underline{A}$ is typically sparse. B is made up of the coefficients associated with discretisation and of known values of $\vec{f}$ (e.g. past and boundary values).

We shall deal first with the case when $\underline{A}$ is independent from $\vec{f}$ so that the related system (D.1) is linear.

D.1 Direct methods for linear systems

The most common direct method is probably the Gauss one. Its starting point is the remark that, should $\underline{A}$ be a triangular matrix, the resolution of $\underline{A}X = B$ is trivial. (A *lower* triangular matrix has nonzero elements only on the diagonal and below, an *upper* triangular matrix has nonzero elements only on the diagonal and above). Consequently, the **Gauss method** consists in transforming the initial linear system into a triangular one by performing linear combination of the various relations of the system. This process is called Gauss elimination or triangulation. In order to ease the triangulation, it may be necessary to interchange rows and columns of matrix $\underline{A}$: this operation is termed *pivoting.*

It appears that the successive transformations which lead to triangulation can be summed up as the process of multiplying the initial system by a matrix $\underline{M}$:

$$\underline{M}\,\underline{A}\,X = \underline{M}\,B$$

such that,

$$\begin{aligned} \underline{M}\,\underline{A} &= \underline{U} \quad \text{upper triangular matrix} \\ \underline{M}\,B &= B' \quad \text{modified right-hand side term} \\ \underline{M} &= \underline{L}^{-1} \quad \text{lower triangular matrix} \end{aligned}$$

$\underline{A}$ is said to have undergone an LU decomposition. If the system has to be solved for several different right-hand side terms, it is useful to store $\underline{L}$ and $\underline{U}$. A complete description of Gauss algorithms can be found in (Goutal & Hérard, 1990) (chap. 1) or in (Press *et al.*, 1989) (chap. 2), in the last book with related samples of computer programs.

The matrix $\underline{A}$ is symmetric, positive and definite if it is regular and satisfies :

- $a_{i,j} = a_{j,i} \quad \forall i, j$
- $X^t\,\underline{A}\,X \geq 0 \quad \forall X$ (nb : X^t transposition of vector X)

In that case, it can be shown that there exists an unique lower triangular matrix $\underline{L}$, which diagonal components are strictly positive, so that $\underline{A} = \underline{L}\,\underline{L}^t$. There is a specific method available to compute $\underline{L}$, called the Cholesky factorisation method, which complete development is reported in (Goutal & Hérard, 1990) (section 1.2), for instance.

For specific types of sparse linear systems, there exist inversion methods more efficient than the Gauss one (cf. (Press *et al.*, 1989), chap. 2). This is the case for instance for tridiagonal, pentadiagonal and block-tridiagonal matrices. We indicate hereafter the **Thomas algorithm** (Roache, 1985; Press *et al.*, 1989; Fletcher, 1991) which applies to tridiagonal matrices such as :

$$\underline{A} = \begin{bmatrix} b_1 & c_1 & 0 & & & 0 \\ a_2 & b_2 & c_2 & & & \\ 0 & a_3 & b_3 & c_3 & & \\ & & . & . & . & \\ & & & a_{\mathcal{N}-1} & b_{\mathcal{N}-1} & c_{\mathcal{N}-1} \\ 0 & & & 0 & a_{\mathcal{N}} & b_{\mathcal{N}} \end{bmatrix}$$

$\underline{A}$ is factorised into $\underline{L}$ and $\underline{U}$ such that

$$\underline{L} = \begin{bmatrix} 1 & 0 & & & & 0 \\ \alpha_2 & 1 & 0 & & & \\ & \cdot & \cdot & & & \\ & & \cdot & \cdot & & \\ & & & \alpha_{\mathcal{N}-1} & 1 & 0 \\ 0 & & & 0 & \alpha_{\mathcal{N}} & 1 \end{bmatrix} \quad \text{and} \quad \underline{U} = \begin{bmatrix} \beta_1 & c_1 & 0 & & & 0 \\ 0 & \beta_2 & c_2 & & & \\ & & \cdot & \cdot & & \\ & & & \cdot & \cdot & \\ & & & 0 & \beta_{\mathcal{N}-1} & c_{\mathcal{N}-1} \\ 0 & & & & 0 & \beta_{\mathcal{N}} \end{bmatrix}$$

where the α_i and β_i obey

$$\begin{array}{lrcl} & \beta_1 & = & b_1 \\ \text{for } 2 \leq i \leq \mathcal{N} & \alpha_i & = & \dfrac{a_i}{\beta_{i-1}} \\ & \beta_i & = & b_i - \alpha_i c_{i-1} \end{array}$$

Solving $\underline{A}X = V$ is then achieved by a "double-sweep" on the computational nodes. In the forward sweep, we solve $\underline{L}Y = V$ by :

$$\begin{array}{lrcl} & y_1 & = & v_1 \\ \text{for } i = 2, 3, \ldots \mathcal{N} & y_i & = & v_i - \alpha_i y_{i-1} \end{array}$$

In the backward sweep, we solve $\underline{U}X = Y$ by :

$$\begin{array}{lrcl} & x_{\mathcal{N}} & = & y_{\mathcal{N}} \,/\, \beta_{\mathcal{N}} \\ \text{for } i = \mathcal{N}-1, \mathcal{N}-2, \ldots 1 & x_i & = & (y_i - c_i x_{i+1}) \,/\, \beta_i \end{array}$$

The Thomas algorithm is particularly economical as it requires only $3\,(\mathcal{N}-1)$ multiplications and $2\mathcal{N}-1$ divisions, whereas the general Gauss algorithm requires about $\mathcal{N}^3/3$ multiplications and $\mathcal{N}^2/2$ divisions (Goutal & Hérard, 1990). However, to prevent contamination by round-off errors, it is necessary that matrix $\underline{A}$ satisfies $|b_i| > |a_i| + |c_i|$ (Fletcher, 1991).

D.2 Iterative methods : General presentation

The general structure of iterative techniques is illustrated by rewriting $\underline{A}X = B$ as

$$(\,\underline{M} - \underline{N}\,)\,X = B \tag{D.2}$$

where $\underline{M}$ is a regular matrix, close to $\underline{A}$ in some sense (i.e. $\|\underline{M}\| \simeq \|\underline{A}\|$), but easy to inverse.

$$\underline{A}X = B \iff X = \underline{M}^{-1}\,\underline{N}\,X + \underline{M}^{-1}\,B$$

We can build a series $X^{(k)}$ approximating the solution X by :

$$X^{(k+1)} = \underline{M}^{-1}\,\underline{N}\,X^{(k)} + \underline{M}^{-1}\,B \tag{D.3}$$

$$\text{or } X^{(k+1)} = X^{(k)} - \underline{M}^{-1}\,R^{(k)} \tag{D.4}$$

where $R^{(k)}$ is the vector of equation residuals at the kth step of the iteration,

$$R^{(k)} = \underline{A}X^{(k)} - B \tag{D.5}$$

We have $R^{(k+1)} = \underline{N}\,\underline{M}^{-1}\,R^{(k)}$, so that the scheme described by D.3 or D.4 converges only if the spectral radius of $\underline{N}\,\underline{M}^{-1}$ is less than 1. (nb : we recall that the spectral radius of a matrix is its norm or, equivalently, its maximum eigenvalue (in absolute value)).

The various iterative methods differ in their choice of $\underline{M}$. Let us write $\underline{A} = D - L - U$ where D is the diagonal matrix, L and U strictly lower and upper triangular matrices respectively. We assume that $\forall i,\ a_{i,i} = d_{i,i} \neq 0$ (provided A is regular, there exists some pivoting which allows to satisfy this property).

- **Jacobi method**

  It amounts to :

  $$\underline{M} = D \quad \text{and} \quad \underline{N} = L + U \tag{D.6}$$

  so that,

  $$x_i^{(k+1)} = \frac{1}{a_{i,i}}\left(b_i - \sum_{j\neq i} a_{i,j}x_j^{(k)}\right) \tag{D.7}$$

- **Gauss - Seidel method**

  constitutes an improvement of the Jacobi method, with a rate of convergence typically twice as fast as Jacobi's one :

  $$\underline{M} = D - L \quad \text{and} \quad \underline{N} = U \tag{D.8}$$

  so that,

  $$x_i^{(k+1)} = \frac{1}{a_{i,i}}\left(b_i - \sum_{j=1}^{i-1} a_{i,j}x_j^{(k+1)} - \sum_{j=i+1}^{\mathcal{N}} a_{i,j}x_j^{(k)}\right) \tag{D.9}$$

- **Successive overrelaxation method (SOR)**

  provides a further improvement. A *relaxation parameter* λ ($\lambda \neq 0$) is introduced :

  $$\underline{M} = \frac{1}{\lambda}D - L \quad \text{and} \quad \underline{N} = \left(\frac{1}{\lambda} - 1\right)D + U \tag{D.10}$$

  so that, $X^{(k+1)}$ appears as a weighted average of $X^{(k)}$ and $X^{(k+1)}$ as it would have been forecasted by Gauss-Seidel method

  $$x_i^{(k+1)} = \frac{\lambda}{a_{i,i}}\left(b_i - \sum_{j=1}^{i-1} a_{i,j}x_j^{(k+1)} - \sum_{j=i+1}^{\mathcal{N}} a_{i,j}x_j^{(k)}\right) + (1-\lambda)\,x_i^{(k)} \tag{D.11}$$

Convergence is not easily studied in general case. However, in special situations, we have the following important results :

- If $\underline{A}$ is diagonally dominant (i.e. $|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$), Jacobi and Gauss-Seidel method do converge.
- The condition $\lambda \in]0,2[$ is necessary for SOR to converge. If $\underline{A}$ is definite positive, this condition proves to be sufficient.
- SOR rate of convergence is sensitive to the choice of λ. Let μ be the spectral radius of matrix $\underline{N}\,\underline{M}^{-1}$. The optimum choice would be

$$\lambda_{\text{opt}} = \frac{2}{1 + (1 - \mu)^{1/2}}$$

  Finding μ explicitly can be expensive. Consequently, the preferred strategy is to obtain an estimate of μ as the iteration proceeds and consequently update λ.

D.3 Iterative methods : acceleration techniques

The iterative methods presented above are effective but can prove rather slow. More efficient algorithms can be devised, either when the matrix $\underline{A}$ has some additional useful properties (cf D.3.1), either by resorting to multigrid techniques (cf D.3.2).

D.3.1 Gradient methods

Gradient methods apply to symmetric positive definite $\underline{A}$.
Then, $\underline{A}\,X - B$ appears to be the gradient of function $\mathcal{J}$:

$$\mathcal{J}(X) = \frac{1}{2} X^t \underline{A}\, X - B^t X \tag{D.12}$$

and solving the linear system is equivalent to finding a minimum of $\mathcal{J}$.

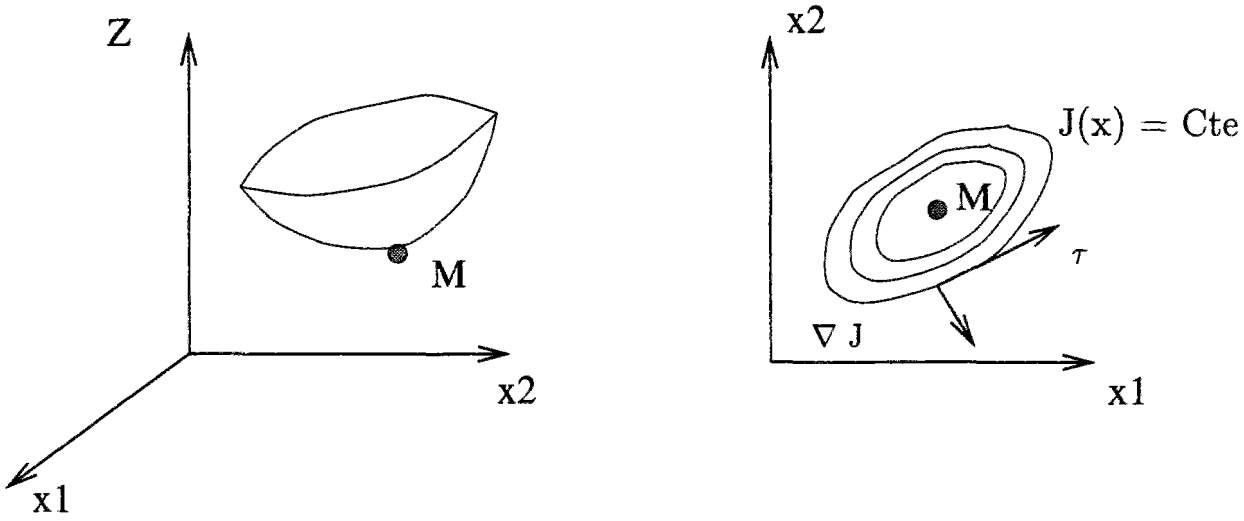


Figure D.1: Representation of a quadratic functional

As can be seen on figure D.1 (extracted from (Labadie, 1986)), $\mathcal{J}$ can be represented either as a surface $Z = \mathcal{J}(x_1, x_2, \ldots x_\mathcal{N})$ in the vectorial space $\Re^{\mathcal{N}+1}$, either by a series of isolines in $\Re^{\mathcal{N}}$. The center of these isolines is the minimum M. By differentiating the equation of isolines, it appears that, in the vicinity of such curve, the maximum variation of $\mathcal{J}$ occurs according to the direction normal to the isoline, i.e. the direction of the $\mathcal{J}$ gradient. Besides, as $\mathcal{J}$ is convex, its modulus decreases in the direction opposite to the gradient. This suggests intuitively that, starting from a point, the fastest way to reach the minimum is to follow the (opposite) local gradient direction. Gradient methods are based on this idea :

- $X^{(k)}$ being known, we look for the following iterate in the gradient direction :

$$\begin{aligned} X^{(k+1)} &= X^{(k)} - \rho \nabla \mathcal{J}\left(X^{(k)}\right) \quad \text{where } \rho \geq 0 \\ \text{i.e. } X^{(k+1)} &= X^{(k)} - \rho\left(\underline{A} X^{(k)} - B\right) \qquad \text{(D.13)} \\ \text{or } X^{(k+1)} &= X^{(k)} - \rho R^{(k)} \qquad \text{(D.14)} \end{aligned}$$

- ρ may be chosen constant or its value can be optimised at each time step, for instance $\rho = \rho_{\text{opt}}$ so that,

$$\mathcal{J}\left(X^{(k+1)}\right) = \min_{\rho} \mathcal{J}\left(X^{(k)} - \rho \nabla \mathcal{J}(X^{(k)})\right) \qquad \text{(D.15)}$$

  Such ρ_{opt} satisfies :

$$\rho_{\text{opt}} = \frac{R^{(k)\,t} R^{(k)}}{R^{(k)\,t} \underline{A} R^{(k)}} \qquad \text{(D.16)}$$

- Let us consider the case when ρ is chosen constant. From D.14, we deduce that :

$$R^{(k+1)} = (\underline{A} - \rho I) R^{(k)} \qquad \text{(D.17)}$$

where I denotes the identity matrix.

Consequently, the rate of convergence of the gradient method depends on the spectral radius r of matrix $\underline{A} - \rho I$. The optimum value for ρ would minimize this spectral radius. Let $\lambda_{\max}$ and $\lambda_{\min}$ be respectively the biggest and smallest eigenvalue of $\underline{A}$ (they are strictly positive as $\underline{A}$ is positive definite). It can be shown that

$$\rho_{\text{opt}} = \frac{2}{\lambda_{\max} + \lambda_{\min}} \tag{D.18}$$

$$\text{for which } r = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \tag{D.19}$$

Unfortunately, in practice, the exact eigenvalues of $\underline{A}$ are not easily evaluated. However, assuming this difficulty can somehow be overcome and formula D.18 applied, eq. D.19 teaches us that the rate of convergence is improved as the ratio $\lambda_{\max}/\lambda_{\min}$ tends to unity.

Simple gradient methods as presented above are based on a *local* optimisation criterion only. **Conjugate gradient methods** represent an improvement : instead of minimizing $\mathcal{J}$ at each iteration in the local gradient direction, we aim to minimize it in the vectorial space generated by the local and all previous gradients. This allows to take advantage of past information and guarantees that the minimum will be reached in a finite number of iterations (maximum number : the total space dimension $\mathcal{N}$). Besides, it can be demonstrated, by proper matrix and vector manipulation (Labadie, 1986), that the method does not require to store all previous gradients and that each iteration can be written in a simple and economic form (Labadie, 1986; Fletcher, 1991) ($X^{(k)}$ known, $d^{(k)}$ direction of descent at iteration k) :

$$R^{(k)} = \underline{A}X^{(k)} - B \tag{D.20}$$

$$d^{(k)} = R^{(k)} + \frac{R^{(k)\,t} R^{(k)}}{R^{(k-1)\,t} R^{(k-1)}} d^{(k-1)} \tag{D.21}$$

$$\rho^{(k)} = \frac{R^{(k)\,t} d^{(k)}}{d^{(k)\,t} \underline{A}\, d^{(k)}} \tag{D.22}$$

$$X^{(k+1)} = X^{(k)} - \rho^{(k)} d^{(k)} \tag{D.23}$$

The heaviest calculations in this sequence are the matrix - vector multiplications $\underline{A}X^{(k)}$ (for computing $R^{(k)}$) and $\underline{A}d^{(k)}$ (for computing $\rho^{(k)}$). The first one can be avoided by noticing that in fact :

$$R^{(k)} = R^{(k-1)} - \rho^{(k-1)} \underline{A} d^{(k-1)}$$

the product $\underline{A}d^{(k-1)}$ having been previously evaluated when computing $\rho^{(k-1)}$.

Preconditioning We mentionned above that, in the case of gradient methods with an optimal constant step ρ, the convergence is better if the spread of $\underline{A}$ eigenvalues is reduced. This proves to be true in a more general case.

Let us consider the system $\underline{A}\,X - B$ and a small pertubation δB of the right-hand side term. It can be shown that the resulting perturbation δX on the system solution satisfies :

$$\frac{\|\delta X\|}{\|X\|} \leq \|\underline{A}\|\,\|\underline{A}^{-1}\|\,\frac{\|\delta B\|}{\|B\|} \tag{D.24}$$

Alternatively, we could study the influence of a perturbation δA of matrix $\underline{A}$. We obtain now that :

$$\frac{\|\delta X\|}{\|X+\delta X\|} \leq \|\underline{A}\|\,\|\underline{A}^{-1}\|\,\frac{\|\delta A\|}{\|\underline{A}\|} \tag{D.25}$$

The influence of any perturbation appears to depend strongly on the *condition number* $\mathrm{cond}(\underline{A}) = \|\underline{A}\|\,\|\underline{A}^{-1}\|$ which should preferably be set as small as possible. $\mathrm{cond}(\underline{A})$ value depends on the chosen norm but, whichever it is, it can be demonstrated that $\mathrm{cond}(\underline{A}) \geq 1$ (Labadie, 1986) (sec. II.1). For symmetric matrix and the euclidian norm, $\mathrm{cond}(\underline{A})$ reduces to the ratio between maximum and minimum eigenvalues.

For most gradient algorithms, explicit formulas, more precise than eq. D.24, can be derived, that relate the rate of convergence of the algorithm to $\mathrm{cond}(\underline{A})$. For instance, for the standard conjugate gradient method, we have (Labadie, 1986; Goutal & Hérard, 1990) :

$$\| X - X^{(k)} \| \leq 2 \left(\frac{\sqrt{\mathrm{cond}(\underline{A})} - 1}{\sqrt{\mathrm{cond}(\underline{A})} + 1} \right)^{k} \| X - X^{(0)} \|$$

As $\mathrm{cond}(\underline{A})$ departs from its minimum value (1), the performance of the algorithm decreases, sometimes drastically. Consequently, it may prove relevant to multiply the linear system by some matrix $\underline{C}$ so that the product matrix $\underline{C}\,\underline{A}$ is more convenient from convergence point of view. $\underline{C}$ must be chosen so that it respects every desirable properties of the original linear system (e.g. positiveness, symmetry ...). Such operation is called *preconditioning*. Different preconditioning methods are reviewed and compared in (Goutal & Hérard, 1990; Sonneveld, 1984; Kaasschieter, 1986; Goutal, 1987; Iliev *et al.*, 1992) for instance. One of the most popular leads to the *biconjugate gradient method* (Fletcher, 1991) (chap. 6), which can be interestingly extended to non-symmetric linear systems (Sonneveld, 1984; Kaasschieter, 1986).

The efficiency of a preconditioning method depends on the balance between the increased complexity its implementation induces and the superior convergence rate it brings. It depends on the properties of the original linear system and consequently on the discretisation and differentiation applied to the original PDE. While some methods (i.e. preconditioning by the diagonal matrix D) are clearly recognized as being less efficient than others, the difference between more sophisticated methods is best appreciated when working on test problems.

D.3.2 Multigrid techniques

Multigrid methods have been originally developped for orthogonal discretisation grids, as frequently used by FDM. However, they are currently undergoing developments, in order to cope with irregular grids as generated by FEM. Hereafter, we content ourselves by indicating the outline of multigrid techniques and deal only with orthogonal grids.

Iterative techniques aim at reducing progressively the residual $R^{(k)}$ (as defined by eq. D.5). $R^{(k)}$ can be interpreted as a Fourier series of different frequency components (see more details in C.3.1). On a given discrete grid, the only allowable components are those which wavelength is a multiple of the grid space step δ. The highest frequency that can be represented corresponds to wavelength $2\,\delta$ (Fletcher, 1991), section 3.4 . It can be shown (Labadie, 1986) (sec. III.2.1) that procedures as Jacobi, Gauss-Seidel or SOR methods remove high frequency (short wavelength) components in a few iterations. Slow convergence is due to the painful removals of low frequency components. Yet, let us assume we consider a second grid with grid size δ' being a multiple of δ : some low frequency components on the fine grid become high frequency components on the coarse grid and thus, they ought to be eliminated quickly, should the resolution proceed on the coarse grid. More precisely, let us consider the equation :

$$\underline{A}_m X = B_m \tag{D.26}$$

which stems from the discretisation of some problem on a grid with gridsize δ_m. Let $X_m^{(k)}$ be an iterative approximation of the exact solution X^*. The corresponding error W_m obeys :

$$\underline{A}_m W_m = B_m - \underline{A}_m X_m^{(k)} = R_m \tag{D.27}$$

If we were able to solve exactly D.27, we could obtain X^* by $X^* = X_m^{(k)} - W_m$. The basic idea of multigrid methods is to solve approximately D.27 on a coarser grid with step $\delta_{m-1} > \delta_m$, by doing so to obtain an approximation $\widetilde{W_m}^{(k)}$ of W_m and then to follow on the iteration with :

$$X_m^{(k+1)} = X_m^{(k)} - \widetilde{W_m}^{(k)} \tag{D.28}$$

If the kth iteration on the fine grid was sufficiently advanced, W_m and R_m should be smooth, i.e. the amplitudes of their high frequency components (which cannot be resolved on the coarse grid) should be small. Consequently, $\widetilde{W_m}^{(k)}$ should closely approximate W_m and thus, iteration D.28 should be quite efficient.

In fact, multigrid methods work with a sequence of grids $m = 1, 2, \ldots M$ where the grid size ratio δ_{m+1}/δ_m is 0.5 . Following (Fletcher, 1991), we introduce the following notations :

- On each grid ($1 \leq m \leq M$) the problem under study is discretised by $\underline{A}_m X = B_m$.

- $Y_m^{n_1} = \text{relax}^{n_1}(Y_m, \underline{A}_m, \xi_m)$ denotes the approximation to the solution of system $\underline{A}_m Y_m = \xi_m$ obtained after n_1 iterations through the simple iterative method (Jacobi, Gauss-Seidel, SOR ...) to which multigrid acceleration is applied.
- I_m^{m+1} denotes an *interpolation* operator from grid number m to finer grid $m+1$. I_m^{m+1} could be linear, cubic (Fletcher, 1991) (sec. 6.3.5) ...
- I_{m+1}^m denotes a *restriction* (projection) operator from grid number $m+1$ to coarser grid m, see examples in (Fletcher, 1991) (sec. 6.3.5) .

Let $X_M^{(k)}$ be the current approximation to solution on the finer grid. The next iterate $X_M^{(k+1)}$ is obtained through the following cycle :

1. Initialisation of the cycle. Set :

$$m = M \qquad W_m = X_M^{(k)} \qquad R_m = B_M$$

2. Proceed from finest to coarsest grid :
   (a) work out simple iterations. obtain
   $$W_m^{n_1} = \text{relax}^{n_1}(W_m, \underline{A}_m, R_m)$$
   (b) make a projection of the residual on the next coarser grid
   $$R_{m-1} = I_m^{m-1}(R_m - \underline{A}_m W_m^{n_1})$$
   (c) set $m-1 \rightarrow m$. If $m > 1$, go on with step (a).
3. Solve $\underline{A}_1 W_1 = R_1$ exactly.
4. Proceed from coarsest to finest grid :
   (a) set $m+1 \rightarrow m$
   (b) Interpolate correction on the finer grid
   $$W'_m = I_{m-1}^m W_{m-1}$$
   (c) work out simple iterations *with initial guess* W'_m. obtain
   $$W_m = \text{relax}^{n_2}(W'_m, \underline{A}_m, R_m)$$
   (d) If $m < M$, go on with step (a).

5. Obtain the next multigrid iterate (MGI) by

$$X_M^{(k+1)} = X_M^{(k)} - W_M$$

From now on, we shall denote $X_M^{(k+1)} = \text{MGI}\left(X_M^{(k)}, \underline{A}_M, B_M \right)$

With a careful choice of the coarsest grid, the solution of step 3 can be achieved economically by a *direct* method.

An even more important acceleration can be achieved if the multigrid cycle is applied to each intermediate grid. Such methods are termed **full multigrid methods (FMG)** and proceed as follows (cf. (Labadie, 1986) (sec. III.2.3) and (Fletcher, 1991) (sec. 6.3.5)) :

1. Initialisation : solve $\underline{A}_1 V_1 = B_1$ exactly.
2. Set $m+1 \rightarrow m$. Set an initial guess ($k = 0$) by

$$X_m^{(0)} = I_{m-1}^m V_{m-1}$$

   (a) set $k+1 \rightarrow k$

   (b) The multigrid cycle proceeds on all grids with numbers m, $m-1, \ldots 1$ i.e. coarser than grid number m :

$$\begin{aligned} X_m^{(k)} &= \text{MGI}\left(X_m^{(k-1)}, \underline{A}_m, B_m \right) \\ R_m^{(k)} &= B_m - \underline{A}_m X_m^{(k)} \end{aligned}$$

   (c) while $\| R_m^{(k)} \| > \epsilon$, go on with step (a)

3. If $m = M$, stop. Else go on with step (2).

In order to apply as efficiently as possible the multigrid techniques, the following points need to be optimised :

- choice of parameters n_1, n_2 and ϵ ;
- choice of the grids and especially of the coarsest one;
- choice of the interpolation and restriction operators;
- treatment of boundaries conditions on the different grids;
- choice of the "inner" iterative method, which smooths out the high frequency components.

Examples of multigrid techniques applications (notably their extension to non-linear systems) may be found in (Phillips & Schmidt, 1984; Phillips & Schmidt, 1985) (advection-diffusion equation), (Fletcher, 1991) (Euler equations, section 14.2) or in more specialized textbooks as (Hackbusch, 1985) .

D.4 Methods suitable for FDM schemes

D.4.1 Alternate direction method

An iterative method particularly suitable for systems produced by multidimensional FDM is the Peaceman-Rachford alternate direction method (ADI) (Labadie, 1986; Goutal & Hérard, 1990). It applies to a matrix $\underline{A}$ definite and positive. Let us assume there exist two positive definite matrix A_1 and A_2 such that $\underline{A} = A_1 + A_2$. Introducing a positive parameter r, we can write two different decompositions of $\underline{A}$:

$$\underline{A} = (A_1 + rI) + (A_2 - rI) \tag{D.29}$$

$$\underline{A} = (A_2 + rI) + (A_1 - rI) \tag{D.30}$$

and we use alternatively the two decompositions, so that the iteration proceeds as :

$$(A_1 + rI)\, X^{(k+1/2)} = B - (A_2 - rI)\, X^{(k)} \tag{D.31}$$

$$(A_2 + rI)\; X^{(k+1)} = B - (A_1 - rI)\, X^{(k+1/2)} \tag{D.32}$$

This is equivalent to a single iteration with the matrix

$$(A_2 + rI)^{-1}\, (A_1 - rI)\, (A_1 + rI)^{-1} (A_2 - rI)$$

which spectral radius is bounded by the product

$$\max_j \left| \frac{r - \lambda_j^1}{r + \lambda_j^1} \right| \times \max_j \left| \frac{r - \lambda_j^2}{r + \lambda_j^2} \right|$$

where the notations λ_j^1 and λ_j^2 denote the eigenvalues of A_1 and A_2 respectively. As these eigenvalues are positive, it is straightforward to check that $\forall r > 0$, the above product is strictly inferior to one, so that the algorithm is convergent. By the way, we can note that the convergence does not require both A_1 and A_2 to be definite : one is enough.

The method is attractive if $\underline{A}$ corresponds to an operator which can be splitted in different directions e.g. $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$: A_1 and A_2 are chosen so that they correspond to *monodimensional operators.* Consequently, the solution of a multidimensional problem can be achieved

through an iteration over much simpler, monodimensional systems.

This method has been applied extensively to the case of shallow water flow equations. However, when applied to transient problems in irregular (as regards bathymetry or geometry) computational domains, some ADI methods display large unaccuracies (Weare, 1979; Stelling *et al.*, 1986) unless the time step is drastically reduced, which makes the methods poorly efficient. In such situations, the remedy is to turn to conjugate gradient methods (e.g. (Wilders *et al.*, 1988)) or to consider an other kind of monodimensional decomposition method as exposed in the following paragraph.

D.4.2 Method of decomposition with coordination

Once again, we consider the decomposition of $\underline{A}$ into two positive definite matrix A_1 and A_2. We assume that the right-hand side term B can also be divided into two parts. Consequently,

$$\underline{A}\,X \;=\; B \quad \Longleftrightarrow \quad (A_1 + A_2)\,X \;=\; B_1 + B_2 \tag{D.33}$$

Let us introduce the functions

$$\mathcal{J}_i(X) \;=\; X^t A_i X \;-\; B_i \quad \text{for } i \;=\; 1,\,2 \tag{D.34}$$

The problem D.33 is equivalent to the minimization of function $\mathcal{J} = \mathcal{J}_1 + \mathcal{J}_2$. The form of $\mathcal{J}$ suggests to split the minimization problem into two parts, corresponding to each $\mathcal{J}_i$, $i = 1, 2$. In fact, solving eq. D.33 amounts to find :

$$\min_{X_1, X_2\ (X_1 = X_2)} \mathcal{J}_1(X_1) \;+\; \mathcal{J}_2(X_2) \tag{D.35}$$

which is a problem of minimization under constraint.

Such problems are usually solved by introducing an additional variable, the Lagrange multiplier q. Let us consider the function

$$\mathcal{L}(X_1, X_2, q) \;=\; \mathcal{J}_1(X_1) \;+\; \mathcal{J}_2(X_2) \;+\; q\,(X_1 - X_2) \tag{D.36}$$

The problem described by relation D.35 is equivalent to

$$\max_{q} \min_{X_1, X_2} \mathcal{L}(X_1, X_2, q) \tag{D.37}$$

which can be solved by usual gradient techniques as it is a problem *without external constraint* and which involves only quadratic and/or convex functions. The equivalence is immediate when

looking at the conditions for optimality corresponding to D.37, as they are obtained by deriving $\mathcal{L}$:

$$\frac{\partial \mathcal{L}}{\partial X_1} = 0 \iff A_1 X_1 = B_1 - q \tag{D.38}$$

$$\frac{\partial \mathcal{L}}{\partial X_2} = 0 \iff A_2 X_2 = B_2 + q \tag{D.39}$$

$$\frac{\partial \mathcal{L}}{\partial q} = 0 \iff X_1 = X_2 \tag{D.40}$$

The initial problem which unknown was X can now be interpreted as a problem which main variable is the Lagrange multiplier : it consists in maximizing the $q-$function $\mathcal{J}^\star(q) = \min_{X_1,X_2} \mathcal{L}(X_1, X_2, q)$. An iterative solution of the transformed problem proceeds as follows :

1. Let $q^{(k)}$ be the present value of the Lagrange multiplier.

2. We determine X_1 and X_2 corresponding to $\mathcal{J}^\star(q)$ with the help of relations D.38 and D.39 :

$$\begin{aligned} A_1 X_1^{(k)} &= B_1 - q^{(k)} \\ A_2 X_2^{(k)} &= B_2 + q^{(k)} \end{aligned}$$

3. Then, the estimation of q is advanced according to the gradient direction :

$$q^{(k+1)} = q^{(k)} + \rho \frac{\partial \mathcal{J}^\star}{\partial q}\left(q^{(k)}\right) \quad (\rho > 0) \tag{D.41}$$

where the gradient of $\mathcal{J}^\star$ appears to be simply

$$\frac{\partial \mathcal{J}^\star}{\partial q}\left(q^{(k)}\right) = X_1^{(k)} - X_2^{(k)}$$

This time, as we are dealing with a maximization problem, we progress truly in the gradient direction, not in the opposite direction.

We can of course substitute to relation D.41 a more sophisticated updating formula, i.e. derived from conjugate gradient methods.

4. The iteration stops when

$$\| X_1^{(k)} - X_2^{(k)} \| \leq \epsilon$$

ϵ is a parameter which controls the accuracy of the iterative solution.

The individual systems described by D.38 and D.39 could be solved iteratively. However, the method of decomposition with coordination really proves to be powerful and computationnally

efficient when A_1 and A_2 are simple matrices, easy to inverse. Such is the case when the related differential operators are monodimensional : when applied to Poisson (Labadie, 1986) or St-Venant two-dimensional equations (Labadie, 1986; Benque *et al.*, 1982) for instance, A_1 and A_2 can be properly reduced to tridiagonal systems.

D.5 Non-linear systems

The system D.1 can be linearised in several ways. The simplest is to set $\underline{A} \simeq \underline{A}(\vec{f}^n)$. This often proves to be a fair enough approximation, provided the time step is kept sufficiently small so that the variations of $\underline{A}$ are smooth at this time scale. However, in order to achieve some cost-effectiveness, one may want to use bigger time steps for which the above approximation is too crude. One may then consider to apply **Newton's methods**. As demonstrated in (Press *et al.*, 1989) (chap. 9), Newton-Raphson method can be a powerful tool for finding the root of a (monovariable or multivariable) function. A typical problem involves N functional relations to be zeroed :

$$g_i(x_1, x_2, \dots x_N) = 0 \quad i = 1, 2, \dots N \tag{D.42}$$

Let X denote the entire vector of values x_i. Then, in the neighborhood of X, each g_i can be expanded in Taylor series

$$g_i(X + \delta X) = g_i(X) + \sum_{j=1}^{N} \frac{\partial g_i}{\partial x_j} \delta x_j + 0\left(\delta X^2\right) \tag{D.43}$$

By neglecting terms of order δX^2 and higher, we obtain a set of linear equations for the corrections δx_j that move each function closer to zero simultaneously. These equations can be written in matrix form :

$$[\nabla G]\, \delta X = -G \tag{D.44}$$

where G denotes the vector of components $g_i(X)$ and $[\nabla G]$ the matrix defined by

$$[\nabla G]_{i,j} = \frac{\partial g_i}{\partial x_j}(X) \tag{D.45}$$

Once the above system is solved, the corrections are added to the solution vector :

$$X^{\text{new}} = X^{\text{old}} + \delta X \tag{D.46}$$

An attractive feature of Newton's method is that, close to the root, it converges quadratically, i.e.

$$\| X^{(k+1)} - X_c \| \leq K \| X^{(k)} - X_c \| \tag{D.47}$$

where X_c denotes the root, $X^{(k)}$ its estimate after k Newton's iterations and K some constant (which value is related to the ratio of second and first-order derivatives of the g_i). Unfortunately,

if the initial estimation $X^{(0)}$ is not nearby the root, the Newton's method may also spectacularly fail to converge !

Newton's method formalism can be applied to the solution of system (D.1) by interpreting (D.1) as the problem of zeroing the residual :

$$R(X) = \underline{A}(X)\,X - B \tag{D.48}$$

Then, Newton's method can be written as

$$X^{(k+1)} = X^{(k)} - \left(\underline{J}^{(k)}\right)^{-1} R^{(k)} \tag{D.49}$$

where $\underline{J}^{(k)}$ is the Jacobian matrix, which elements are :

$$J_{i,j}^{(k)} = \frac{\partial R_i^{(k)}}{\partial X_j^{(k)}} \tag{D.50}$$

Introducing $\delta X^{(k)} = X^{(k+1)} - X^{(k)}$, we can rewrite D.49 as :

$$\underline{J}^{(k)}\,\delta X^{(k)} = -R^{(k)} \tag{D.51}$$

$$X^{(k+1)} = X^{(k)} + \delta X^{(k)} \tag{D.52}$$

Consequently, the solution of non-linear system D.1 can be reached through an iterative process where each step involves the solution of a **linear** system of equations as described by eq. D.51. Usually, the iteration starts with $X^{(0)} = \vec{f}^{n}$. If the time step is not too large with respect to the transient features of the problem, this should prove convenient.

In fact, Newton's method updating is seldom applied as indicated by D.52. Indeed, the corrections, notably in the first steps, can be too large, so large that the current estimation eventually falls too far apart from the root and that the method consequently fails to converge. Thus, the corrections are not implemented in full but rather :

$$X^{(k+1)} = X^{(k)} + \rho^{(k)}\,\delta X^{(k)} \tag{D.53}$$

In order to accelerate the convergence, $\rho^{(k)}$ can be optimised. The principle of such optimisation is exposed in (Fletcher, 1991) (section 6.1) :

- Three different values of $\rho^{(k)}$, namely ρ_i, $i = 1, 2, 3$, are chosen. For each one, we compute new solutions

  $$X_m^{(k+1)} = X^{(k)} + \rho_m\,\delta X^{(k)}$$

  and the corresponding new residual $R_m^{(k+1)}$

- We compute the (euclidian) norm of each residual (more exactly, its square). We assume now that the residual mean squares (rms) are a quadratic function of parameter ρ. As the rms are known for three different values of ρ, this quadratic function is entirely defined.
- We look for the ρ value which minimizes the rms and use it when applying D.53.

However, this strategy can prove expensive if the residuals are complicated.

The main contribution to the execution time of Newton's method stems generally from the solution of system D.51. The simplest way to alleviate this stage is to "freeze" $J^{(k)}$ on a few (δk) subsequent iterations. Thus, it needs to be factorised only once every δk steps. Yet, such simplification may lower the accuracy and hinder the convergence. An alternative approach is provided by the **Quasi-Newton methods** (Fletcher, 1991) (section 6.1), (Goutal & Hérard, 1990) (section 3.9) : the combination of D.51 and D.53 is replaced by

$$X^{(k+1)} = X^{(k)} - \rho^{(k)} \underline{H}^{(k)} R^{(k)} \tag{D.54}$$

where $\underline{H}^{(k)}$ is an approximation to $\left(\underline{J}^{(k)}\right)^{-1}$. $\underline{H}^{(k)}$ is modified at each iteration so that it approaches more and more closely $\left(\underline{J}^{(k)}\right)^{-1}$ as $X^{(k)}$ converges to the exact solution. To different updating formulas for $\underline{H}$ correspond different quasi-newton methods.

As reported by (Fletcher, 1991), the effectiveness of quasi-newton methods often depends on $\underline{J}$ having special properties, e.g. being positive and definite. Consequently, their applicability needs to be examined on a case-by-case basis : it depends on the equations considered and on a relevant discretisation of these. Process splitting (cf 3.6) may contribute to produce well-behaved $\underline{J}$. One popular quasi-newton method is the BFGS algorithm (Fletcher, 1991) (section 6.1.4). It can be shown that applying a preconditioned conjugate gradient method to the problem of minimizing function $\mathcal{J}$ (as defined by D.12) with the condition matrix $H^{(0)}$ is equivalent to applying a BFGS algorithm initialised by the same couple $\left(X^{(0)}, H^{(0)}\right)$ (Goutal & Hérard, 1990) (section 3.9).

In summary, it appears that non-linear systems can be either solved by specific techniques as the Quasi-Newton methods, either replaced by the subsequent resolution of linear systems to which the methods exposed in sections D.1 to D.4 can be applied.

Appendix E

Additional informations about advection-diffusion algorithms

E.1 Backward characteristic methods

E.1.1 Backtracking the fluid particles

The characteristic lines obey equation

$$\frac{\mathrm{D}\vec{X}}{\mathrm{D}t} = \vec{U} \tag{E.1}$$

where $\vec{X}$ and $\vec{U}$ denote respectively the coordinates and velocity vector. $\vec{U}$ is a function of space and time. In two-dimensional situations, we denote u and v the components of $\vec{U}$ in the $x-$ and $y-$ directions respectively.
At each time level, we need to solve E.1 backward in time for each node of the computational grid.

The procedure we use is a multiple-steps one, i.e. we split the time step Δt into several ones $(\delta t)_k, k = 1, \ldots N$ and integrate E.1 over each smaller time interval. Before explaining on what criterion we split the integration, let us explain what is done within each time interval. Let Q_k be the location on the characteristic lines after k steps, consequently at time $t_k = t^{n+1} - \sum_{i=1}^{k} (\delta t)_i$. The next location Q_{k+1} is determined with a second-order accurate Runge-Kutta method :

1. compute $\vec{V} = U(Q_k, t_k)$ by linear (bilinear in two dimensions) interpolation in space, as well as in time;
2. compute intermediary point A which location is $\vec{X}(Q_k) - \vec{X}(A) = \frac{1}{2}(\delta t)_{k+1}\vec{V}$;

3. update $\vec{V}$ by $\vec{V} = U(A, t_k - \frac{1}{2}(\delta t)_{k+1})$ if A lies inside the computational domain, otherwise do not modify $\vec{V}$;

4. determine the location of Q_{k+1} by $\vec{X}(Q_k) - \vec{X}(Q_{k+1}) = (\delta t)_{k+1}\vec{V}$.

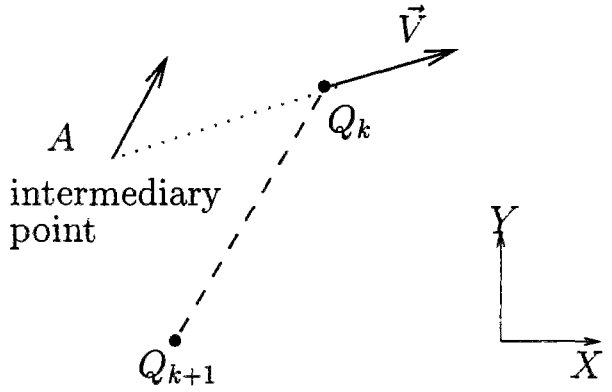


Figure E.1: One stage in trajectory computation

When the characteristic is entirely contained within one grid cell, it can be shown that splitting the time step (and hence the computation of the trajectory) in two or three intervals of same length allows to determine very accurately the location of its departure point. Let m be the number of cells crossed during Δt by a particle arriving at node $\mathcal{N}$: dividing Δt in $2m$ or $3m$ steps should warrant a good precision in the location of the departure point of the characteristic line. What we need is therefore an evaluation of m. This is done as follows :

1. We consider the sign of $\vec{U}(\mathcal{N})$ components : it indicates which cell of the computational grid do the particles cross immediatly before reaching $\mathcal{N}$. We consider the maximum velocity, according to each coordinate direction, over the nodes defining this cell (respectively 2 and 4 nodes in one-dimensional and two-dimensional cases).

2. Neglecting further variations of the velocity or of the grid spacing, we estimate the number of cells crossed in each direction by a particle reaching $\mathcal{N}$ by

$$m_x = \frac{|u|_{\max}\,\Delta t}{\Delta x} \qquad \text{and} \qquad m_y = \frac{|v|_{\max}\,\Delta t}{\Delta y}$$

   where Δx and Δy denote the grid size respectively in the $x-$ and $y-$ directions.

3. m is finally set to $m = \max(m_x, m_y)$

Then, we can determine an average sub-time step by $\overline{\delta t} = \Delta t/2m$ or $\overline{\delta t} = \Delta t/3m$. The computation begins setting $(\delta t)_1 = \overline{\delta t}$. Then we apply,

$$(\delta t)_{k+1} = \min\left(\overline{\delta t},\, t_k - t^n\right)$$

However, the effective time interval corresponding to step $k+1$ of the trajectory integration can be shorter than proposed above, because of possible intersection with boundaries.

When the next point Q_{k+1} of the trajectory traces outside the computational domain, special algorithms are needed. First, the location of the intersection I between segment $[Q_k, Q_{k+1}]$ and the domain boundary is computed, as is the time of intersection. Then, we examine the boundary type.

- If it is an open boundary, further integration is canceled. The boundary is obviously an inflow boundary, where the scalar value has to be prescribed. Knowing the location where and time when the particle enters the domain, the scalar value at I and hence at node $\mathcal{N}$ can readily be computed.
- If it is a land boundary, the time interval $(\delta t)_{k+1}$ is reduced so that Q_{k+1} matches I. Then, we follow on the integration. If the velocity field satisfies consistent hydraulic conditions at closed boundaries, $\vec{U}(Q_{k+1})$ should be parallel to the boundary. If it is not, we constrain the particle to travel along the grid boundary by taking into account only the velocity component parallel to boundary.

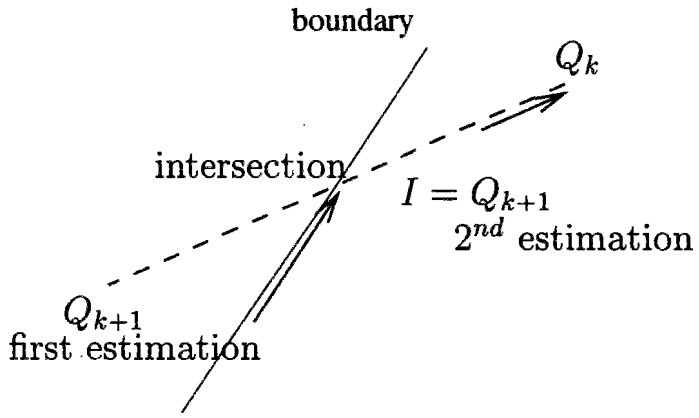


Figure E.2: Intersection with land boundary

E.1.2 Two-dimensional extension of interpolating forms

In one-dimensional case, interpolating forms such as the Hermite cubic polynomial used in Holly-Preissmann and Rasch-Williamson scheme or the quadratic polynomial used in Li's minimax method can all be written :

$$P_i(O) = a\,C_{i-1} + b\,C_i + d\,C_{x\,i-1} + e\,C_{x\,i} \tag{E.2}$$

where C_j, $j = i, i-1$ and C_{xj}, $j = i, i-1$ denote respectively the value of the dependent variables and its derivative estimates at the endpoints of the cell containing the characteristic foot O and a, b, d, e are polynomial functions of the relative position of O within the cell, namely $\theta = (x_i - x(O)) / (x_i - x_{i-1})$. Let us now consider the two dimensional situation as sketched in figure E.3.

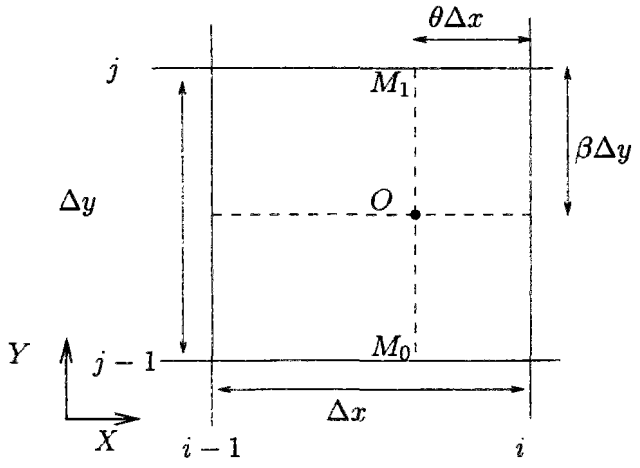


Figure E.3: Definition sketch for two-dimensional interpolation

The dependent variables and its derivative estimates are known at each grid node. Then, the value $C(O)$ can be evaluated by applying successively one-dimensional interpolators along each coordinate direction (Holly & Usseglio-Polatera, 1984) :

1. Using C and C_x at nodes $(i-1, j)$ and (i, j) calculate C at point M_1 by applying E.2;
2. Estimate similarly C_y at point M_1;
3. Using C and C_x at nodes $(i-1, j-1)$ and $(i, j-1)$ calculate C at point M_0 by applying E.2;
4. Estimate similarly C_y at point M_0;
5. Using C and C_y at points M_0 and M_1, calculate C at point O, applying E.2 (with $\beta = (y_j - y(O)) / (y_j - y_{j-1})$ instead of θ).

There are different ways for achieving steps 2 and 4. (Holly & Usseglio-Polatera, 1984) suggest to compute the C_y by applying again interpolator E.2. To do so, it is necessary to get an estimate of the cross-derivatives $\frac{\partial^2 C}{\partial x \partial y} = C_{xy}$. (Holly & Usseglio-Polatera, 1984) propose to use the centered approximation :

$$\left(C^n_{xy}\right)_{i,j} = \frac{1}{2}\left[\frac{(C_x)_{i,j+1} - (C_x)_{i,j-1}}{y_{j+1} - y_{j-1}}\right] + \frac{1}{2}\left[\frac{(C_y)_{i+1,j} - (C_y)_{i-1,j}}{x_{i+1} - x_{i-1}}\right] \qquad \text{(E.3)}$$

As regards Li and Rasch-Williamson method, the first-order derivatives at one node are a combination (a differentiation) based on C values at the surrounding nodes. Thus, instead of using E.2 and E.3, the derivatives $C_y(M_0)$ and $C_y(M_1)$ could be obtained by the following approach :

1. Define points M_{k+1} by $x(M_{k+1}) = x(0)$ and $y(M_{k+1}) = y_{j+k}$. For Li scheme, it is necessary to do so for $k = 1$ and $k = -2$; for Rasch-Williamson method, we need to take $k = 2, 1, -2 - 3$.

2. Using C and C_x at nodes $(i-1, j+k)$ and $(i, j+k)$ calculate C at points M_{k+1} by applying E.2;

3. Calculate C_y at M_0 and M_1 by applying the arithmetic or Akima estimate (respectively for Li and Rasch scheme) between points M_{k+1}.

For Holly-Preissman method, it is also necessary to compute the first-order spatial derivatives at O. We proceed as for the concentration itself, except that the interpolator is slightly different (see eq. 4.20).

The extension of Dan N'Guyen scheme is slightly different. If, as sketched in figure E.3, the origin O of the characteristic lies within the cell which vertices are (i, j), $(i-1, j)$, $(i-1, j-1)$ and $(i, j-1)$, $C(O)$ is interpolated from the second-order Taylor series expansions of distribution C in the vicinity of each vertex :

$$\begin{aligned} C(O) &= a(1-\theta)\,a(1-\beta)\,\Upsilon_{i,j}(O) + a(\theta)\,a(1-\beta)\,\Upsilon_{i-1,j}(O) \\ &+ a(\theta)\,a(\beta)\,\Upsilon_{i-1,j-1}(O) + a(1-\theta)\,a(\beta)\,\Upsilon_{i,j-1}(O) \end{aligned} \tag{E.4}$$

where $a(\theta) = \theta^2\,(3 - 2\theta)$ and the Taylor series read :

$$\begin{aligned} \Upsilon_{i,j}(O) &= C_{i,j} + (x(O) - x_i)\,dx_{i,j} + \frac{1}{2}(x(O) - x_i)^2\,dxx_{i,j} \\ &\quad + (y(O) - y_j)\,dy_{i,j} + \frac{1}{2}(y(O) - y_j)^2\,dyy_{i,j} \\ &\quad + (x(O) - x_i)\,(y(O) - y_j)\,dxy_{i,j} \end{aligned} \tag{E.5}$$

where the symbols dx, dy, dxx, dyy, dxy denote centred finite difference approximations to the first-order, second-order and cross spatial derivatives. Thus, if the grid spacing is uniform, $\Upsilon_{i,j}$ reduces to :

$$\begin{aligned} \Upsilon_{i,j}(O) &= C_{i,j} + \frac{1}{2}(\theta - 1)\,(C_{i+1,j} - C_{i-1,j}) + \frac{1}{2}(\theta - 1)^2\,(C_{i+1,j} - 2C_{i,j} + C_{i-1,j}) \\ &\quad + \frac{1}{2}(\beta - 1)\,(C_{i,j+1} - C_{i,j-1}) + \frac{1}{2}(\beta - 1)^2\,(C_{i,j+1} - 2C_{i,j} + C_{i,j-1}) \\ &\quad + \frac{1}{4}(\theta - 1)(\beta - 1)\,(C_{i+1,j+1} - C_{i+1,j-1} - C_{i-1,j+1} + C_{i-1,j-1}) \end{aligned} \tag{E.6}$$

E.1.3 Computation of corrective terms (Holly-Preissmann method)

As explained in 4.1.2.1, the first-order derivatives of the transported scalar obey slightly different advection equations, respectively

$$\frac{\partial C_x}{\partial t} + U\frac{\partial C_x}{\partial x} = -\frac{\partial U}{\partial x}.C_x \tag{E.7}$$

in one dimensional situation and

$$\frac{\partial C_x}{\partial t} + U\frac{\partial C_x}{\partial x} + V\frac{\partial C_x}{\partial y} = -\frac{\partial U}{\partial x}.C_x - \frac{\partial V}{\partial x}.C_y \tag{E.8}$$

$$\frac{\partial C_y}{\partial t} + U\frac{\partial C_y}{\partial x} + V\frac{\partial C_y}{\partial y} = -\frac{\partial U}{\partial y}.C_x - \frac{\partial V}{\partial y}.C_y \tag{E.9}$$

in two dimensional case (C_x and C_y denoting the first-order derivatives respectively along the $x-$ and $y-$ directions). The right-hand side terms in equations E.8 and E.9 couple them.

Let M and O be two points along the same characteristic line, O being reached at time t and M at $t + \Delta t$. While, in case of pure advection, $C(\mathrm{M}) = C(O)$, we obtain as regards the derivatives :

$$C_x(M) - C_x(O) = -\int_t^{t+\Delta t} (\, U_x C_x + V_x C_y \,)\,(\vec{x},\tau)\, d\tau \tag{E.10}$$

$$C_y(M) - C_y(O) = -\int_t^{t+\Delta t} (\, U_y C_x + V_y C_y \,)\,(\vec{x},\tau)\, d\tau \tag{E.11}$$

As detailed in E.1.1, the method used to backtrack the characteristic lines breaks the total trajectory between O and M into a series of segments. The right-hand side corrective integrals are evaluated by applying the trapezium rule on each segment. Let Q_k and Q_{k+1} be the end-points of segment number k, traveled in time $(\delta t)_k$. Then :

$$\int_{Q_{k+1}}^{Q_k} (\, U_x C_x + V_x C_y \,)\,(\vec{x},\tau)\, d\tau \simeq \frac{(\delta t)_k}{2}\, [\, (U_x C_x + V_x C_y)(Q_k) + (U_x C_x + V_x C_y)(Q_{k+1}) \,] \tag{E.12}$$

Thus, computing the corrective terms implies to assess the derivatives of velocities and scalar at several locations along the trajectory.

- As, while performing this integration, we do not know the concentration and derivative values at the next time level, the corrective terms evaluation is necessarily explicit, i.e. it relies on derivative values at the preceding time level only.
- The flow field derivatives at Q_k are estimated with a first-order approximation (notations indicated in figure E.4) :

$$U_x(Q_k) = (1-\beta)\frac{U(M_1) - U(M_2)}{\Delta x} + \beta\frac{U(M_4) - U(M_3)}{\Delta x}$$

$$U_y(Q_k) = (1-\theta)\frac{U(M_1) - U(M_4)}{\Delta y} + \theta\frac{U(M_2) - U(M_3)}{\Delta y}$$

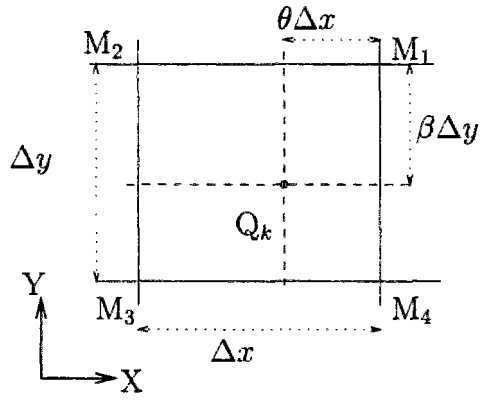


Figure E.4: Estimation of flow velocity derivatives : notation

- The scalar derivatives at Q_k are estimated by interpolation between the surrounding grid nodes. The simplest interpolator is bilinear. Alternatively, the derivatives may be computed with the same bicubic interpolator used at the foot of each trajectory. The consequences of this choice are discussed further down (see Appendix F.4).

E.2 Treatment of the diffusion operator

The transport equation reads (in tensor notations) :

$$\text{Conservative form} \qquad \frac{\partial \psi}{\partial t} + \frac{\partial u_j \psi}{\partial x_j} = S \tag{E.13}$$

$$\text{Non-conservative form} \qquad \frac{\partial C}{\partial t} + u_j \frac{\partial C}{\partial x_j} = S' \tag{E.14}$$

where u_j denotes the flow velocity along the coordinate direction j, C stands for the scalar concentration and ψ is respectively equal to C, hC and AC in three, two and one-dimensional cases, with h denoting the water-depth and A the wet section.

Without external or biogeochemical sink/source terms, S and S' account solely for dispersion effects. Their general expressions are then :

$$\text{in one-dimensional case,} \quad S = \frac{\partial}{\partial x}\left(A\Gamma \frac{\partial C}{\partial x} \right) \tag{E.15}$$

$$\begin{aligned} \text{in two-dimensional case} \quad S = & \frac{\partial}{\partial x}\left(h\Gamma_{xx} \frac{\partial C}{\partial x} \right) + \frac{\partial}{\partial y}\left(h\Gamma_{yy} \frac{\partial C}{\partial y} \right) \\ & + \frac{\partial}{\partial x}\left(h\Gamma_{xy} \frac{\partial C}{\partial y} \right) + \frac{\partial}{\partial y}\left(h\Gamma_{yx} \frac{\partial C}{\partial x} \right) \end{aligned} \tag{E.16}$$

where Γ ($m^2.s^{-1}$) denotes the longitudinal dispersion coefficient (one-dimensional situation) and Γ_{xx}, Γ_{yy}, Γ_{xy}, Γ_{yx} ($m^2.s^{-1}$) its two-dimensional counterparts. S' is respectively equal to

S/A and S/h in one and two-dimensional situations.
The cross-terms Γ_{xy} and Γ_{yx} are equal. Their occurence is due to the fact that the streamlines are not always aligned with the coordinates direction (cf section 2.4.5). Let $\Gamma_{\|}$ and $\Gamma_{\perp}$ be the dispersion coefficients respectively in the streamwise direction and in the direction perpendicular to streamlines. Those are the quantities for which numerous empirical formulae have been developped (cf Appendix B). Let θ be the angle (positive counter-clockwise) of the local streamline with respect to the x axis. We have :

$$\Gamma_{xx} = \Gamma_{\|} \cos^2\theta + \Gamma_{\perp} \sin^2\theta \tag{E.17}$$

$$\Gamma_{yy} = \Gamma_{\|} \sin^2\theta + \Gamma_{\perp} \cos^2\theta \tag{E.18}$$

$$\Gamma_{xy} = \Gamma_{yx} = \left(\Gamma_{\|} - \Gamma_{\perp}\right) \sin\theta \cos\theta \tag{E.19}$$

The advective terms and the diffusion operators corresponding to S or S' can be treated either simultaneously, either successively. In the latter case we follow a fractional step approach (cf section 3.6).

- Backwards characteristics methods (cf sec. 4.1.2) rely naturally on process splitting into advective and dispersive terms. Similarly, in Lagrangian polynomial fitting methods (cf sec. 4.1.1), the determination of the polynomial coefficients is based on consideration of pure transport.

  In summary, algorithms developped to deal with the non-conservative form of the transport equation (cf sec. 4.1) lend themselves naturally to the application of a fractional step approach.

- The fractional step method may also be applied to algorithms dealing with the conservative form of the transport equation. However, these algorithms boil down to relationships of the form

  $$\psi_i^{n+1} = \psi_i^n - \frac{1}{\Delta x_i}\left(F_{i+1/2} - F_{i-1/2}\right) \tag{E.20}$$

  and it is just as simple to include directly the dispersive contribution in quantities $F_{i+1/2}$ and $F_{i-1/2}$, which are the mass fluxes respectively at the left and right boundaries of the cell centred around node i.

  Following the Crank-Nicholson approximation (Roache, 1985), the dispersive flux reads simply (in one dimensional case) :

  $$F_{di+1/2} = \left(A\Gamma \right)_{i+1/2} \frac{C_{i+1} - C_i}{x_{i+1} - x_i} \tag{E.21}$$

  where $\left(A\Gamma \right)_{i+1/2}$ is an estimation of the product of the wet section by the dispersion coefficient at midpoint $x_{i+1/2}$.

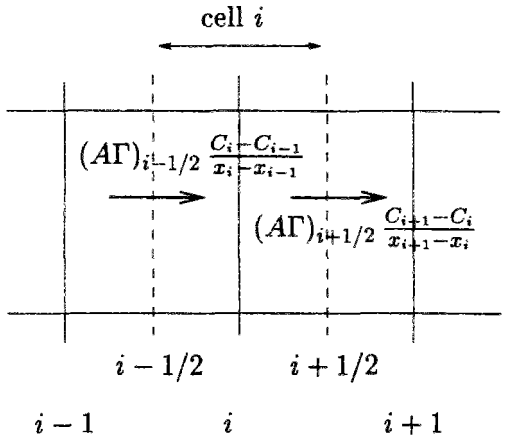


Figure E.5: Diffusive fluxes at cell i (1D case)

A and Γ depend on the flow field. When hydraulic variables and transported scalar are estimated on staggered grids, so that hydraulic variables are effectively computed at midnodes $x_{i+1/2}$, the estimation of $(A\Gamma)_{i+1/2}$ is straightforward. If they are computed on the same grid, the most common approximations to $(A\Gamma)_{i+1/2}$ are either

$$\frac{A_i + A_{i+1}}{2} \cdot \frac{\Gamma_i + \Gamma_{i+1}}{2} \quad \text{or} \quad \frac{1}{2}\left[A_i\Gamma_i + A_{i+1}\Gamma_{i+1} \right]$$

In eq. E.21, we have not indicated which time level is used for the computation of $F_{di+1/2}$. In fact, the $F_{di+1/2}$ may just as well be estimated explicitly or implicitly, i.e. as a function of C, A and Γ values at both time levels n and $n+1$. However, most algorithms introduced in section 4.2 are explicit, so that, in order to keep consistent with the evaluation of the advective fluxes and to preserve the simplicity of the scheme, it would be more natural to adopt an explicit evaluation of the diffusive fluxes. A plain explicit evaluation yields :

$$F_{di+1/2} = (A\Gamma)^n_{i+1/2} \frac{C^n_{i+1} - C^n_i}{x_{i+1} - x_i}$$

Yet, it is possible to improve this evaluation by taking somehow into account the downstream displacement of the concentration profile during the time step.
The upwind estimation of concentration at time level $n+1$ at node i after pure advection is $C_i^{\mathrm{ad}} = C_i - c_r (C_i - C_{i-1})$ (nb : case of positive velocity; c_r local Courant number). A fairer approximation to the diffusive flux is thus given by

$$F_{di+1/2} = (A\Gamma)^{n+1/2}_{i+1/2} \frac{C^*_{i+1} - C^*_i}{x_{i+1} - x_i}$$

where C_i^* is the arithmetic mean between C_i^n and C_i^{ad} (Leonard, 1979) and subscript $n+1/2$ denotes the arithmetic mean between $A\Gamma$ evaluations at time levels n and $n+1$.

An explicit treatment of the diffusive fluxes induces some stability requirement (Leonard, 1979; Holly & Usseglio-Polatera, 1984), namely

$$\forall i, \quad \frac{\Gamma_{i+1/2}\,\Delta t}{(x_{i+1}-x_i)^2} \leq \frac{1}{2} \tag{E.22}$$

We have finally chosen to apply the same numerical treatment to the dispersive terms, whatever the advective scheme may be : this implied to adopt a **fractional step approach**, *which is detailed hereafter.*

E.2.1 One-dimensional algorithm

From now on, let $C^{n+1/2}$ denote the values of the scalar concentration estimated after the advection process. The concentrations at next time level $n+1$ are obtained by solving at each node :

$$A_i^{n+1/2}\,\frac{C_i^{n+1}-C_i^{n+1/2}}{\Delta t} = S_i \tag{E.23}$$

The dispersion term S_i is expressed as :

$$S_i = \alpha \left[\frac{\partial}{\partial x}\left(A\Gamma \frac{\partial C}{\partial x} \right) \right]_i^{n+1} + (1-\alpha) \left[\frac{\partial}{\partial x}\left(A\Gamma \frac{\partial C}{\partial x} \right) \right]_i^{n+1/2} \tag{E.24}$$

where α is a an implicitation parameter ($0 \leq \alpha \leq 1$) and

$$\left[\frac{\partial}{\partial x}\left(A\Gamma \frac{\partial C}{\partial x} \right) \right]_i = \frac{2}{x_{i+1}-x_{i-1}} \left[(A\Gamma)_{i+\frac{1}{2}} \frac{C_{i+1}-C_i}{x_{i+1}-x_i} - (A\Gamma)_{i-\frac{1}{2}} \frac{C_i - C_{i-1}}{x_i - x_{i-1}} \right] \tag{E.25}$$

By developping E.23 at each node with the help of relations E.24 and E.25, we find that

$$[T_1]\,[C]^{n+1} = [T_2]\,[C]^{n+1/2}$$

where $[C]^{n+1/2}$ and $[C]^{n+1}$ denote the node values vectors corresponding respectively to advected concentrations and concentrations at time level $n+1$, and $[T_1]$ and $[T_2]$ are both tridiagonal matrices. The dimension of these vectors and matrices is equal to the total number of grid points in the domain. At open boundaries, the diffusion is usually neglected so that, on first and last lines of $[T_1]$ and $[T_2]$, the diagonal term is set to 1 and all other components are null. If the upstream or downstream boundary is a solid one, the components values reflect a condition of pure reflection of the concentration.

The tridiagonal system is easily solved with a classical double-sweep algorithm (cf Thomas algorithm in sec. D.1). In case of uniform hydraulics, the above scheme is unconditionally stable if $\alpha \geq 1/2$. Otherwise, the space and time steps need to satisfy criterion E.22.

E.2.2 Two-dimensional algorithm

The two dimensional numerical treatment of the dispersive terms combines process and space splitting (cf 3.6) and is borrowed from (Holly & Usseglio-Polatera, 1984).

The solution proceeds in two stages, with the computation of an intermediate variable C^* :

$$h^{n+1/2}\frac{C^* - C^{n+1/2}}{\Delta t} = \alpha\frac{\partial}{\partial x}\left(h\Gamma_{x,x}\frac{\partial C^*}{\partial x}\right) + (1-\alpha)\frac{\partial}{\partial x}\left(h\Gamma_{x,x}\frac{\partial C^{n+1/2}}{\partial x}\right) + \frac{\partial}{\partial x}\left(h\Gamma_{x,y}\frac{\partial C^{n+1/2}}{\partial y}\right) \quad \text{(E.26)}$$

$$h^{n+1/2}\frac{C^{n+1} - C^*}{\Delta t} = \alpha\frac{\partial}{\partial y}\left(h\Gamma_{y,y}\frac{\partial C^{n+1}}{\partial y}\right) + (1-\alpha)\frac{\partial}{\partial y}\left(h\Gamma_{y,y}\frac{\partial C^*}{\partial y}\right) + \frac{\partial}{\partial y}\left(h\Gamma_{y,x}\frac{\partial C^*}{\partial x}\right) \quad \text{(E.27)}$$

The first stage accounts for $x-$ diffusion of the advected values, the second stage adds $y-$diffusion. The first stage is solved independently along each line $y =$ constant, the second stage along lines $x =$ constant. In order to allow for this space splitting, terms involving cross-derivatives have to be written explicitly.

Let us assume that node (i,j) has coordinates (x_i, y_j). Terms of the form $\frac{\partial}{\partial x}\left(h\Gamma\frac{\partial f}{\partial x}\right)$ at node (i,j) are developped as :

$$\frac{2}{x_{i+1} - x_{i-1}}\left[(h\Gamma)_{i+\frac{1}{2},j}\frac{f_{i+1,j} - f_{i,j}}{x_{i+1} - x_i} - (h\Gamma)_{i-\frac{1}{2},j}\frac{f_{i,j} - f_{i-1,j}}{x_i - x_{i-1}} \right]$$

with

$$(h\Gamma)_{i+\frac{1}{2},j} = \frac{1}{2}\left[h_{i,j}\Gamma_{i,j} + h_{i+1,j}\Gamma_{i+1,j} \right]$$

A similar discretization is applied to terms $\frac{\partial}{\partial y}\left(h\Gamma\frac{\partial f}{\partial y}\right)$.

The cross-derivative terms $\frac{\partial}{\partial x}\left(h\Gamma\frac{\partial f}{\partial y}\right)$ are estimated as

$$\frac{2}{x_{i+1} - x_{i-1}}\left[(h\Gamma)_{i+\frac{1}{2},j}\left(\frac{\partial f}{\partial y}\right)_{i+1/2,j} - (h\Gamma)_{i-\frac{1}{2},j}\left(\frac{\partial f}{\partial y}\right)_{i-1/2,j} \right]$$

with

$$\left(\frac{\partial f}{\partial y}\right)_{i+1/2,j} = \frac{1}{2}\left[\frac{f_{i,j+1} - f_{i,j-1}}{y_{j+1} - y_{j-1}} + \frac{f_{i+1,j+1} - f_{i+1,j-1}}{y_{j+1} - y_{j-1}} \right]$$

The above centred estimation of $\frac{\partial f}{\partial y}_{i+1/2,j}$ is second-order accurate only if the grid spacing is uniform. Whenever the cross-diffusion is important, it can easily be modified in order to retain always second-order spatial accuracy (cf eq. 3.11 in section 3.2.2).

This kind of discretization leads to solve for each stage E.26 and E.27 a set of monodimensional, tridiagonal, linear systems. The treatment of boundaries conditions proceeds as indicated for the one-dimensional case. As the cross-derivative terms are estimated explicitly, the usual unconditional stability obtained with $\alpha \geq 1/2$ cannot be guaranteed.

Appendix F

Supplementary results - advection/diffusion tests

F.1 Comparison of different flux form advection schemes

F.1.1 Presentation of the schemes

We have been investigating the performance of different flux-form advection schemes based on a polynomial approximation of the scalar field. The methodology of these algorithms is indicated in 4.2.4. We restrict our analysis to schemes using third or fourth-order polynomials. From previous work (Simon, 1990b), it stems that the more efficient scheme relying on a second-order polynomial approximation is the QUICKEST one (cf sec. 4.1.1 and 4.2.4).

The approximation polynomial of scalar ψ on cell i of length Δx_i reads :

$$\phi_i(x) = \sum_k a_{i,k}\,\xi^k \qquad \text{with} \quad \xi = \frac{x - x_i}{\Delta x_i}$$

It may be built following different approaches :

1. **polynomial fitting**
   The $(a_{i,k})$ are chosen so that

   $$\phi_i\,(x_{i+l}) = \psi_{i+l} \quad \text{for} \quad l = 0, \pm 1, \pm 2 \ldots \text{ according to the desired order of approximation} \tag{F.1}$$

2. **Taylor series expansion**

ϕ_i corresponds to a Taylor series expansion of ψ in the vicinity of node i

$$\phi_i(x) \;=\; \sum_k a_{i,k}\,\xi^k \;=\; \sum_k \frac{(x-x_i)^k}{k!}\,\frac{\partial^k \psi}{\partial x^k}_{|i}$$

so that

$$a_{i,k} \;=\; \frac{(\Delta x_i)^k}{k!}\,\frac{\partial^k \psi}{\partial x^k}_{|i} \tag{F.2}$$

The partial derivatives at node i are estimated by finite differences approximations involving the node neighbours. In order for the expansion to be m-order accurate (i.e. all truncation errors of order up to and including m are null), the truncation error in the estimation of the k-order derivative must have a leading term of order $m+1-k$ at least.

It can be checked that, in order to estimate the needed derivatives with the adequate truncation error, the Taylor Series expansion of order $2p-1$ and $2p$ need both to use nodes $i-p$ to $i+p$. For Taylor Series expansion of order $2p-1$ it is one node mode than required for building a same order polynomial by discrete data fitting.

Even-order Taylor Series expansion are found to coincide with their counterparts based on discrete data fitting; odd-order Taylor Series do not coincide.

3. **Area preserving polynomials** are such that they satisfy

$$\int_{x_{i+l-\frac{1}{2}}}^{x_{i+l+\frac{1}{2}}} \phi_i(x)\,dx \;=\; \psi_{i+l}\,\Delta x_l \qquad \text{for} \quad l \;=\; 0,\,\pm 1,\,\pm 2\ldots \tag{F.3}$$

i.e. represent the same mass on each cell $i+l$ than does the discrete scalar field.

The algorithms derived from approach (1) are termed "integrated flux-form schemes" in the terminology of (Tremback *et al.*, 1987). The corresponding expressions of fluxes $F_{i+\frac{1}{2}}$ for orders up to 10 are given in (Tremback *et al.*, 1987) in case of uniform grid spacing. The polynomial approximations themselves are mentionned in (Bott, 1989a) up to order 4.

The algorithms derived from approach (3) are termed "constant grid flux form" in (Tremback *et al.*, 1987) where the corresponding fluxes (up to order 10) are presented too. Yet, the introduction to this approach delivered in (Bott, 1989b) is far more general than it is in (Tremback *et al.*, 1987). In (Bott, 1989b), we find the derivation of the polynomial approximations for orders 2 and 4.

The possible expressions for cubic Taylor Series expansion (approach (2)) are indicated in (Simon, 1990b) both for uniform and variable grid spacing.

We have retained only approximations which allow to estimate the advected value ψ_i^{n+1}, computed as

$$\psi_i^{n+1} \;=\; \psi_i^n \;-\; \frac{1}{\Delta x_i}\left(F_{i+1/2} - F_{i-1/2}\right)$$

with the help of 6 neighbouring grid nodes at the very most. Consequently, each ϕ_i has to be build from ψ values no further downstream nor upstream than nodes $i-2$ and $i+2$. We indicate hereafter in tables F.1 and F.2 the acceptable polynomials, with Δx constant and U positive. (nb : the coefficients of the cubic Taylor Series expansion (table F.1) derive from approximations to the first, second and third-order derivatives with a leading truncation error of orders 4, 2 and 2 respectively. A fourth-order approximation could be used for $a_{i,2}$: the resulting expression is $a_{i,2}$ in table F.2, first column. $a_{i,4}$ given in table F.2 is an approximation to $\frac{1}{24}.\Delta x^4.\frac{\partial^4 \psi}{\partial x^4}$ with a 6^{th}-order leading truncation term.)

Table F.1: Third-order approximation polynomials

	Polynomial fitting	Taylor Series Expansion
$a_{i,0}$	ψ_i	ψ_i
$a_{i,1}$	$\frac{-2\psi_{i-1} - 3\psi_i + 6\psi_{i+1} - \psi_{i+2}}{6}$	$\frac{\psi_{i-2} - 8\psi_{i-1} + 8\psi_{i+1} - \psi_{i+2}}{12}$
$a_{i,2}$	$\frac{\psi_{i-1} - 2\psi_i + \psi_{i+1}}{2}$	$\frac{\psi_{i-1} - 2\psi_i + \psi_{i+1}}{2}$
$a_{i,3}$	$\frac{-\psi_{i-1} + 3\psi_i - 3\psi_{i+1} + \psi_{i+2}}{6}$	$\frac{-\psi_{i-2} + 2\psi_{i-1} - 2\psi_{i+1} + \psi_{i+2}}{12}$

Table F.2: Fourth-order approximation polynomials

	Polynomial fitting = Taylor Series exp.	Area-Preserving polynomial
$a_{i,0}$	ψ_i	$\frac{9\psi_{i-2} - 116\psi_{i-1} + 2134\psi_i - 116\psi_{i+1} + 9\psi_{i+2}}{1920}$
$a_{i,1}$	$\frac{\psi_{i-2} - 8\psi_{i-1} + 8\psi_{i+1} - \psi_{i+2}}{12}$	$\frac{5\psi_{i-2} - 34\psi_{i-1} + 34\psi_{i+1} - 5\psi_{i+2}}{48}$
$a_{i,2}$	$\frac{-\psi_{i-2} + 16\psi_{i-1} - 30\psi_i + 16\psi_{i+1} - \psi_{i+2}}{24}$	$\frac{-\psi_{i-2} + 12\psi_{i-1} - 22\psi_i + 12\psi_{i+1} - \psi_{i+2}}{16}$
$a_{i,3}$	$\frac{-\psi_{i-2} + 2\psi_{i-1} - 2\psi_{i+1} + \psi_{i+2}}{12}$	$\frac{-\psi_{i-2} + 2\psi_{i-1} - 2\psi_{i+1} + \psi_{i+2}}{12}$
$a_{i,4}$	$\frac{\psi_{i-2} - 4\psi_{i-1} + 6\psi_i - 4\psi_{i+1} + \psi_{i+2}}{24}$	$\frac{\psi_{i-2} - 4\psi_{i-1} + 6\psi_i - 4\psi_{i+1} + \psi_{i+2}}{24}$

For uneven grid spacings, the coefficient expressions are far more complicated. As regards approach (1) and (3), the $(a_{i,k})$ are obtained by solving a linear system of $m+1$ equations (m being the order of the polynomial).
Let us assume this polynomial is built with the help of nodes $i+l$ with $l_{\min} \leq l \leq l_{\max}$. ξ_l and $\xi_{l+1/2}$ denote respectively $(x_{i+l} - x_i)/\Delta x_i$ and $\left(x_{i+l+1/2} - x_i\right)/\Delta x_i$, $x_{i+l+1/2}$ being the midpoint between nodes $i+l$ and $i+l+1$. Let $\underline{C}$ be the vector made of scalar values ψ_{i+l}, $l = l_{\min}, \ldots l_{\max}$ and $\underline{A}$ the vector made of coefficients $a_{i,0}$ to $a_{i,m}$. The polynomial developped according to approach (1) is obtained by solving $\underline{M}.\underline{A} = \underline{C}$ where the elements on row number k of $\underline{M}$ are the ξ_l^k, $l = l_{\min}, \ldots l_{\max}$. For approach (3), the $\underline{M}$- matrix coefficients are more complex; they read :

$$\frac{\Delta x_{i+l}}{\Delta x_i} \frac{\xi_{l+1/2}^{k+1} - \xi_{l-1/2}^{k+1}}{k+1}$$

F.1.2 Performance evaluation

The detailed Fourier analysis of schemes derived according approaches (1) and (3) may be found in (Tremback *et al.*, 1987; Bott, 1989a; Bott, 1989b).

We illustrate hereafter the performances of the different algorithms with respect to one of the monodimensional tests applied in chapter 6 : the transport of a narrow gaussian in an uniform velocity field (cf section 6.3.1).

- The gaussian distribution is initially centred in x =2000 and is transported during 10800s with velocity 0.5 m/s. The initial standard deviation is σ_0 = 264m, so that the initial source length is 8 Δx (Δx being constant and equal to 200m).
- Simulations are made for 4 different diffusivities (0, 2, 5, 50 $m^2.s^{-1}$) corresponding to Peclet numbers $Pe = U\Delta x/\Gamma = \infty$, 50, 20, 2 and eight different values for the time step : the resulting Courant numbers range from 0.12 to 1 and the total number of time levels between the start and the end of computation ranges between 225 and 27.
- The definitions of the error measures related to this test are given in sec. 6.2.

The schemes have been applied with and without the positiveness-enforcing limiter suggested by Bott (Bott, 1989a; Bott, 1989b) (cf section 4.2.5). The results of this test are representative of the overall ranking of the schemes. Thus, discussion relative to other tests has been omitted for the sake of brevity.

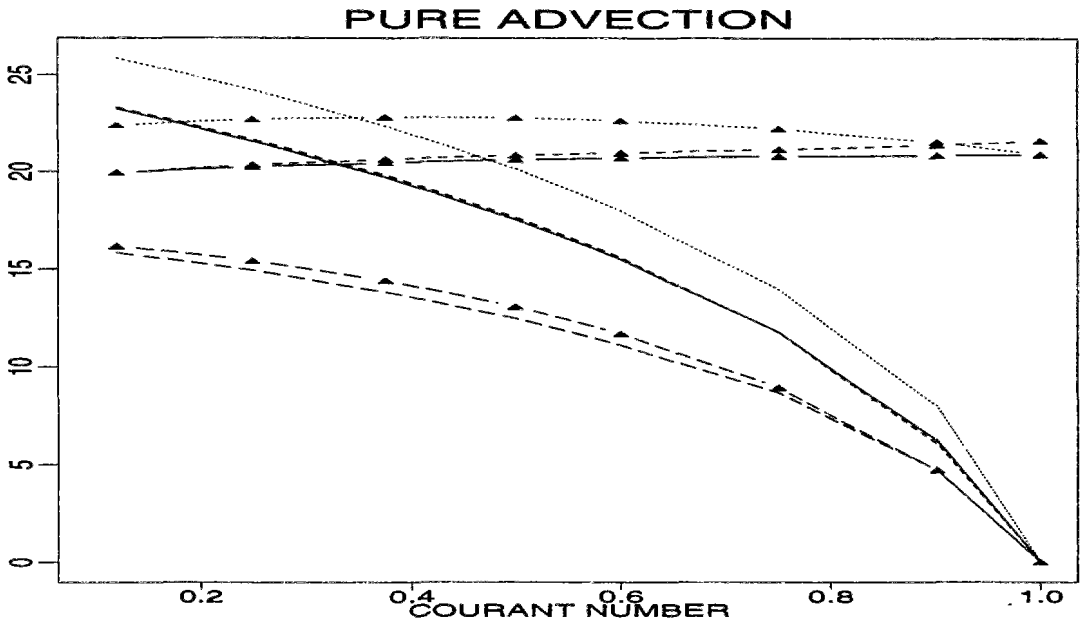


Figure F.1: Damping : relative error on peak value (%)

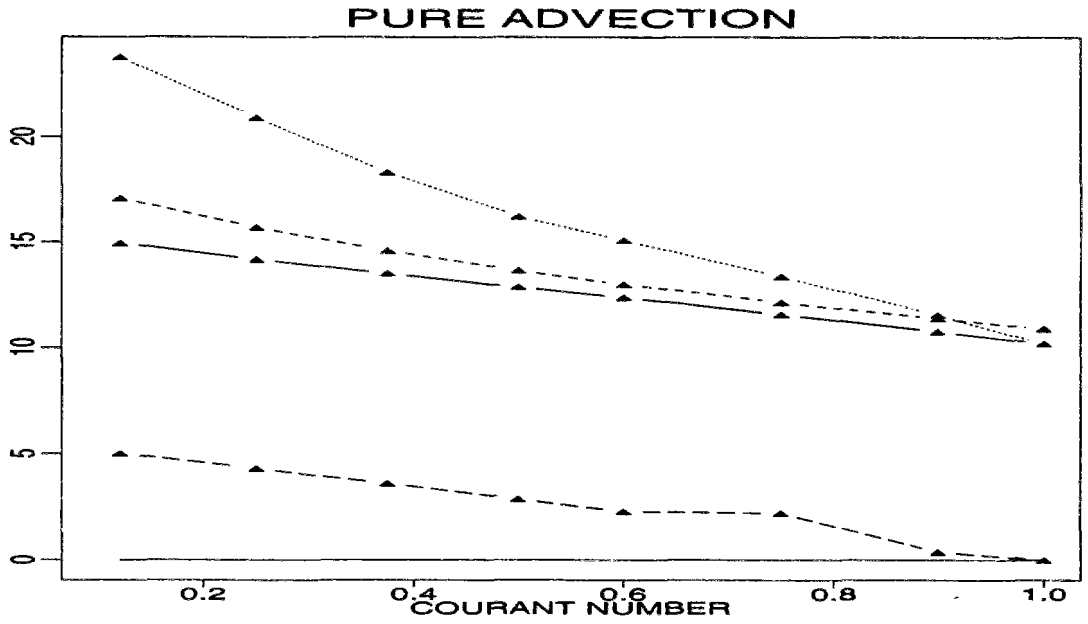


Figure F.2: Maximum negative concentration (in % of peak value)

Figures F.1 to F.5 illustrate the evolution of the error measures according to the Courant number in case of pure advection ($\Gamma = 0$). The legend of these plots is the following :

- Solid lines refer to results of the third-order Taylor Series expansion.
  Dotted lines refer to results with the cubic obtained by discrete data fitting.
  Short-dashed lines refer to results with the fourth-order polynomial corresponding both to data fitting and Taylor Series expansion.
  Long-dashed lines refer to results of the fourth-order area-preserving polynomial.
- Lines with marks denote the unlimited versions of the algorithms, lines without marks their limited versions.

The following comments can be made :

1. **Limited versions are far superior to their unlimited counterparts.**
   Notably, the limiter ensures that an exact solution is obtained for $Cr = 1$ (exact displacement of one mesh size during each time step).

   The only unlimited version which performs correctly is the one corresponding to the fourth-order area-preserving polynomial. This is because for that particular scheme the limited and unlimited versions are in fact quite close. Indeed, the limiter has two stages (cf sec. 4.2.5) : cancelation of negative fluxes, normalization in order for the approximation to be

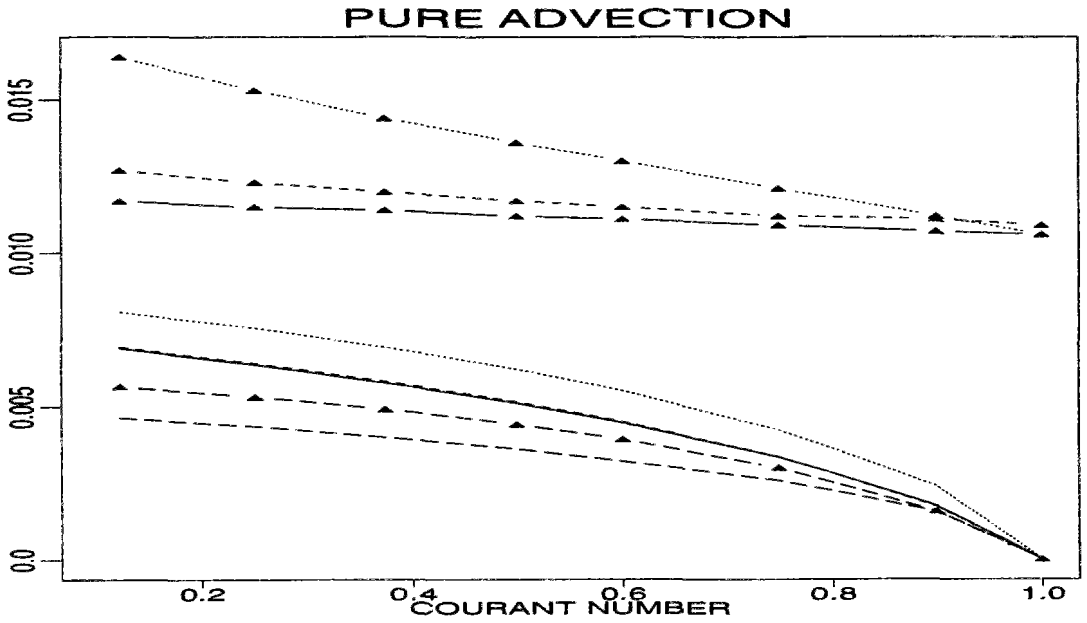


Figure F.3: L2 Error Norm (normalized by total mass)

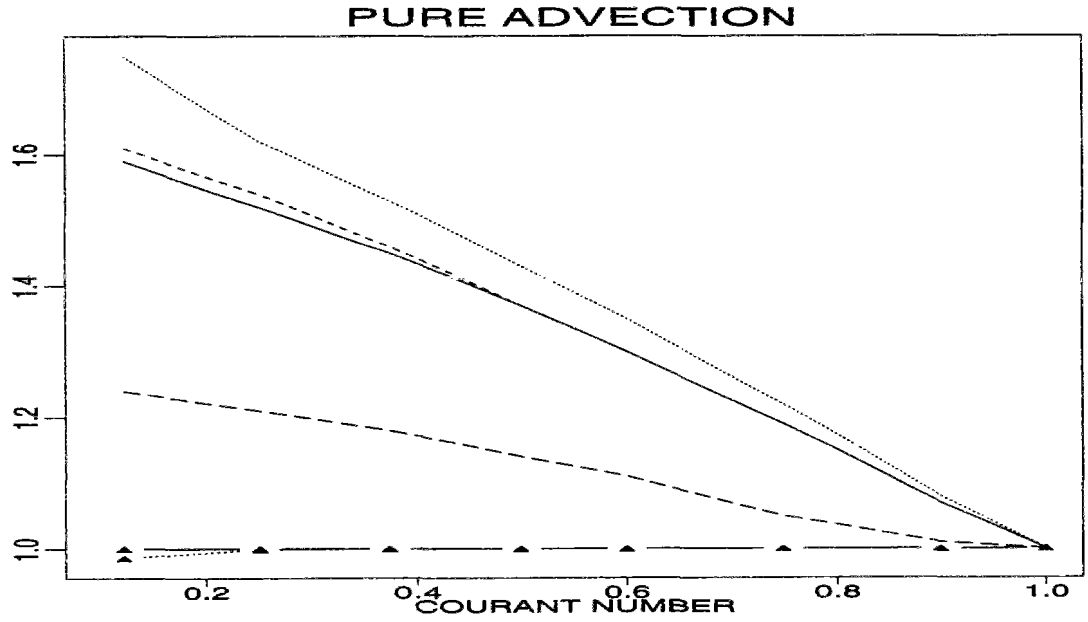


Figure F.4: Numerical spreading : ratio of numerical vs. exact variance

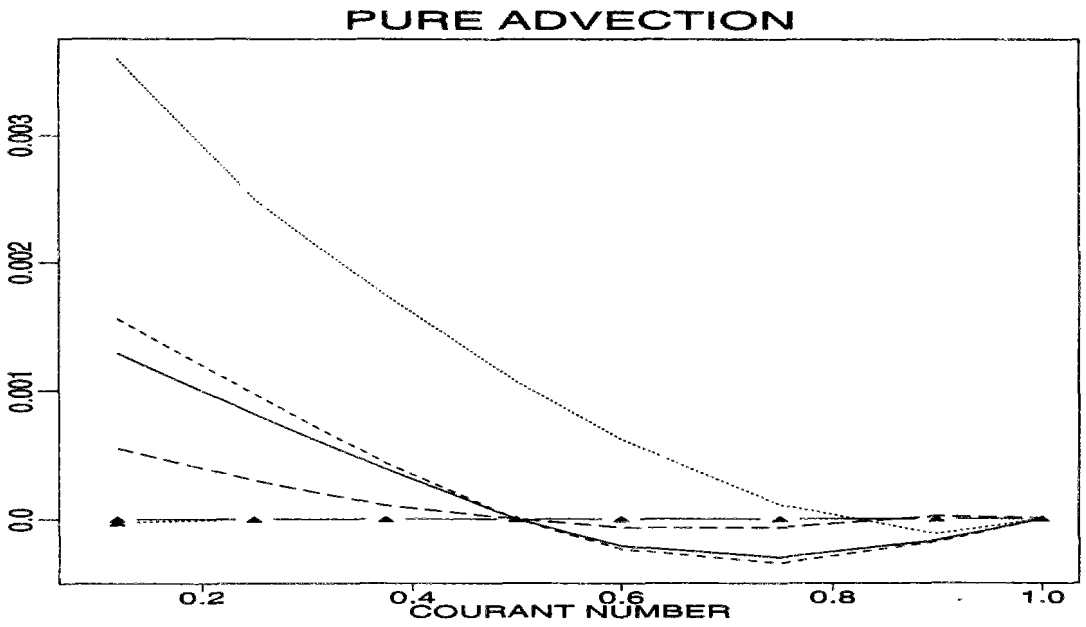


Figure F.5: Global phase shift (normalized by exact travel distance)

mass preserving. By definition, the area-preserving polynomials need no normalization : the only difference between the unlimited and limited versions is that, in the first one, computed negative fluxes are not systematically set to zero.

The most costly stage of the limiter is the normalization, which implies (except for area-preserving polynomials) the computation of one supplementary flux integral. On the other hand, canceling negative values requires but a limited amount of computational effort. Hence, while unlimited versions of schemes based on polynomial fitting or Taylor Series expansion are about twice as fast as their limited counterparts, the cost of both area-preserving versions is nearly the same, the limited version raising better results. Consequently, we may state that the unlimited version of the area-preserving scheme is rather irrelevant and we shall not make any further comment on it.

2. **Some error measures are deceiptive.**
They indicate that there is no difference between the exact and computed center of mass, nor between the exact and computed variance for unlimited versions of the schemes (see respectively figures F.5 and F.4). However, these measures are *integral* error measures : it turns out that the strong negative concentations generated by unlimited schemes help to cancel out part of the integrals. A look at the computed pollutograms allows to check that they are indeed plagued both by dissipation and dispersion errors (see figures F.6 to F.9).

An other measure of the phase shift is the error on the peak location. While limited versions

forecast correctly the peak position, unlimited versions (except the area-preserving one) misplace it : they locate it systematically one cell upstream. This result confirms that the integral measure of phase shift need to be manipulated and interpreted with caution !

3. **With limited versions of the algorithms, damping** (fig F.1), **error norm** (fig F.3), **numerical spreading** (fig F.4) **decrease as the Courant number increases**, i.e. as the number of iterations performed between the start and the end of computation decreases.

   The errors generated by the unlimited versions also tend to decrease with the number of iterations, except the peak damping, which remains quite stable. Yet, they decrease much more slowly and keep superior to the errors of their limited counterparts (e.g. for schemes based on polynomial fitting and Taylor Series expansion, the L2 error norm of unlimited versions is about three times as big than for the corresponding limited algorithms).

4. **The behaviour of the phase shift is slightly more complicated.**
   The scheme based on cubic polynomial fitting has a predominantly lagging phase error. Other schemes have a lagging phase error for $c_r \leq 0.5$, then a leading phase error for bigger Courant numbers.

   However, it must be pointed out that, in any case, the error is quite small : less than 0.3 % of the total travel distance.

5. As regards overall performance, **limited schemes rank as follows :**

   Area-preserving form > 3^{rd}-order Taylor Series exp. > 4^{th}-order one > 3^{rd}-order fitting

   The results obtained by the third-order Taylor Series expansion and the 4^{th}-order one are quite close : increasing the order of the approximation is not very effective, at the contrary !

   This grading is confirmed by the tests performed with other diffusivities, as can be deduced from tables F.3 to F.6.

   - In these tables we have merely been indicating the lower and upper bounds of the error mesaures : they display the same dependency with respect to the Courant number value than for pure advection.
   - Results for the bigger Γ value ($Pe = 2$) are not given : all schemes perform perfectly for this diffusion-dominated situation.
   - A slight overestimation of the peak value can be observed in table F.5 for $Pe = 20, 50$. This can be improved by a better resolution of the diffusion operator. We have been working with an implicitation parameter of 0.6 (cf sec. E.2.1) : lowering it closer to the optimal value 0.5 reduces the overshoot.

In conclusion, considering that this test is a severe one, **the performances of the algorithms based on third-order Taylor Series expansion and above all on the fourth-order area-preserving polynomial appear to be quite encouraging.** We have thus decided to extend our testing of these two schemes.

Why haven't we restricted our study to the best scheme only ? This is because we have been considering an other indicator of the schemes performance : their computational cost.

- When the grid spacing is uniform, the computation of the polynomial coefficients is straightforward and the most time consuming stage of the computation lies in the evaluation of the flux integrals. As mentionned above, with area-preserving forms, only one flux integral needs to be estimated, instead of two for other methods. Consequently, the area-preserving forms yield cheaper algorithms, even when based on higher-order polynomial approximations : e.g. the scheme based on 3^{rd}-order Taylor Series expansion requires 30 % more CPU time than with the 4^{th}-order area-preserving polynomial.

- However, preliminary tests with uneven grid spacings (Simon, 1990b) taught us that the cost of area-preserving forms increases dramatically in such situations (by a factor 3 approximately), while the generalization of the 3^{rd}-order Taylor Series to uneven grids is just slightly more costly (a mere 5 % !).

  The sharp increase has to be ascribed to the stage dealing with polynomial coefficients evaluation, which requires now the solution of a (5×5) linear system at each grid node and for each time step. This stage remains costly, even when, in order to spare time, the corresponding matrices are assembled and inversed once, at the beginning of the computation.

  (nb : all this is discussed more thoroughly in section 6.6.)

Table F.3: L2 error norm (in % of exact mass)

Scheme	$Pe = \infty$		$Pe = 50$		$Pe = 20$	
	min	max	min	max	min	max
	Limited versions					
Taylor Series	0	0.69	0.04	0.35	0.04	0.16
Cubic fitting	0	0.81	0.04	0.42	0.04	0.20
4^{th}-order fitting	0	0.69	0.04	0.35	0.04	0.16
Area-Preserving	0	0.46	0.04	0.20	0.04	0.07
	Un-limited versions					
Taylor Series	1.06	1.17	0.61	0.68	0.33	0.36
Cubic fitting	1.06	1.64	0.61	0.98	0.33	0.51
4^{th}-order fitting	1.09	1.27	0.64	0.74	0.34	0.40
Area-Preserving	0	0.57	0.04	0.23	0.04	0.07

Table F.4: Ratio of numerical to exact variance

Scheme	$Pe = \infty$		$Pe = 50$		$Pe = 20$	
	min	max	min	max	min	max
	Limited versions					
Taylor Series	1	1.59	1	1.29	1	1.14
Cubic fitting	1	1.75	1	1.40	1	1.17
4^{th}-order fitting	1	1.61	1	1.30	1	1.14
Area-Preserving	1	1.24	1	1.08	1	1.01

Table F.5: Relative error on peak value (in %)

Scheme	$Pe = \infty$		$Pe = 50$		$Pe = 20$	
	min	max	min	max	min	max
	Limited versions					
Taylor Series	0	23.26	-1.8	13.72	-1.9	7.09
Cubic fitting	0	25.86	-1.8	15.18	-1.9	7.51
4^{th}-order fitting	0	23.32	-1.8	13.73	-1.9	7.08
Area-Preserving	0	15.83	-1.8	7.53	-1.9	2.51
	Un-limited versions					
Taylor Series	19.95	20.91	11.33	12.83	6.03	7.37
Cubic fitting	20.91	22.42	12.33	13.39	5.81	7.60
4^{th}-order fitting	19.93	21.64	11.12	13.26	5.73	7.78
Area-Preserving	0	16.17	-1.8	7.59	-1.9	2.51

(NB : In the following figures (F.6 to F.9), we have been plotting for 4 different values of the Courant number c_r the concentration fields computed by the algorithms based respectively on the 4^{th}-order area-preserving polynomial (dashed lines) and the 3^{rd}-order Taylor Series expansion (solid lines). The black triangle marks denote the unlimited versions of the algorithms. The exact solution passes through the crosses.)

Table F.6: Maximum Negative Concentration (in % of peak value)

Scheme	$Pe = \infty$		$Pe = 50$		$Pe = 20$	
	min	max	min	max	min	max
	Un-limited versions					
Taylor Series	10.22	14.95	6.50	9.57	2.97	4.82
Cubic fitting	10.22	23.79	6.50	19.02	2.97	10.28
4^{th}-order fitting	10.92	17.07	7.16	11.79	3.21	5.80
Area-Preserving	0	5.02	0	2.06	0	0.05

F.2 Influence of derivative initialization on Holly-Preissmann method

As the Holly-Preissmann algorithm (cf 4.1.2.1) considers both the scalar concentration and its first-order derivative as dependent variables, it is necessary to initialize this derivative before running the scheme. Most often, the initial scalar field only is known. The initial derivatives can be computed by the application of any derivative estimate. In order to assess the sensitivity of HOLLY with respect to derivative initialization, we have been comparing the scheme outcomes when the initial derivatives are set to zero and when they are evaluated with the Akima estimate (cf 4.1.2.1). Two pure advection tests were performed : the first deals with the transport of a small gauss-hill in uniform flow (cf sec. 6.3.1), the second one with the propagation of a steep front (cf sections 6.5.1 and F.3). The error measures related to these tests are illustrated in figures F.10 to F.14, with the following legend :

- solid lines refer to the gauss-hill test results;
- dashed lines refer to the steep front test results;
- marks refer to results with initially null derivatives.

Results obtained with the Akima estimate are slightly better but, on the whole, **HOLLY appears to be rather insensitive to the derivative initialization.**

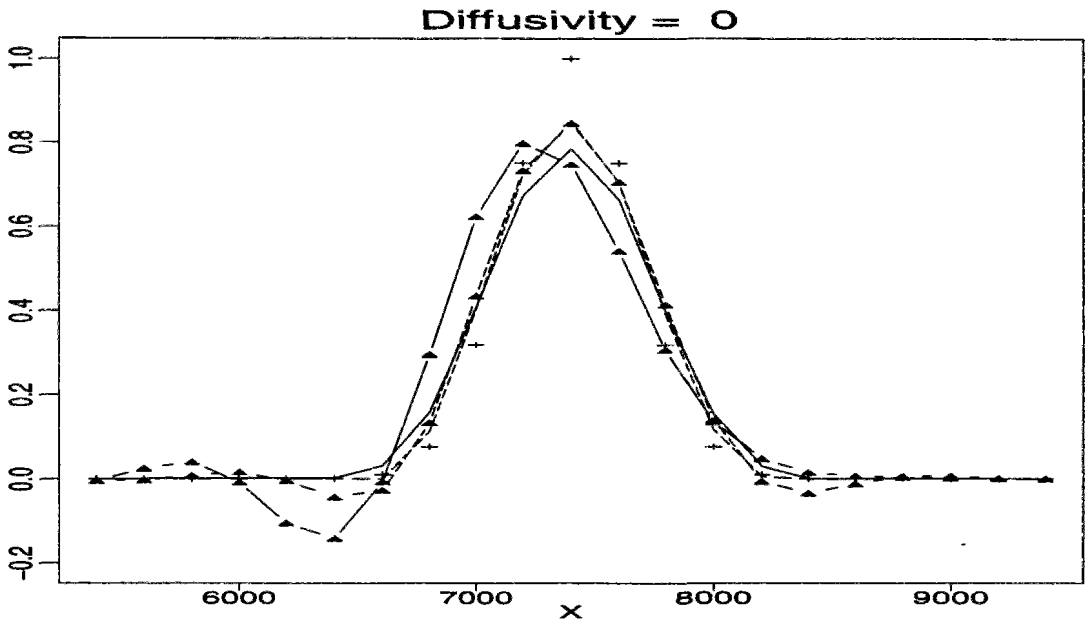


Figure F.6: Comparison of limited and unlimited algorithms : $cr = 0.25$

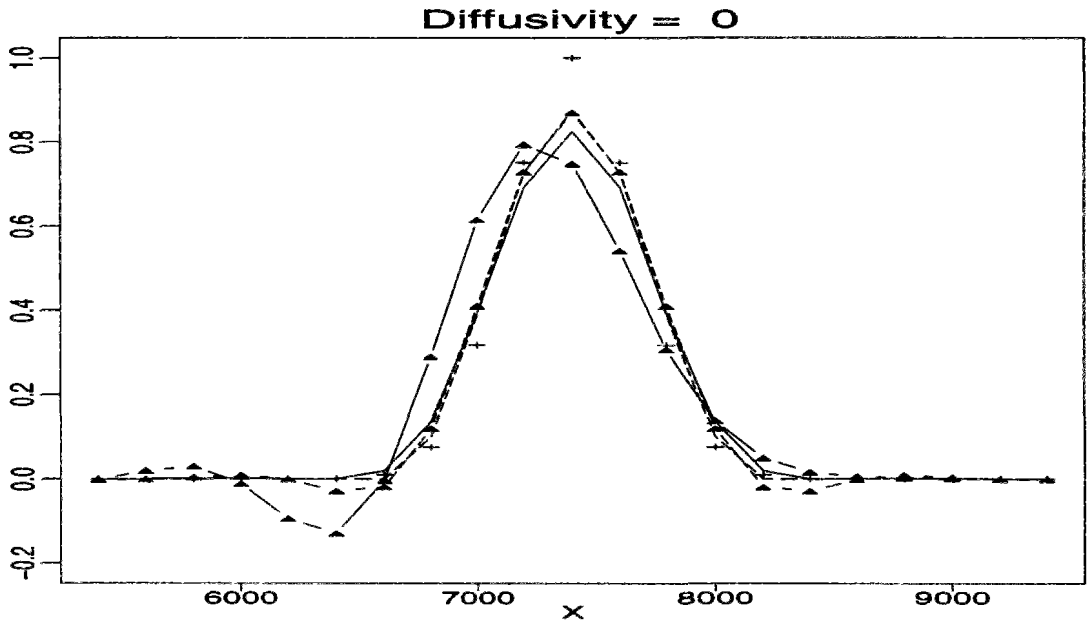


Figure F.7: Comparison of limited and unlimited algorithms : $cr = 0.50$

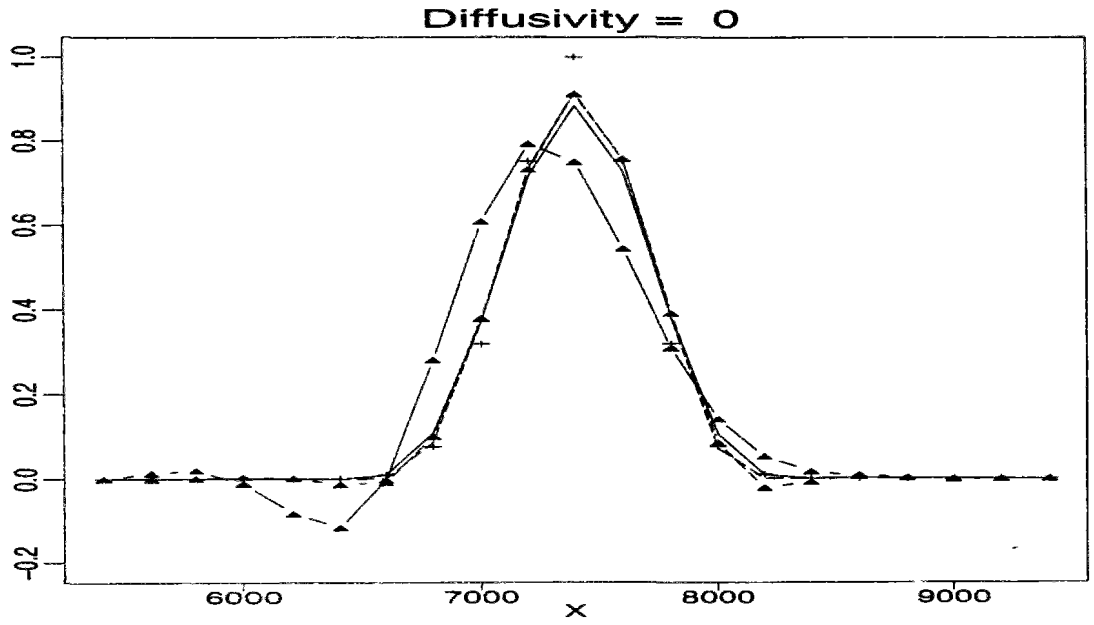


Figure F.8: Comparison of limited and unlimited algorithms : $cr = 0.75$

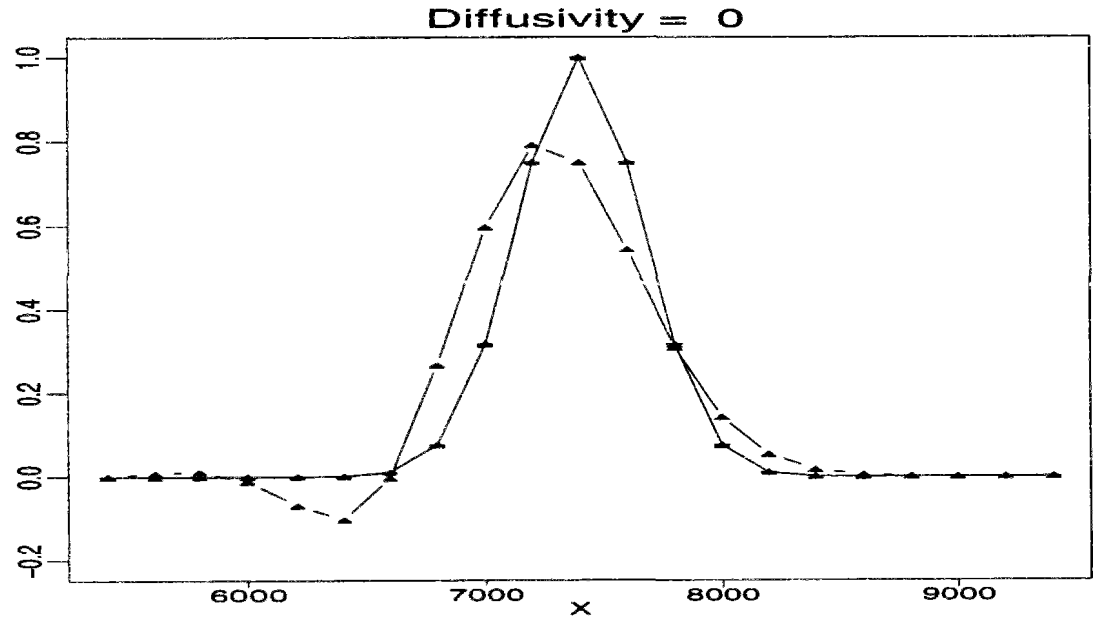


Figure F.9: Comparison of limited and unlimited algorithms : $cr = 1.00$

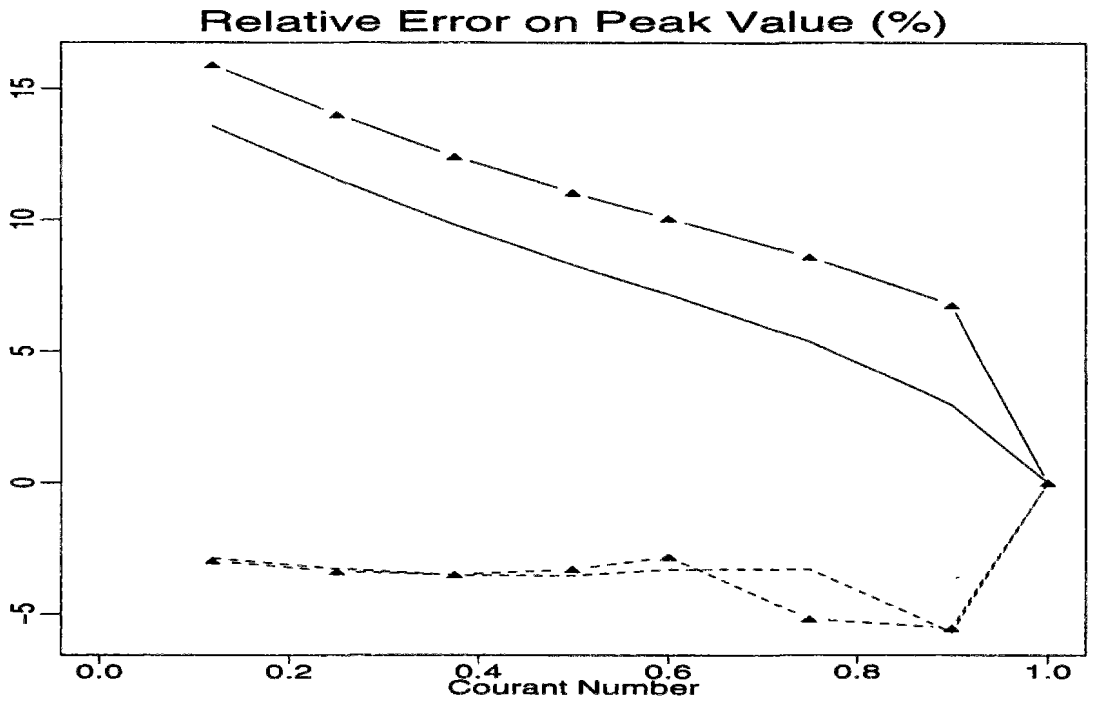


Figure F.10: Influence of derivative initialization on HOLLY : Damping & Overshoots

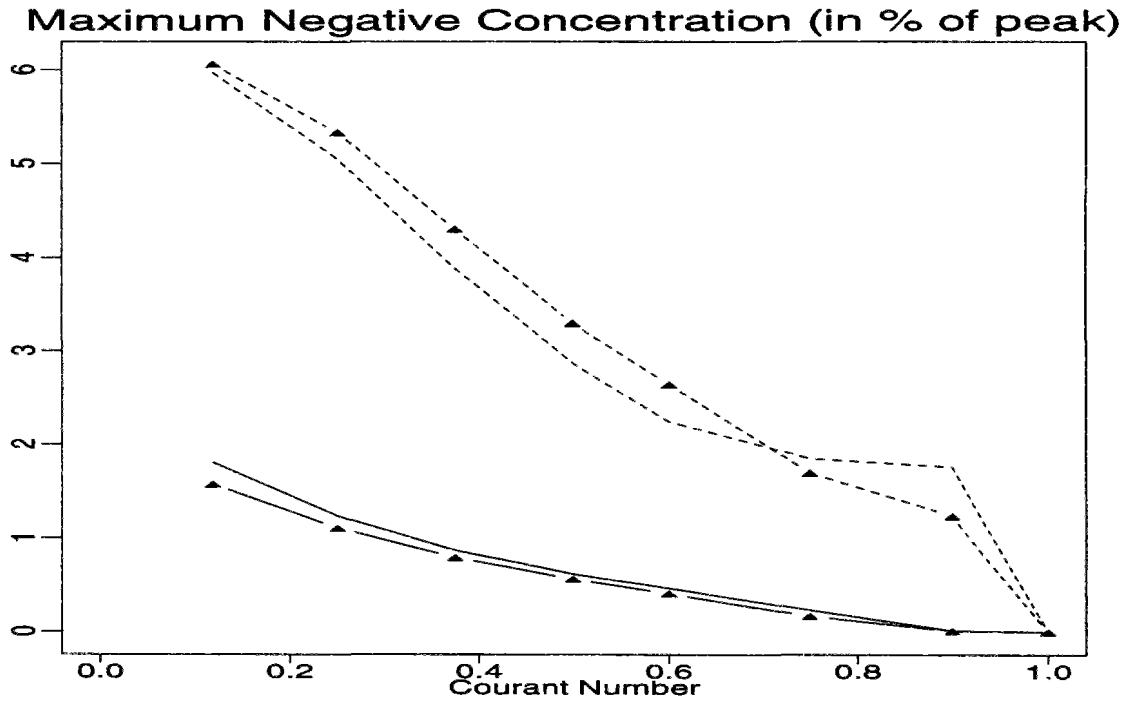


Figure F.11: Influence of derivative initialization on HOLLY : Undershoots

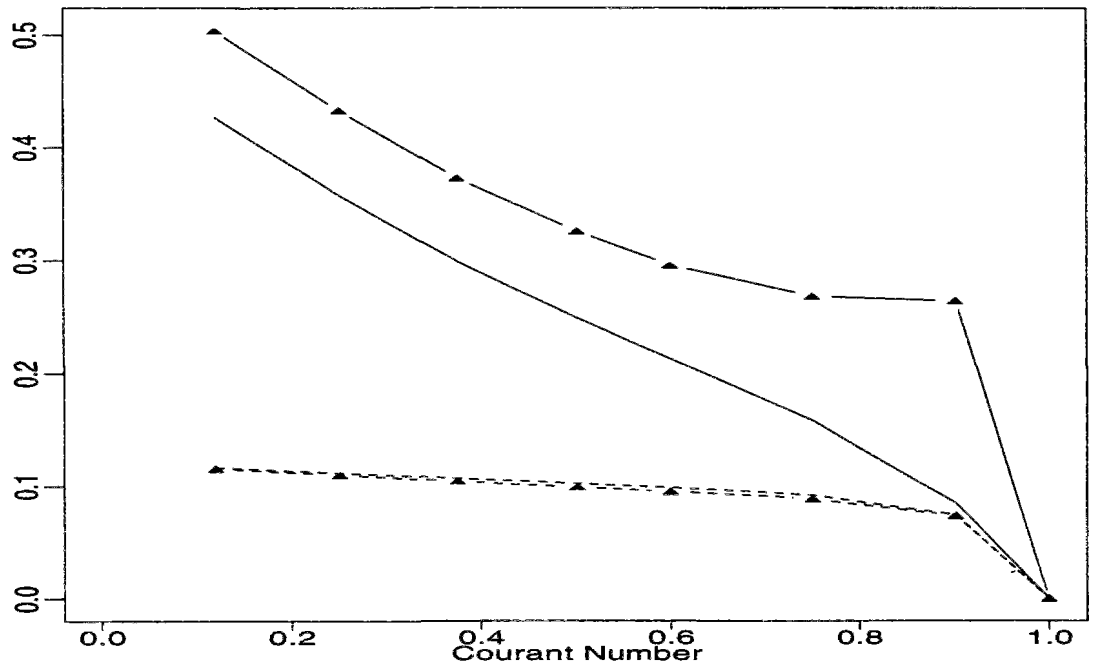


Figure F.12: Influence of derivative initialization on HOLLY : normalized L2 Error norm

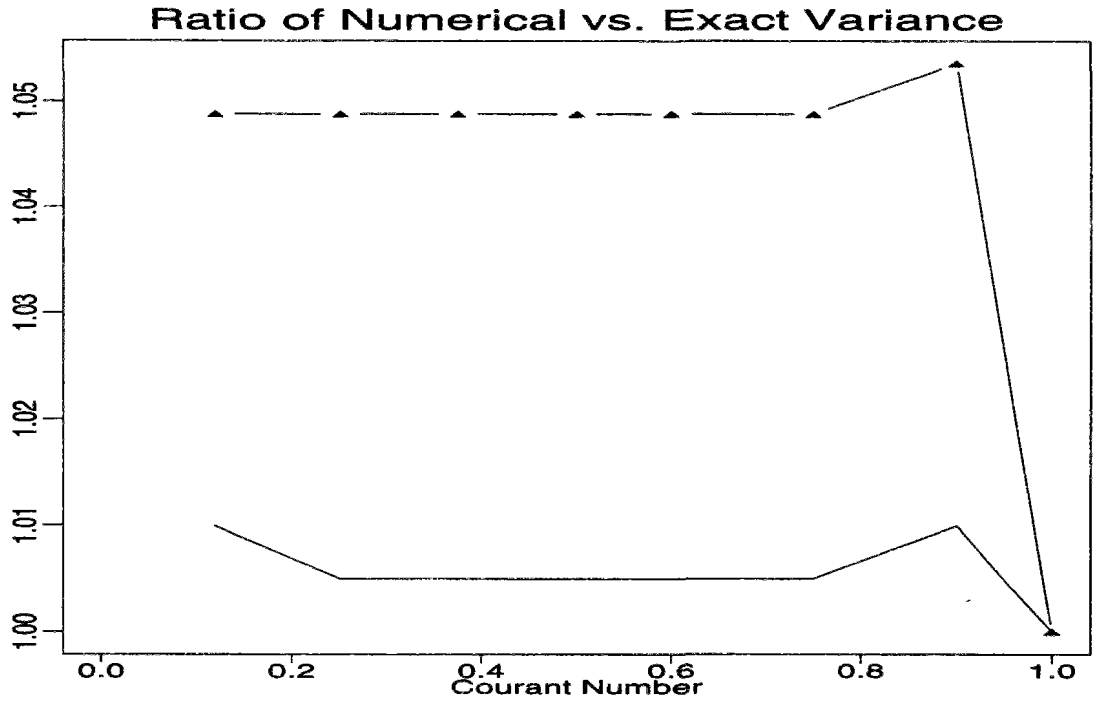


Figure F.13: Influence of derivative initialization on HOLLY : Numerical spreading

F.3 Supplementary results for the advancing front test

The details relevant to this test are given in section 6.5.1. The figures F.15 to F.18 display the error measures for the pure advection case.

From their inspection it stems that once again the use of Dan N'Guyen interpolator must be avoided, since it induces unacceptable under- and over-shoots. On the other hand, by comparing respectively fig F.15 to 6.35, fig F.16 to 6.36, fig F.17 to 6.37 and fig F.18 to 6.38, it appears that the MINIMAX and QUICKEST algorithms raise results quite similar to those of the best schemes so far, namely BOTT3 and BOTT4, RASCH and HOLLY. Besides, the schemes results appear insensitive to the grid variability.

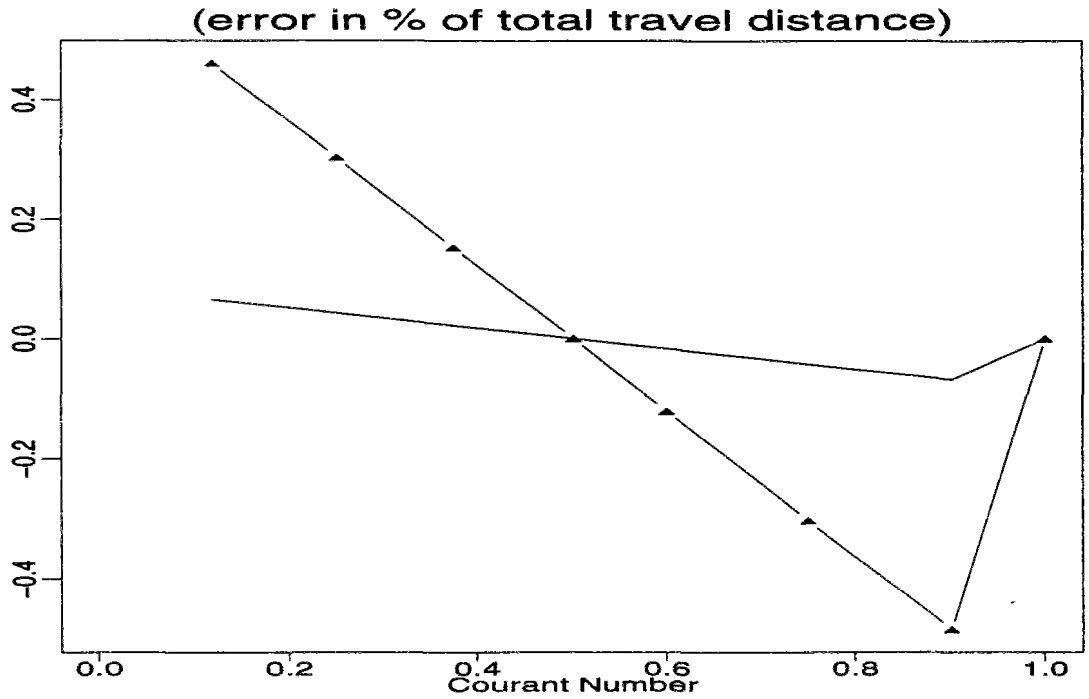


Figure F.14: Influence of derivative initialization on HOLLY : Global phase shift

F.4 Influence of corrective terms estimation in Holly-Preissmann scheme

In appendix E.1.3, we presented the method applied to compute the corrective terms in the equations governing the advection of the scalar derivatives. Two facts influence the evaluation of these terms : how we break the computation of the particles trajectories and what kind of interpolator is used to assess the intermediary derivative values. We have been studying the consequences of these different choices in the frame of the first series of two-dimensional tests (pure advection tests of group A, cf 7.2). Three options were compared :

(i) The trajectory is broken into two segments per mesh cell and the intermediate derivative interpolator is bilinear.

(ii) Same as (i), except that the trajectory is broken into three segments per mesh cell.

(iii) Same as (i), except that the interpolator is bicubic.

Let us first consider the solution of these problems with an uniform discretization. The relevant error measures are displayed in figures F.19 to F.22. Results raised by options (i), (ii) and (iii) are respectively plotted in solid, dotted and dashed lines.

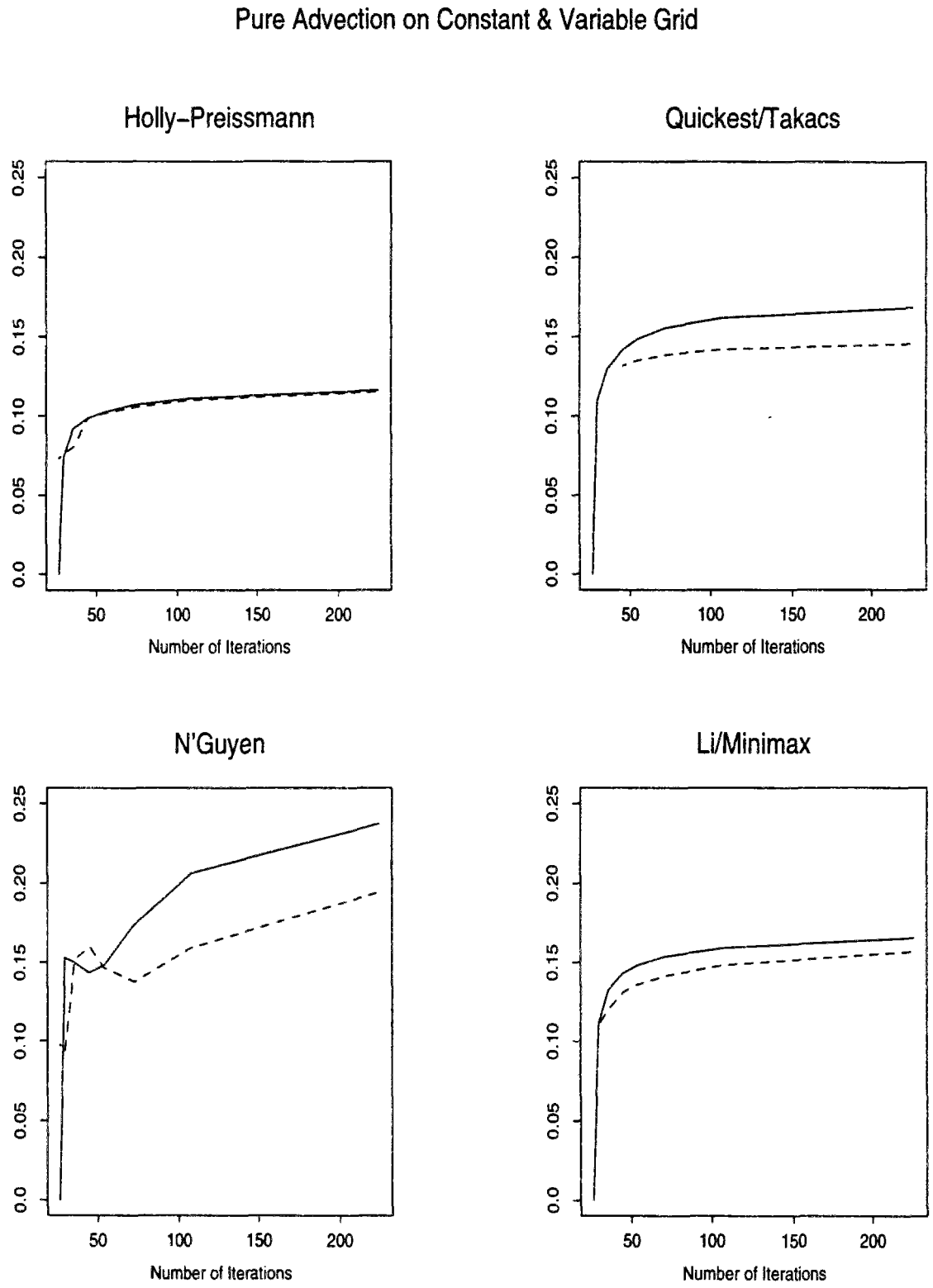


Figure F.15: Advancing front : L2 error norm (in % of total mass)

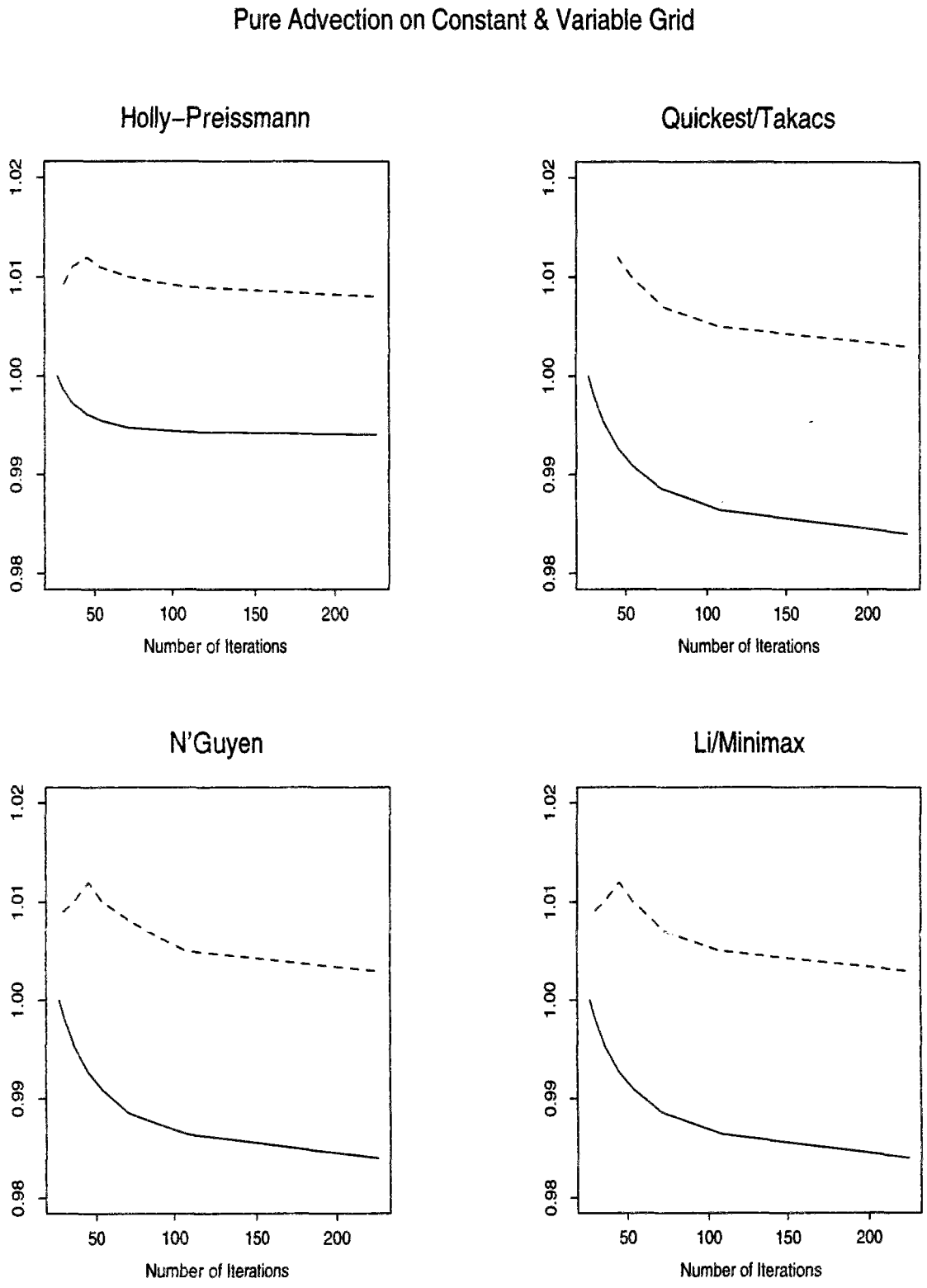


Figure F.16: Advancing front : ratio of computed by exact mass

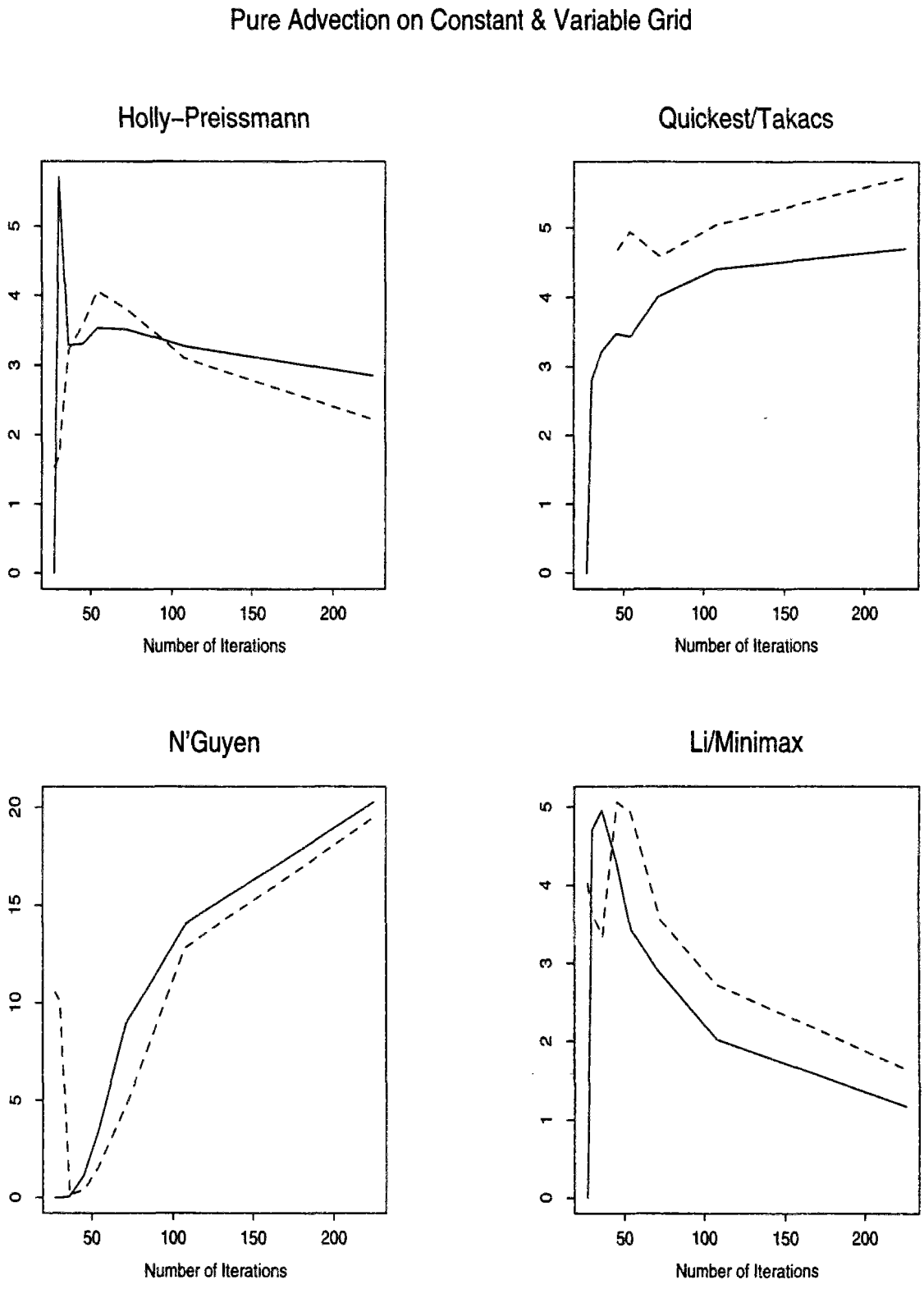


Figure F.17: Advancing front : Maximum overshoot (in % of front concentration)

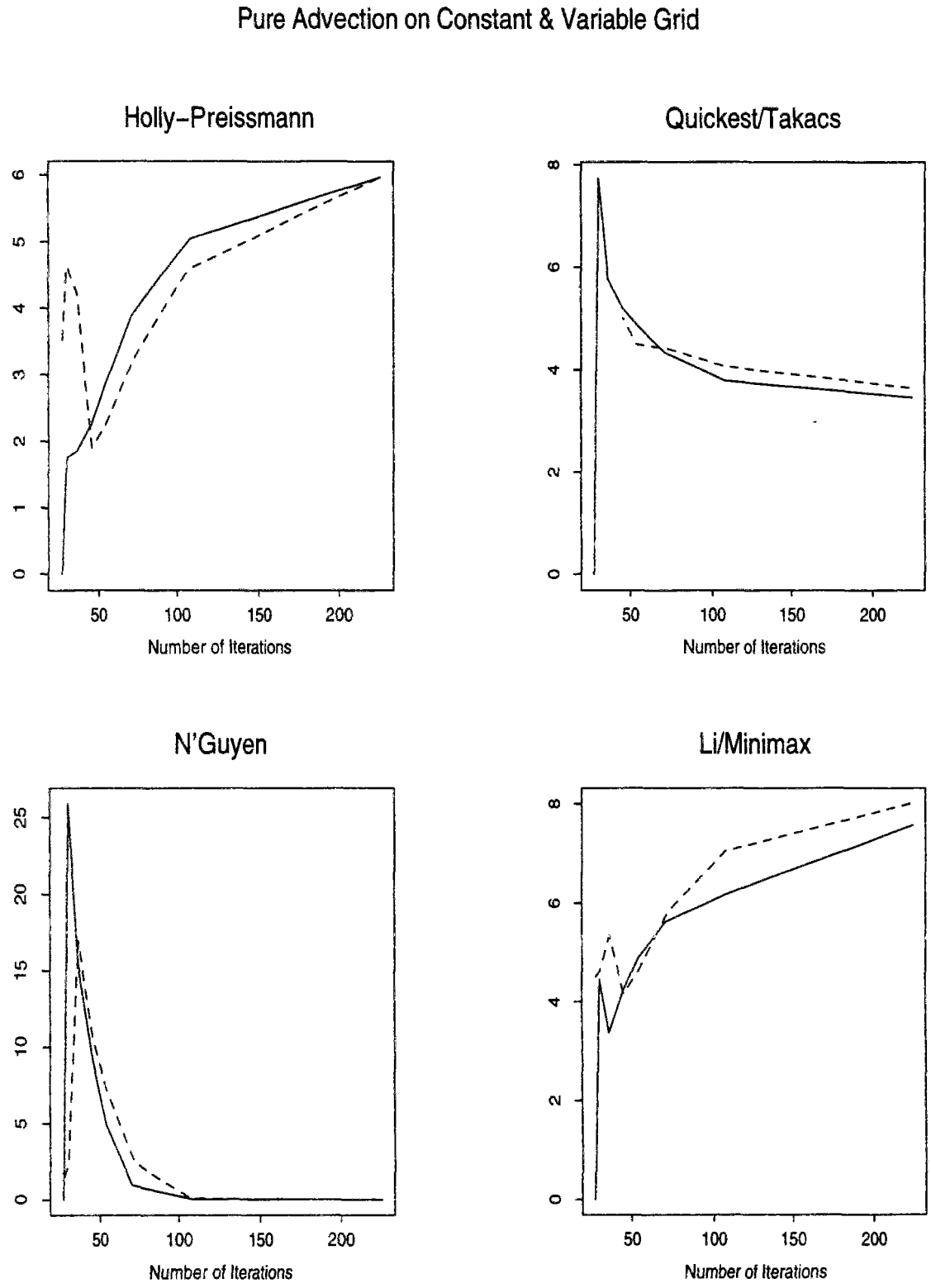


Figure F.18: Advancing front : Maximum undershoot (in % of front concentration)

1. Options (i) and (ii) raise the same level of overall error (see the L2 error norm on fig. F.19). The third option has a similar outcome for large numbers of iterations but differs as this number is reduced : then, its errors are significantly lower.

2. The different options perform well and similarly as regards mass preservation, with a relative mass error less than 0.4 %.

3. The peak damping does not behave monotonically (cf figure F.20) : it first decreases as the number of iterations decreases, then grows again. Yet, this growth is less important with option (iii).

4. The maximum undershoot has the same trend than the peak damping. Once again, troubles arising when the number of iterations is small are more limited when the third option is applied.

5. Errors on the center of mass location have the same order of magnitude whatever the chosen option (see fig. F.21). There is never any error concerning the peak location.

6. Numerical spreading is evaluated in polar coordinates with the help of two second-order moments ratios, one relative to the radius (μ_{rr}), the other one relative to the angle ($\mu_{\theta\theta}$) (cf table 7.2). μ_{rr} levels are similar for the different options. $\mu_{\theta\theta}$ worsens for the smallest iterations numbers. This degradation is significantly less severe for the third option. In fact, negative concentrations appear to play a big part in the evaluation of second-order moments. As these undershoots are more limited with option (iii) this solution is better. This is still more obvious when comparing, on figure F.23 for instance, the concentration profiles obtained for the largest time step $\Delta t = 500$s (solutions with option (i) and (iii) are plotted in solid and dashed lines respectively, reference solution with cross marks).

The reader may check with the help of figures F.24 to F.27 that on variable grid the different options behave as on uniform grid.

Of course, these options have different computer time requirements, as they use more or less complex procedures. The method applied to estimate CPU needs is outlined in 7.4. The more economical version corresponds to option (i). Options (ii) and (iii) are respectively 22 and 4 % more expensive.

Considering both the schemes accuracy and their computational requirements it appears that the best choice consists in applying option (iii), namely breaking the trajectory computation into two segments per mesh cell and computing the intermediate derivatives with the help of a bicubic interpolator.

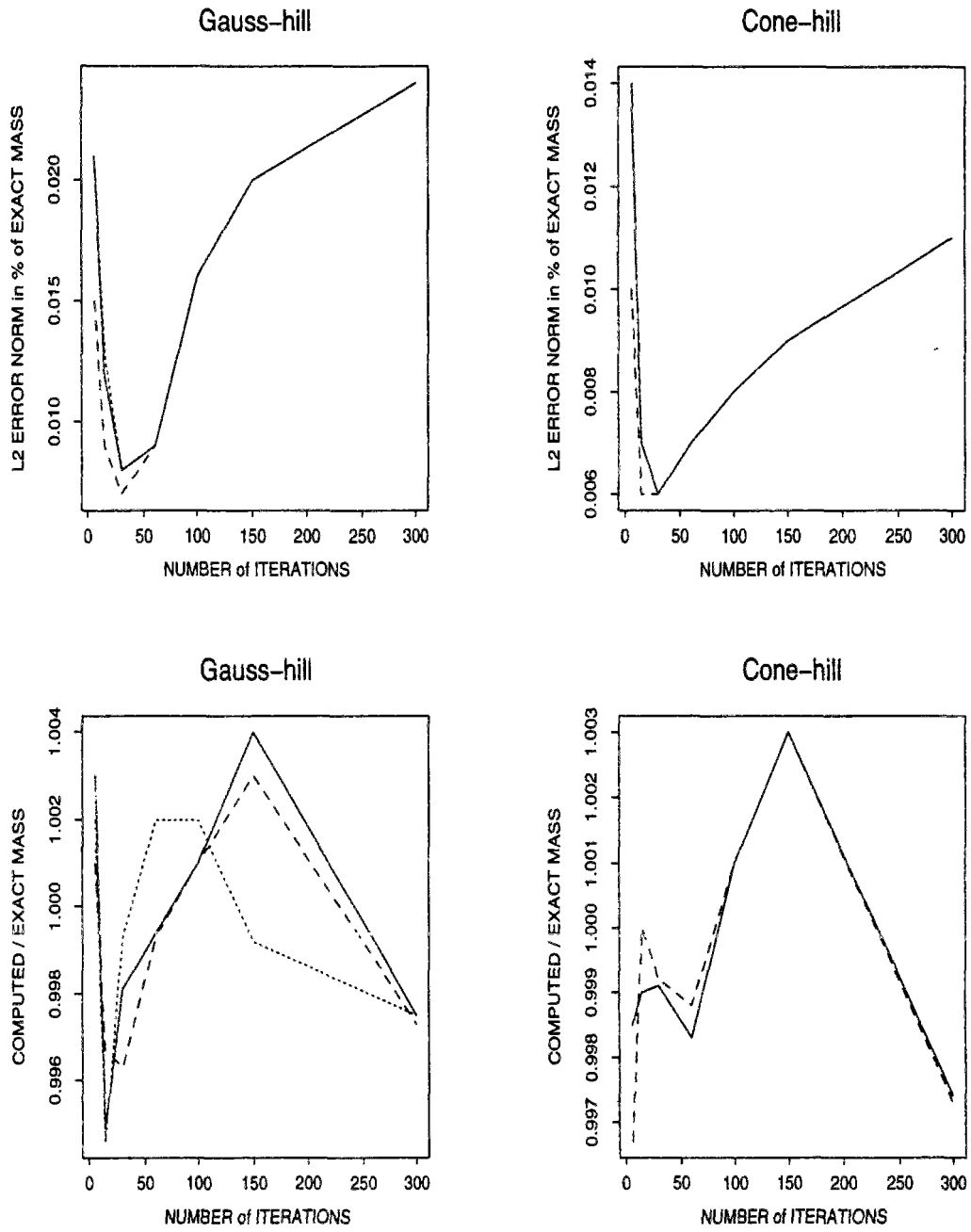


Figure F.19: Influence of corrective terms evaluation on overall & mass conservation errors

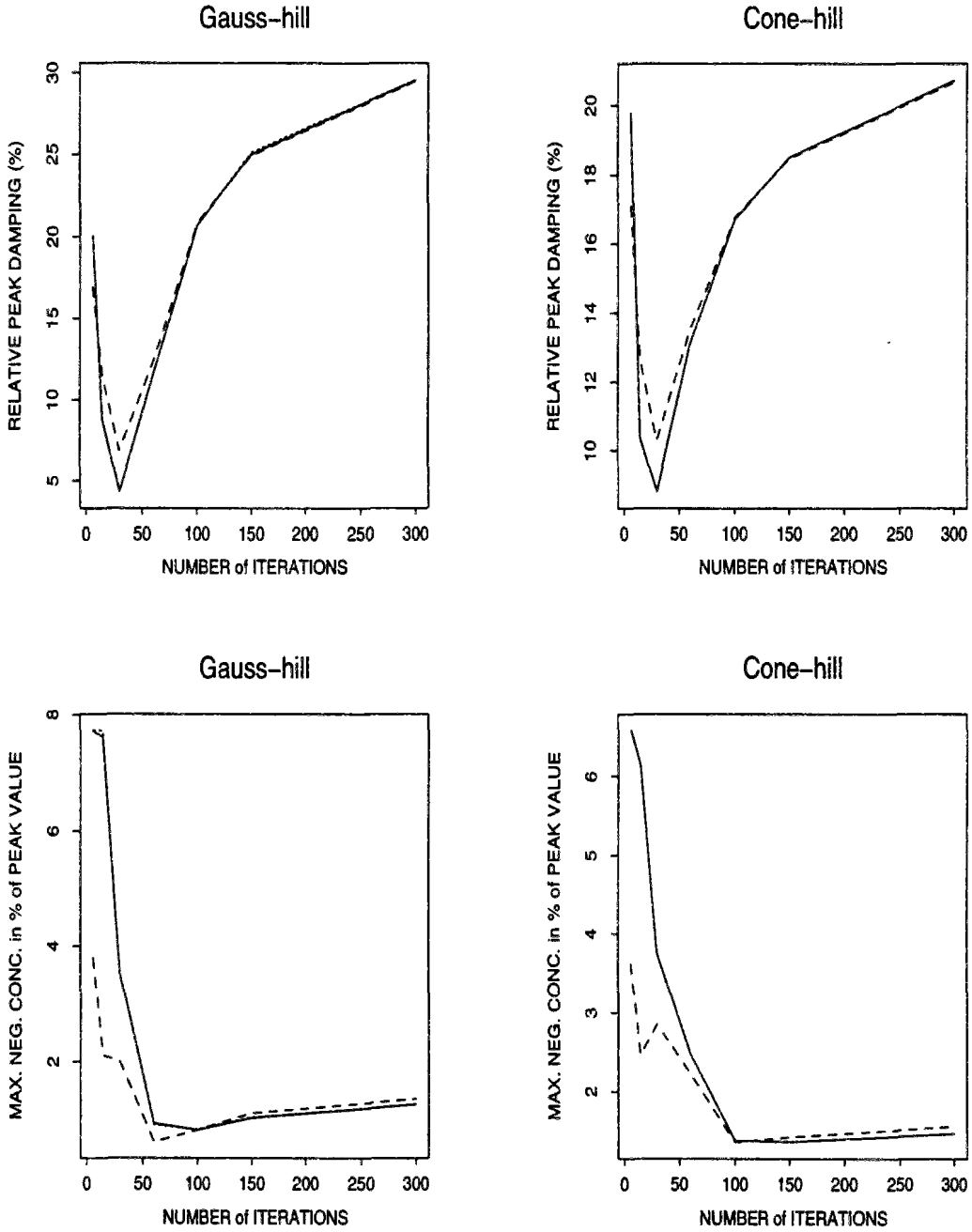


Figure F.20: Influence of corrective terms evaluation on damping & undershoots

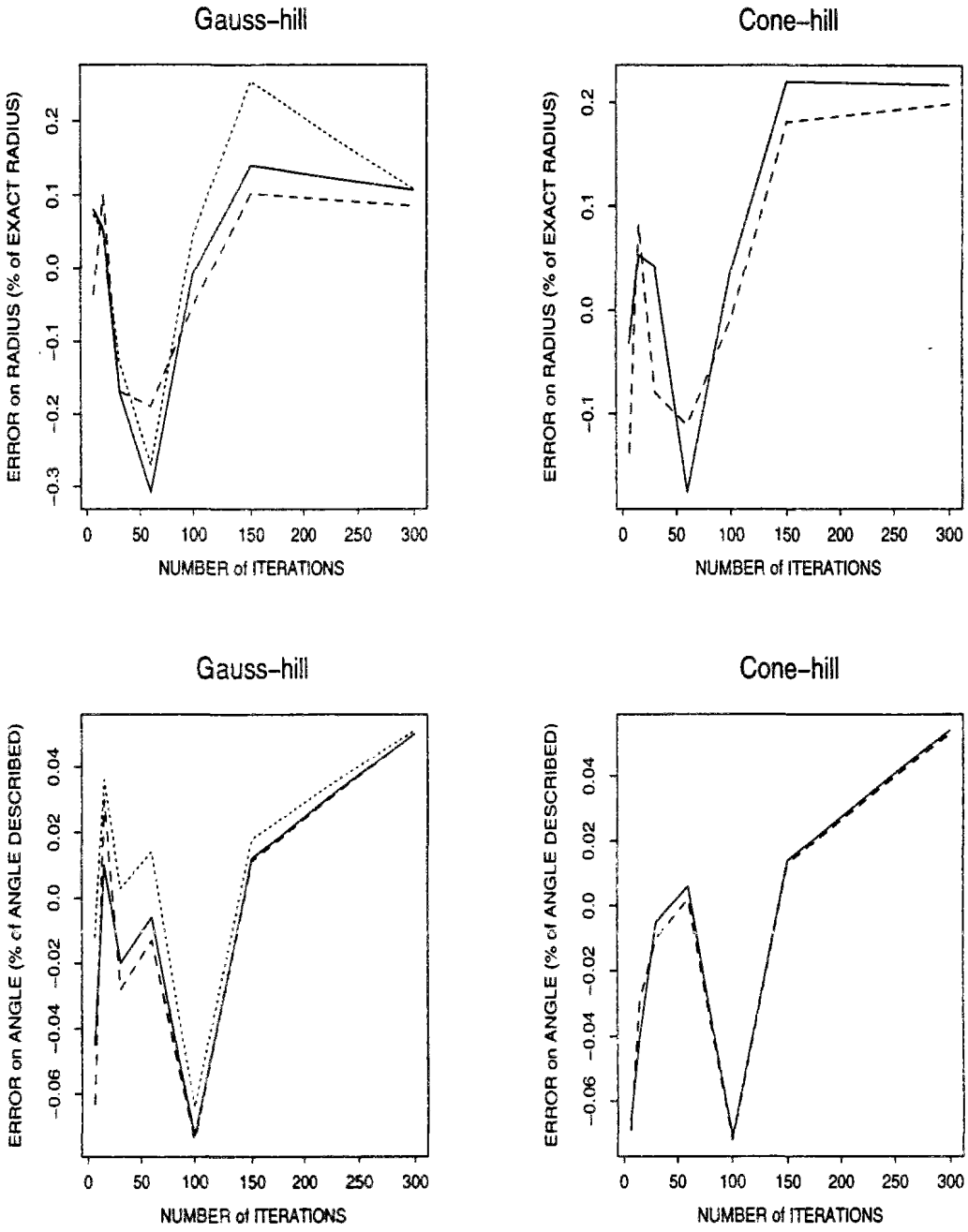


Figure F.21: Influence of corrective terms evaluation on center of mass location

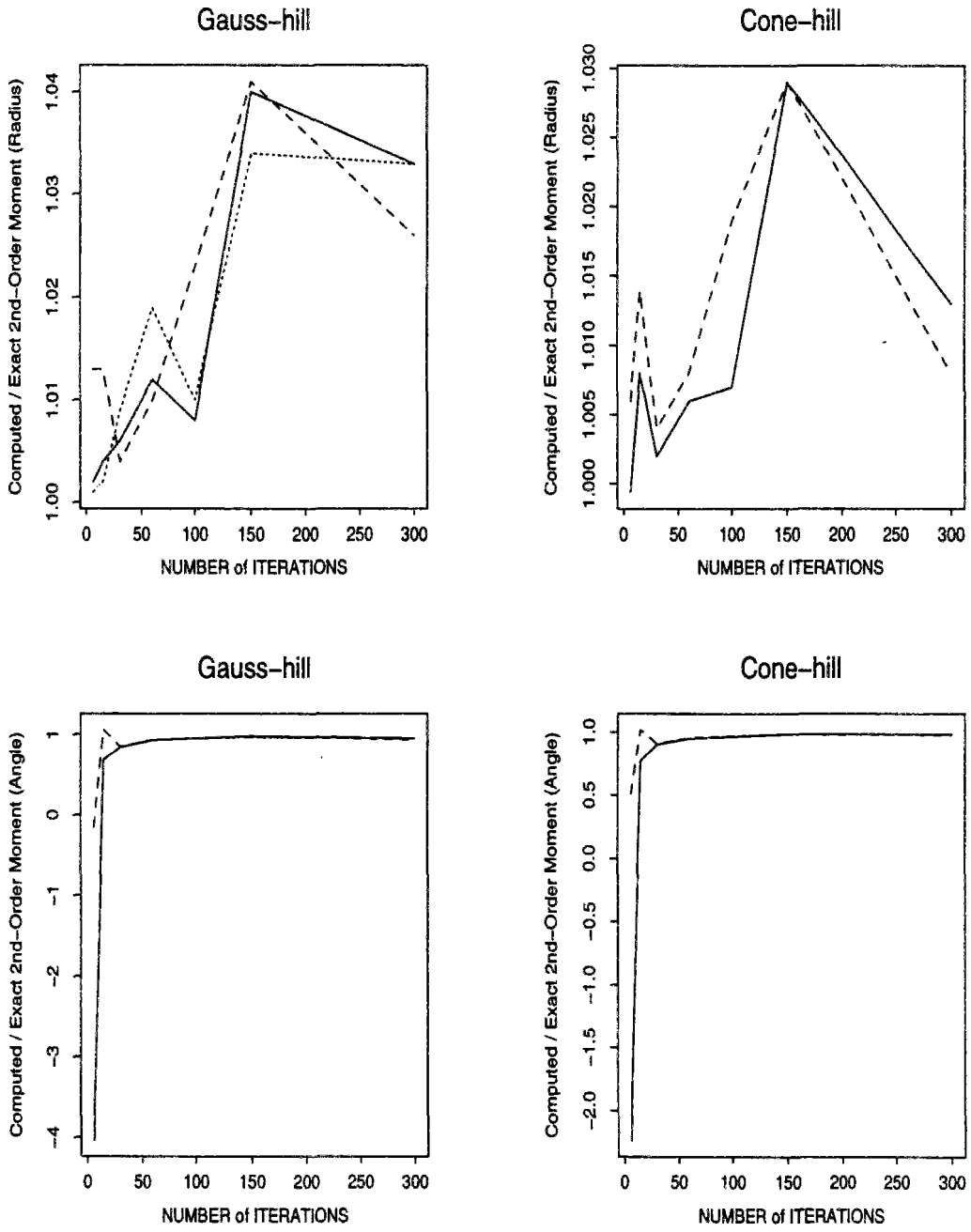


Figure F.22: Influence of corrective terms evaluation on numerical spreading

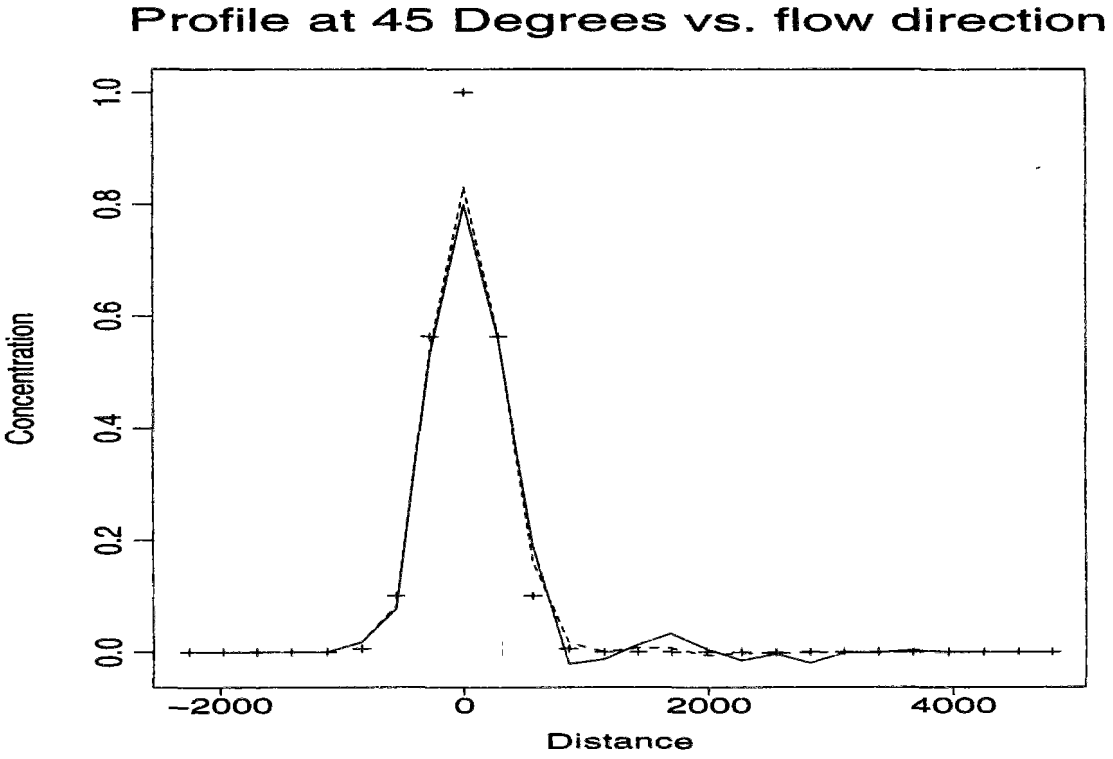


Figure F.23: Gauss-hill rotation on uniform grid : Holly-Preissmann scheme $\Delta t = 500$s

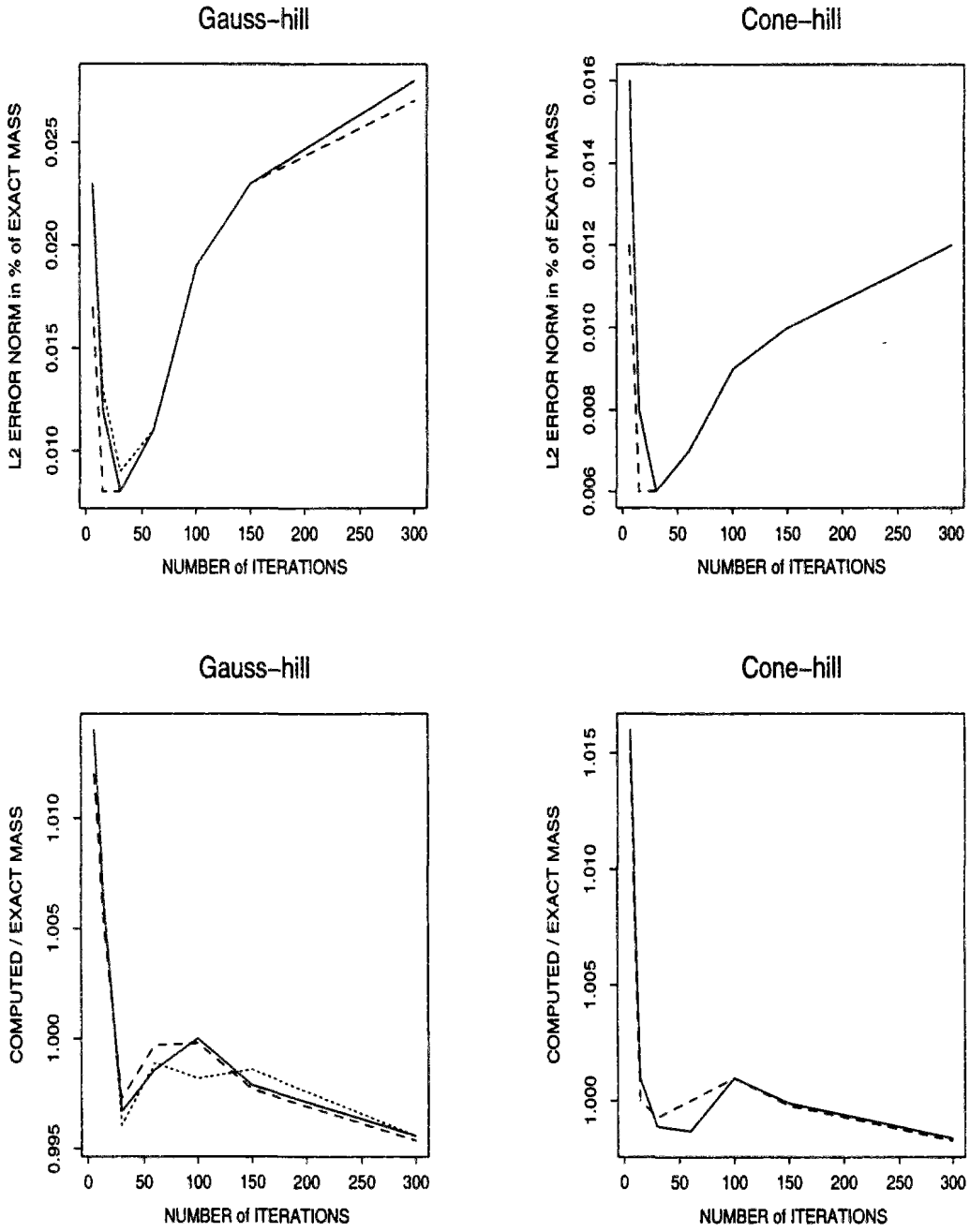


Figure F.24: Variable grid : overall & mass conservation errors

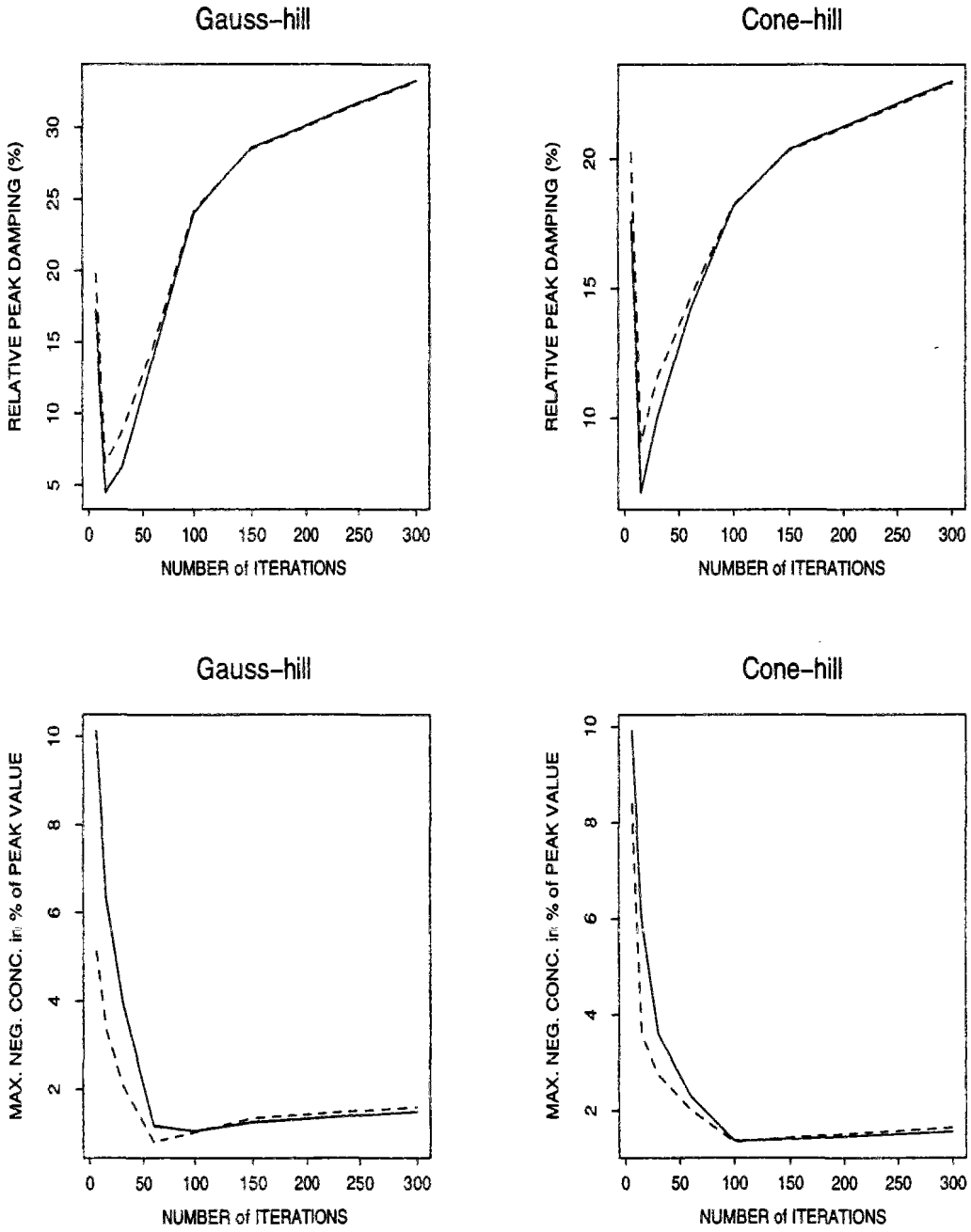


Figure F.25: Variable grid : damping & undershoots

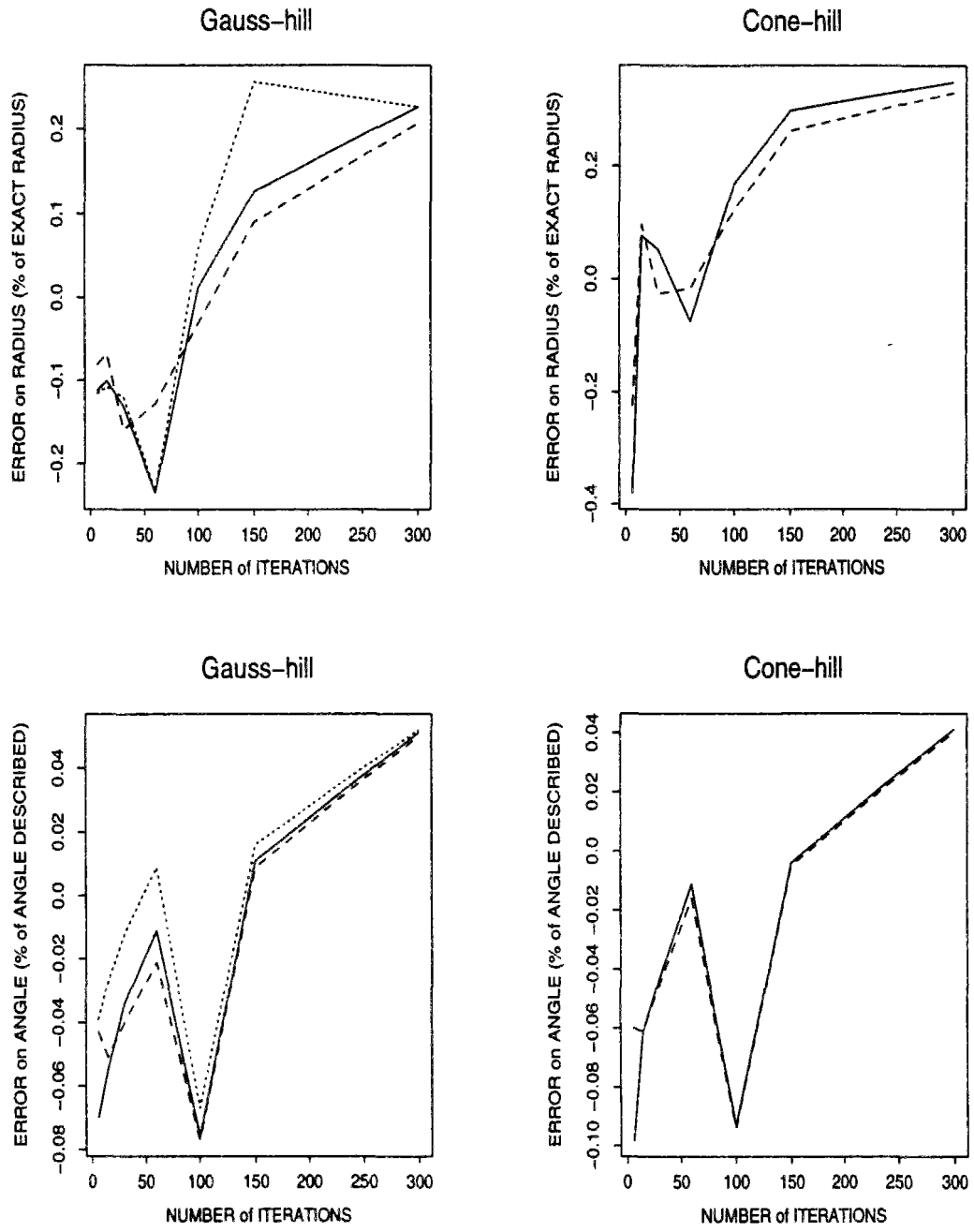


Figure F.26: Variable grid : center of mass location

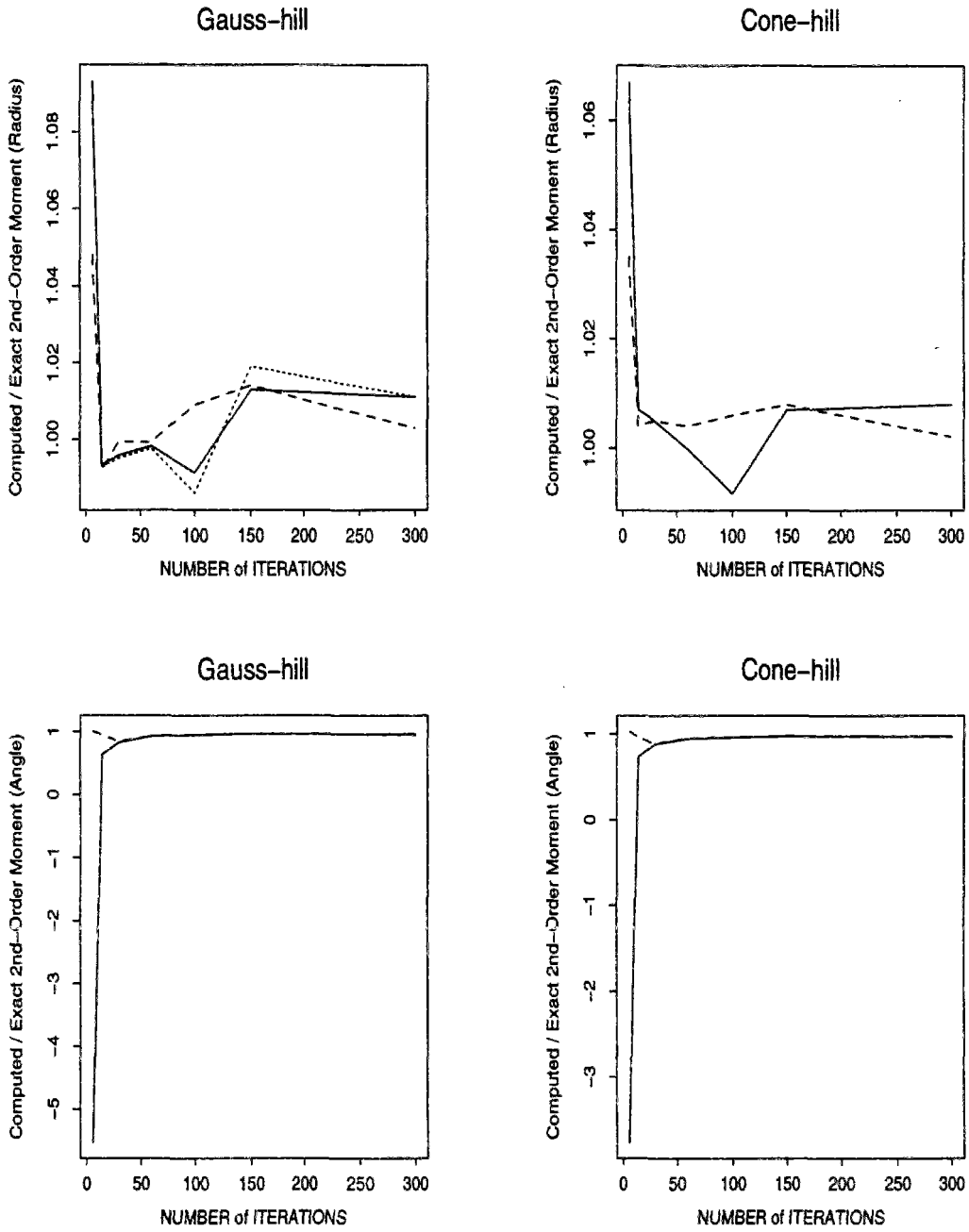


Figure F.27: Variable grid : numerical spreading

Appendix G

Supplementary results - Hydraulic tests

G.1 Backwater curve calculation

Table G.1: Water-depths in sloping channel at $t = 4500$s ($\Delta t = 10$ then 5s)

X	Runge-Kutta	TELEMAC	UH2	QH2
0	1.0000	1.0000	1.0000	1.0000
500	1.0000	1.0000	1.0000	1.0000
1000	1.0000	1.0001	1.0001	1.0000
1500	1 .0001	1.0003	1.0002	1.0001
2000	1.0005	1.0009	1.0009	1.0006
2500	1.0031	1.0041	1.0038	1.0033
3000	1.0172	1.0194	1.0185	1.0176
3500	1.0835	1.0867	1.0850	1.0839
3600	1.1102		1.1117	1.1106
3700	1.1430		1.1443	1.1432
3800	1.1820		1.1832	1.1822
3900	1.2273		1.2283	1.2275
4000	1.2786	1.2807	1.2794	1.2787
4100	1.3354		1.3361	1.3355
4200	1.3971		1.3976	1.3971
4300	1.4630		1.4634	1.4631
4400	1.5326		1.5329	1.5326
4500	1.6053	1.6060	1.6056	1.6053
4600	1.6805		1.6807	1.6805
4700	1.7580		1.7581	1.7580
4800	1.8372		1.8373	1.8372
4900	1.9180		1.9180	1.9180
5000	2.0000	2.0000	2.0000	2.0000

G.2 Fluvial flow over a sill

In order to identify the main source of trouble in this test, we applied two kinds of tests.

First, we run the models with different upstream conditions regarding the free surface elevation. In a first series of trials, we made use of "clamped" boundaries conditions, that is to say we set $\zeta = 2\text{m}$ at the inflow boundary. In a second series of trials, we computed the upstream depth by making an explicit development of the continuity equation : $\delta z = -\Delta t \, \text{div}\vec{Q}^n$. None of these modifications prevents the development of an extra elevation of the free surface profile upstream the sill. Over and downstream the sill, the surface profiles computed under the three different upstream boundary conditions stand no farther apart than 0.2 mm ! The overall errors on the flow rate are similar but errors are distributed slightly differently along the channel. For instance, clamped boundaries conditions induce systematically an underestimation of the flow rate uptream the sill and slightly improves its preservation over it. We also noticed that clamped boundaries conditions were accelerating the models convergence rate.

In fact, the main source of errors does not lie in the approximation of upstream boundary conditions but in the unability of the model to deal perfectly with bathymetric discontinuities. With the original parabolic sill, we have indeed :

$$\lim_{x \to 12^-} \frac{\partial z_b}{\partial x} = -0.2 \quad \text{while} \quad \lim_{x \to 12^+} \frac{\partial z_b}{\partial x} = 0$$

$$\lim_{x \to 8^-} \frac{\partial z_b}{\partial x} = 0 \quad \text{while} \quad \lim_{x \to 8^+} \frac{\partial z_b}{\partial x} = +0.2$$

These brutal changes in the bottom slope induce the dependent variables to display sharp gradients at each extremity of the sill. This represents a challenge not only when solving the advection step but also within the propagation step as its governing equations are discretized with centred differences, ill-suited to account for gradient discontinuities. Most errors decrease when the bathymetric discontinuities are somewhat smoothed out, as demonstrate the following test.

Let us consider a smoother sill which differs from the original one in areas $7.6 \leq x \leq 8.2$ and $11.8 \leq x \leq 12.4$. In interval [7.6, 8.2] $z_b(x)$ is a cubic defined by :

$$z_b(7.6) = \frac{\partial z_b}{\partial x}(7.6) = 0, \qquad z_b(8.2) = 0.038 \quad \text{and} \quad \frac{\partial z_b}{\partial x}(8.2) = 0.18$$

We proceed similarly for interval [11.8, 12.4]. Thus, the bottom slope remains continuous all over the sill.

When models QH2 and UH2 are applied to compute the surface profile over this new bottom profile, we notice the following (cf table G.2) :

- Flow preservation with model QH2 is clearly improved : the average flow error is halved and the maximum error is approximately reduced three times. Improvement is less impressive with model UH2, except for $\Delta t = 0.01$ where errors are halved.
- The upstream extra elevation is not modified with the two biggest time steps but it is disminished when $\Delta t = 0.05$ (by 60 and 20 % for QH2 and UH2 respectively).
- The dowstream spurious overshoot is fairly reduced in case $\Delta t = 0.05$ (by a factor 6 and 3 for QH2 and UH2 respectively) and completely disappear when simulations are performed at $\Delta t = 0.01$.

Table G.2: Error measures : smoother sill test

Δt (s)	$\overline{\delta Q_x}$ (%)	$\max \delta Q_x$ (%)	$\overline{\Delta H}$ (mm)	$\max \Delta H$ (mm)	$\overline{\delta h}$ (%)	$\max \delta h$ (%)
	Model QH2 (bicubic interpolant)					
0.20	0.04	0.18	12.3	25.9	0.67	1.43
0.10	0.03	0.13	6.1	12.7	0.33	0.68
0.05	0.03	0.10	3.1	7.2	0.17	0.38
0.01	0.02	0.06	0.7	1.7	0.04	0.09
	Model UH2 (bicubic interpolant)					
0.20	1.25	3.17	23.0	47.0	1.25	2.66
0.10	0.68	1.74	12.1	24.8	0.66	1.36
0.05	0.36	0.90	6.3	14.3	0.34	0.75
0.01	0.08	0.22	1.3	3.8	0.07	0.20

Table G.3: Water-depths computed for sill test - Model QH2

X	Exact	QH2 (bicubic interpolant)			
	solution	Δt = 0.2s	Δt = 0.1s	Δt = 0.05s	Δt = 0.01s
8.0	2.0000	2.0052	2.0018	2.0015	2.0000
8.2	1.9486	1.9652	1.9598	1.9578	1.9522
8.4	1.9014	1.9241	1.9146	1.9092	1.9018
8.6	1.8585	1.8838	1.8711	1.8644	1.8595
8.8	1.8202	1.8461	1.8320	1.8258	1.8214
9.0	1.7868	1.8122	1.7979	1.7924	1.7878
9.2	1.7587	1.7828	1.7689	1.7639	1.7596
9.4	1.7363	1.7584	1.7452	1.7407	1.7370
9.6	1.7200	1.7392	1.7271	1.7233	1.7204
9.8	1.7100	1.7256	1.7149	1.7121	1.7102
10.0	1.7067	1.7178	1.7089	1.7073	1.7066
10.2	1.7100	1.7162	1.7094	1.7092	1.7096
10.4	1.7200	1.7208	1.7165	1.7177	1.7193
10.6	1.7363	1.7319	1.7303	1.7329	1.7353
10.8	1.7587	1.7495	1.7505	1.7544	1.7576
11.0	1.7868	1.7734	1.7770	1.7818	1.7853
11.2	1.8202	1.8036	1.8093	1.8149	1.8189
11.4	1.8585	1.8397	1.8471	1.8530	1.8571
11.6	1.9014	1.8814	1.8898	1.8961	1.9001
11.8	1.9486	1.9281	1.9371	1.9432	1.9471
12.0	2.0000	1.9798	1.9893	1.9953	1.9983
12.2	2.0000	1.9914	1.9987	2.0030	2.0019
12.4	2.0000	1.9977	2.0006	2.0014	1.9988
12.6	2.0000	2.0002	1.9998	1.9997	1.9995
12.8	2.0000	2.0008	1.9992	1.9996	1.9995
13.0	2.0000	2.0007	1.9990	2.0000	1.9994
13.5	2.0000	2.0000	1.9996	1.9998	1.9996
14.0	2.0000	1.9997	1.9999	1.9997	1.9994

Table G.4: Water-depths computed for sill test - Model UH2

X	Exact	UH2 (bicubic interpolant)			
	solution	Δt = 0.2s	Δt = 0.1s	Δt = 0.05s	Δt = 0.01s
8.0	2.0000	2.0106	2.0041	2.0019	2.0000
8.2	1.9486	1.9728	1.9668	1.9642	1.9554
8.4	1.9014	1.9370	1.9253	1.9170	1.9035
8.6	1.8585	1.9014	1.8834	1.8713	1.8616
8.8	1.8202	1.8662	1.8447	1.8317	1.8228
9.0	1.7868	1.8340	1.8105	1.7978	1.7891
9.2	1.7587	1.8054	1.7810	1.7690	1.7606
9.4	1.7363	1.7811	1.7565	1.7453	1.7377
9.6	1.7200	1.7613	1.7372	1.7270	1.7209
9.8	1.7100	1.7464	1.7235	1.7146	1.7103
10.0	1.7067	1.7366	1.7155	1.7085	1.7063
10.2	1.7100	1.7323	1.7137	1.7089	1.7091
10.4	1.7200	1.7336	1.7182	1.7159	1.7185
10.6	1.7363	1.7408	1.7293	1.7297	1.7345
10.8	1.7587	1.7541	1.7468	1.7500	1.7566
11.0	1.7868	1.7736	1.7709	1.7766	1.7846
11.2	1.8202	1.7994	1.8013	1.8091	1.8179
11.4	1.8585	1.8313	1.8376	1.8470	1.8563
11.6	1.9014	1.8693	1.8795	1.8899	1.8993
11.8	1.9486	1.9132	1.9265	1.9375	1.9467
12.0	2.0000	1.9627	1.9781	1.9889	1.9977
12.2	2.0000	1.9755	1.9929	2.0030	2.0037
12.4	2.0000	1.9872	1.9992	2.0030	1.9991
12.6	2.0000	1.9949	2.0003	2.0001	1.9996
12.8	2.0000	1.9990	2.0000	1.9991	2.0000
13.0	2.0000	2.0012	1.9997	1.9993	1.9998
13.5	2.0000	2.0015	1.9997	2.0003	2.0001
14.0	2.0000	2.0005	2.0000	2.0001	2.0000

G.3 Waves propagation in a closed basin

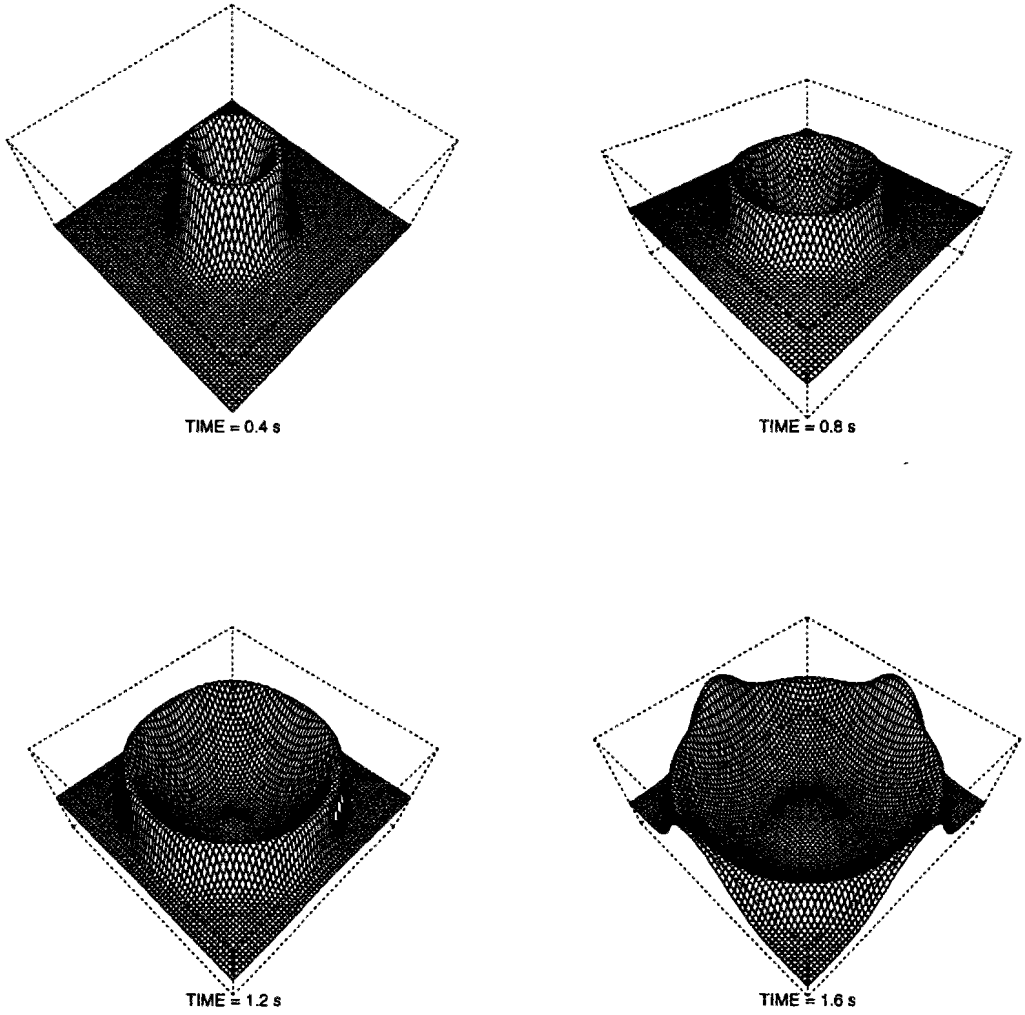


Figure G.1: Computed water profiles for $0.4 \leq t \leq 1.6$ s ($\Delta t = 0.04, \gamma = 0.6$)

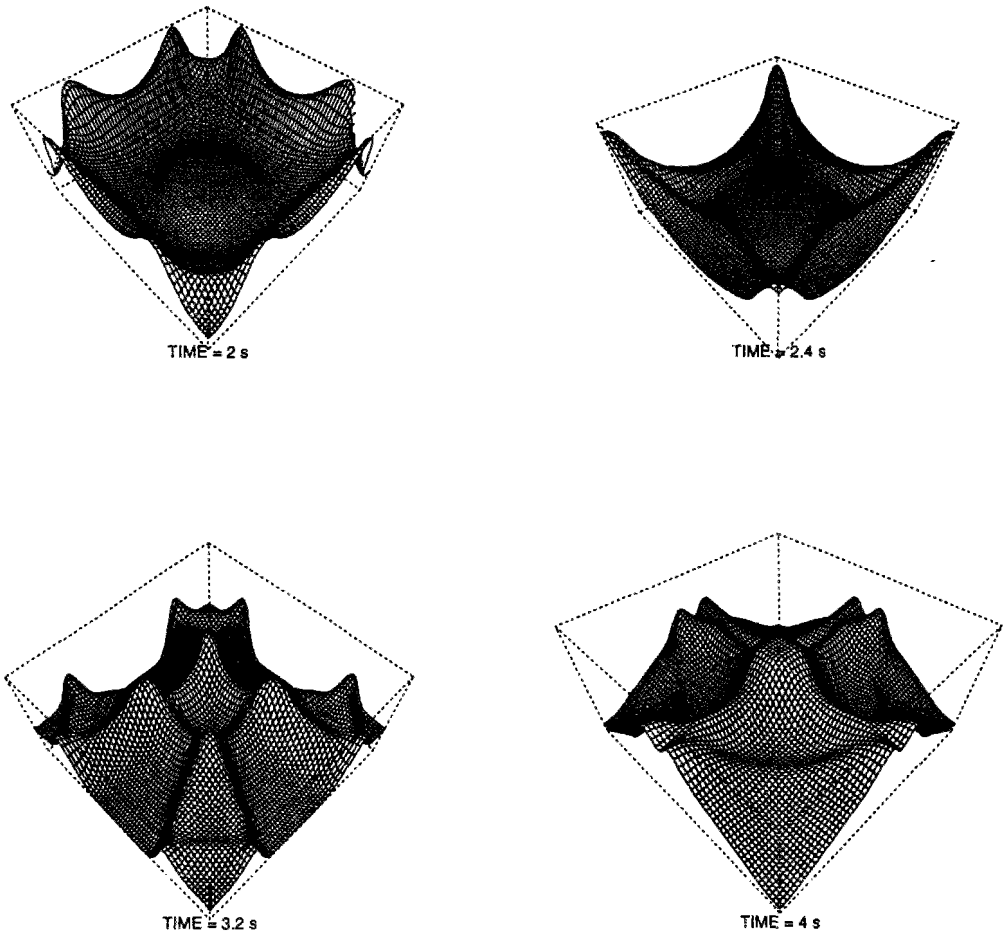


Figure G.2: Computed water profiles for $2.0 \leq t \leq 4.0$ s ($\Delta t = 0.04, \gamma = 0.6$)

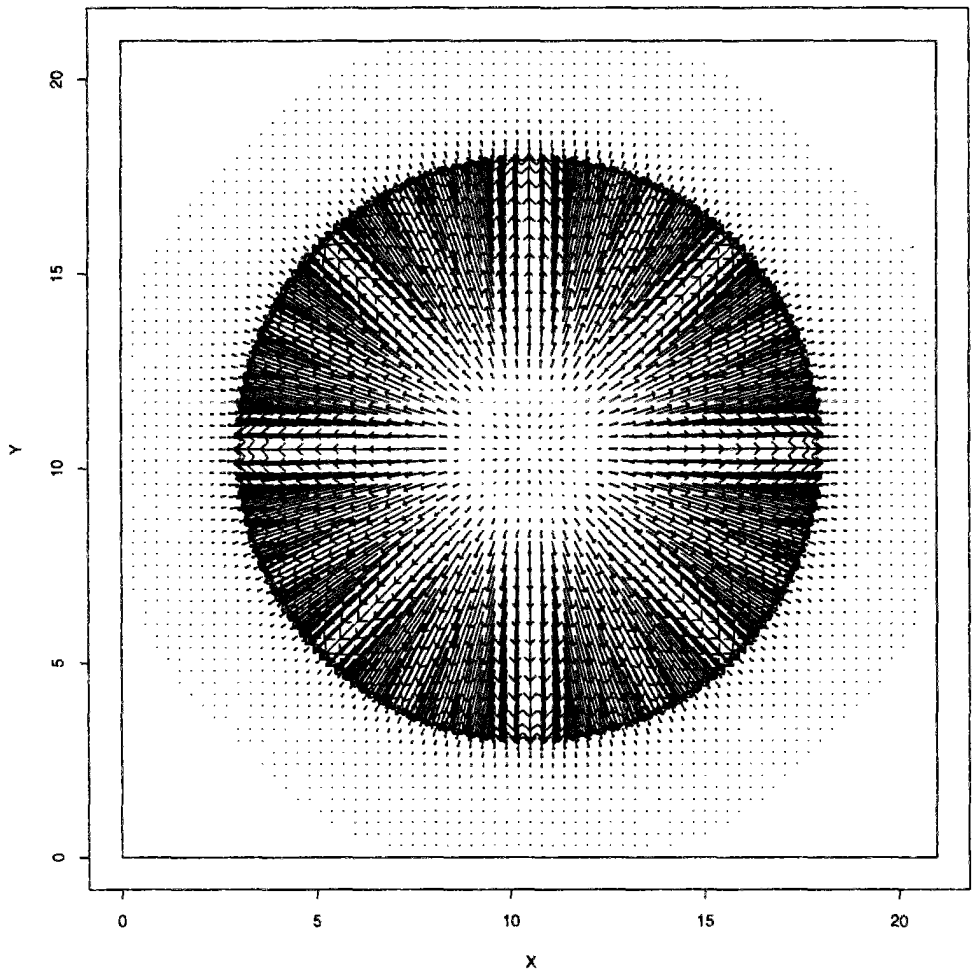


Figure G.3: Computed velocity field at $t = 0.8$ s ($\Delta t = 0.04$, $\gamma = 0.6$)

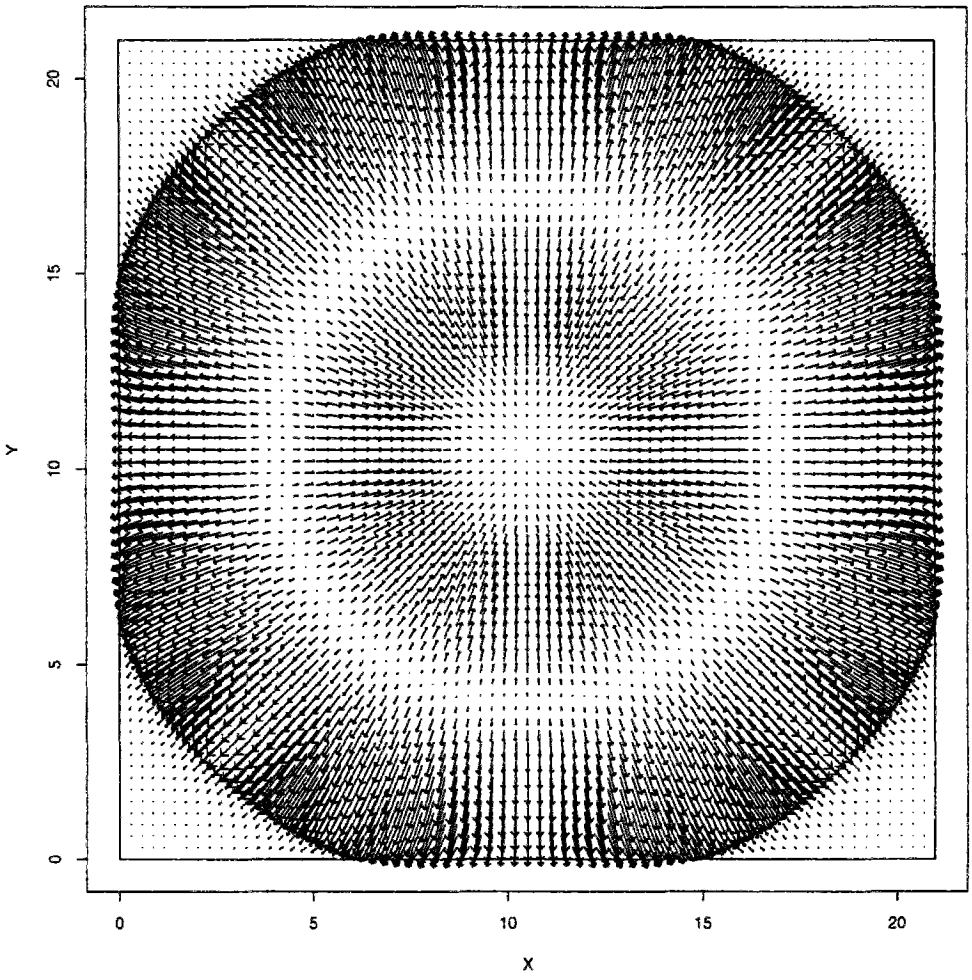


Figure G.4: Computed velocity field at $t = 1.6$ s ($\Delta t = 0.04$, $\gamma = 0.6$)

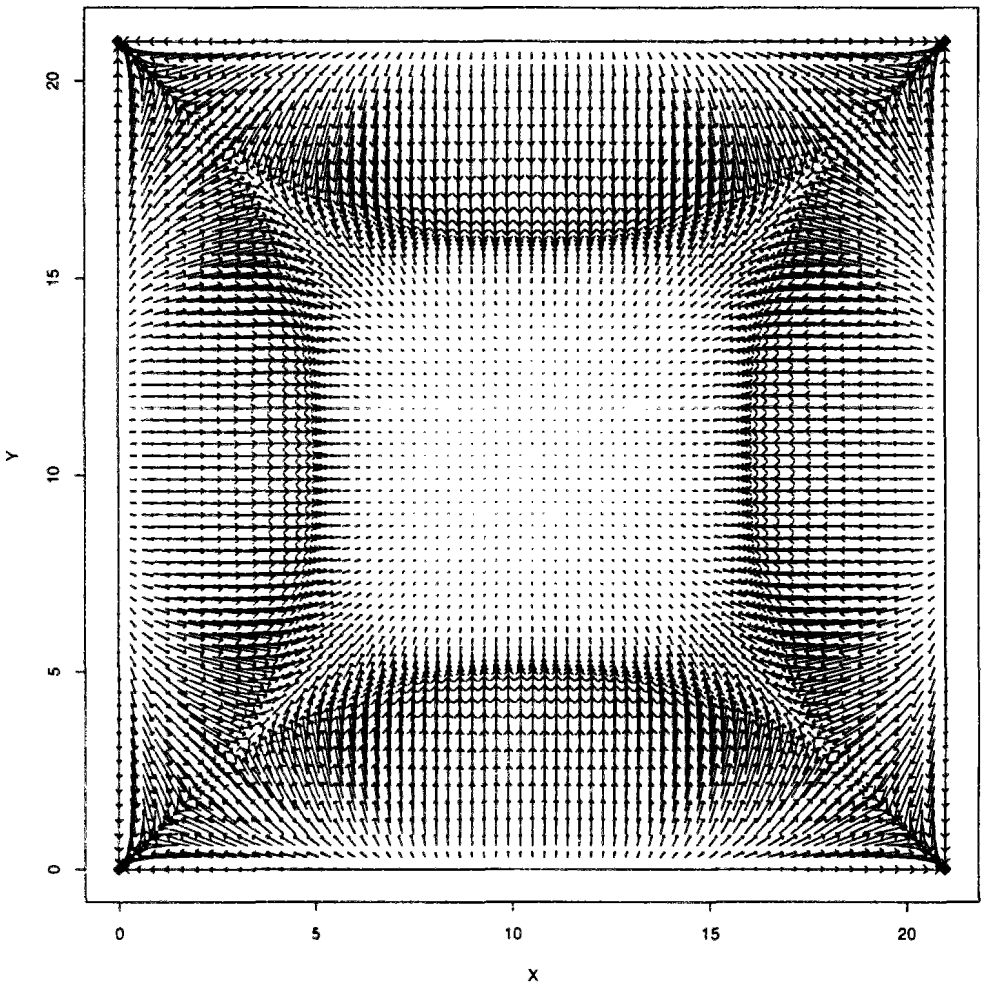


Figure G.5: Computed velocity field at $t = 2.4$ s ($\Delta t = 0.04$, $\gamma = 0.6$)

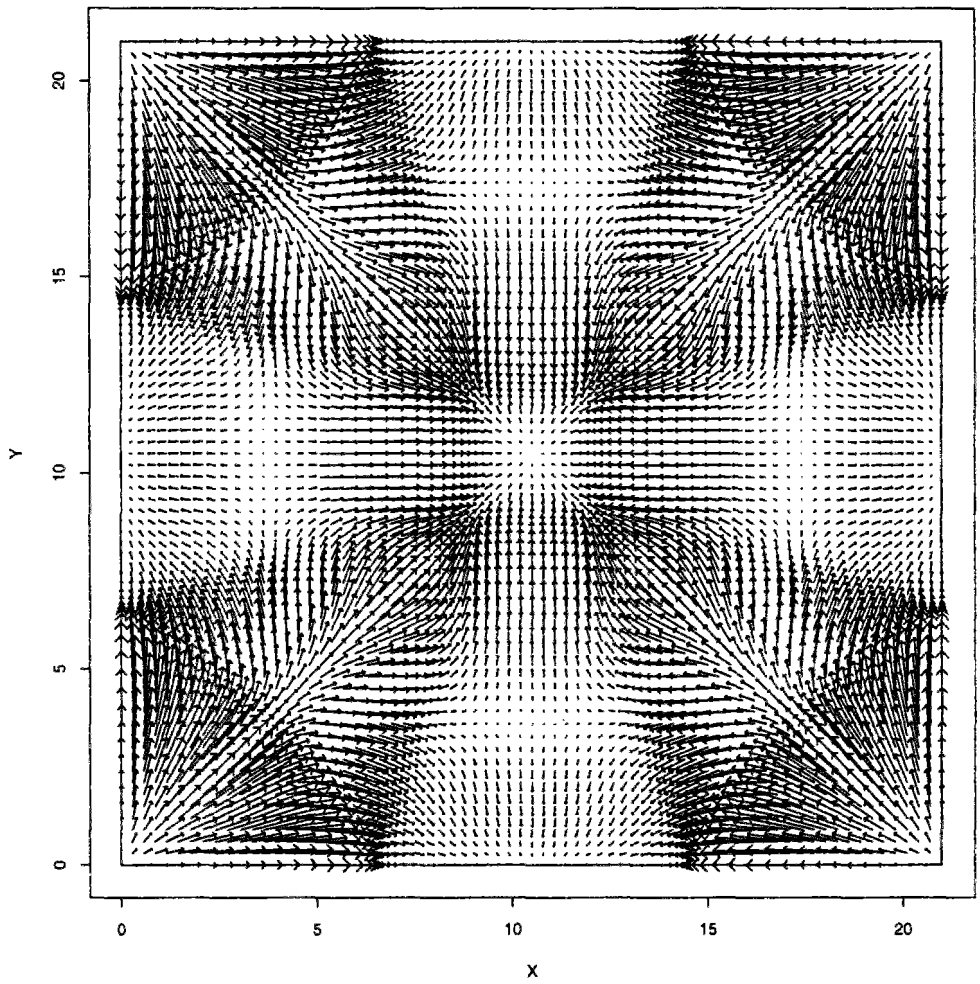


Figure G.6: Computed velocity field at $t = 3.2$ s ($\Delta t = 0.04$, $\gamma = 0.6$)

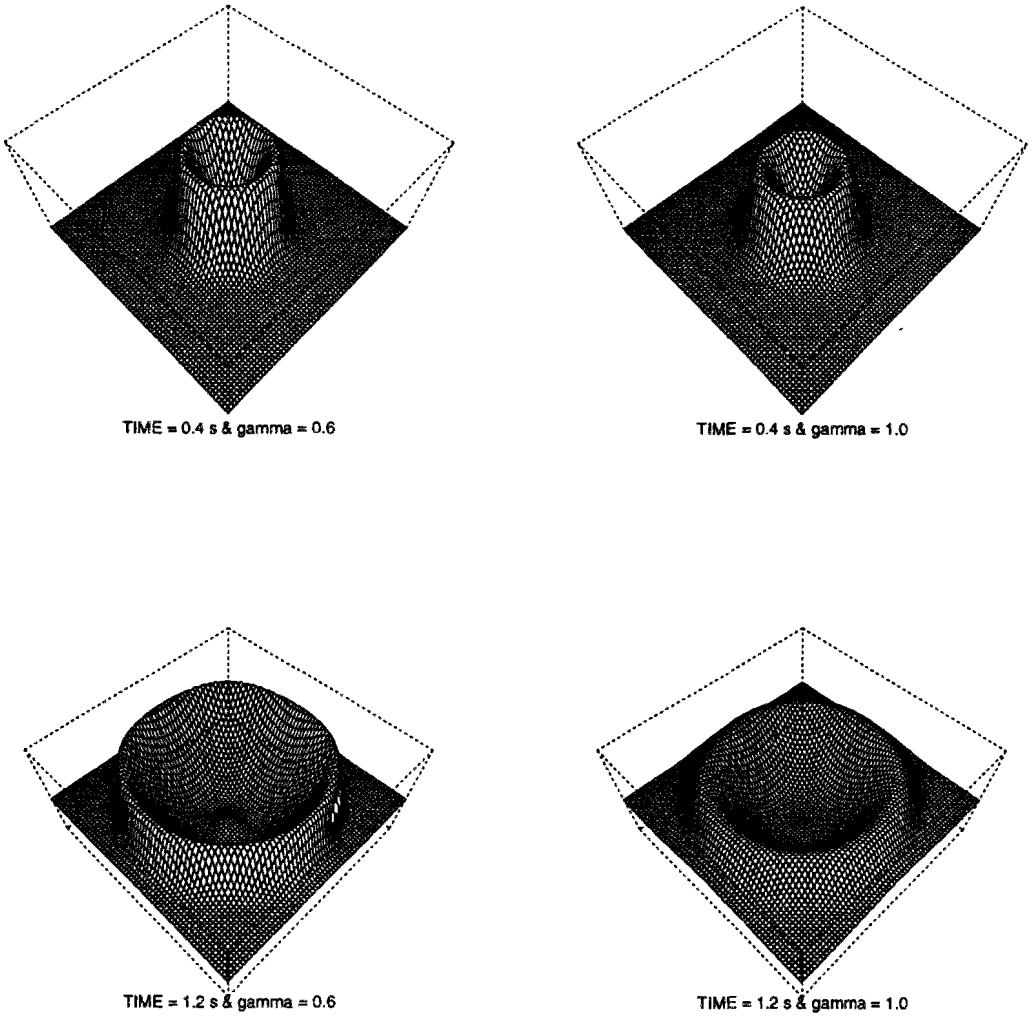


Figure G.7: Influence of γ choice on forecasted surface profiles for $t = 0.4$ and 1.2 s ($\Delta t = 0.04$)

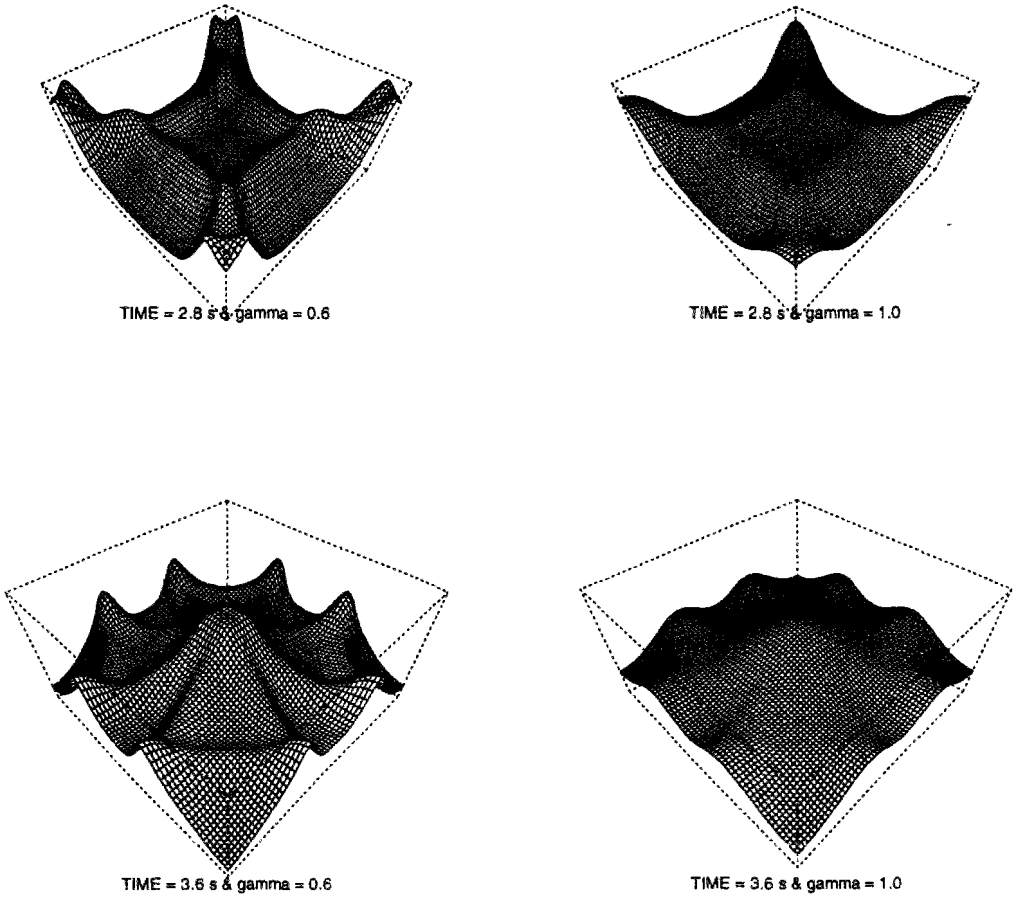


Figure G.8: Influence of γ choice on forecasted surface profiles for $t = 2.8$ and 3.6 s ($\Delta t = 0.04$)

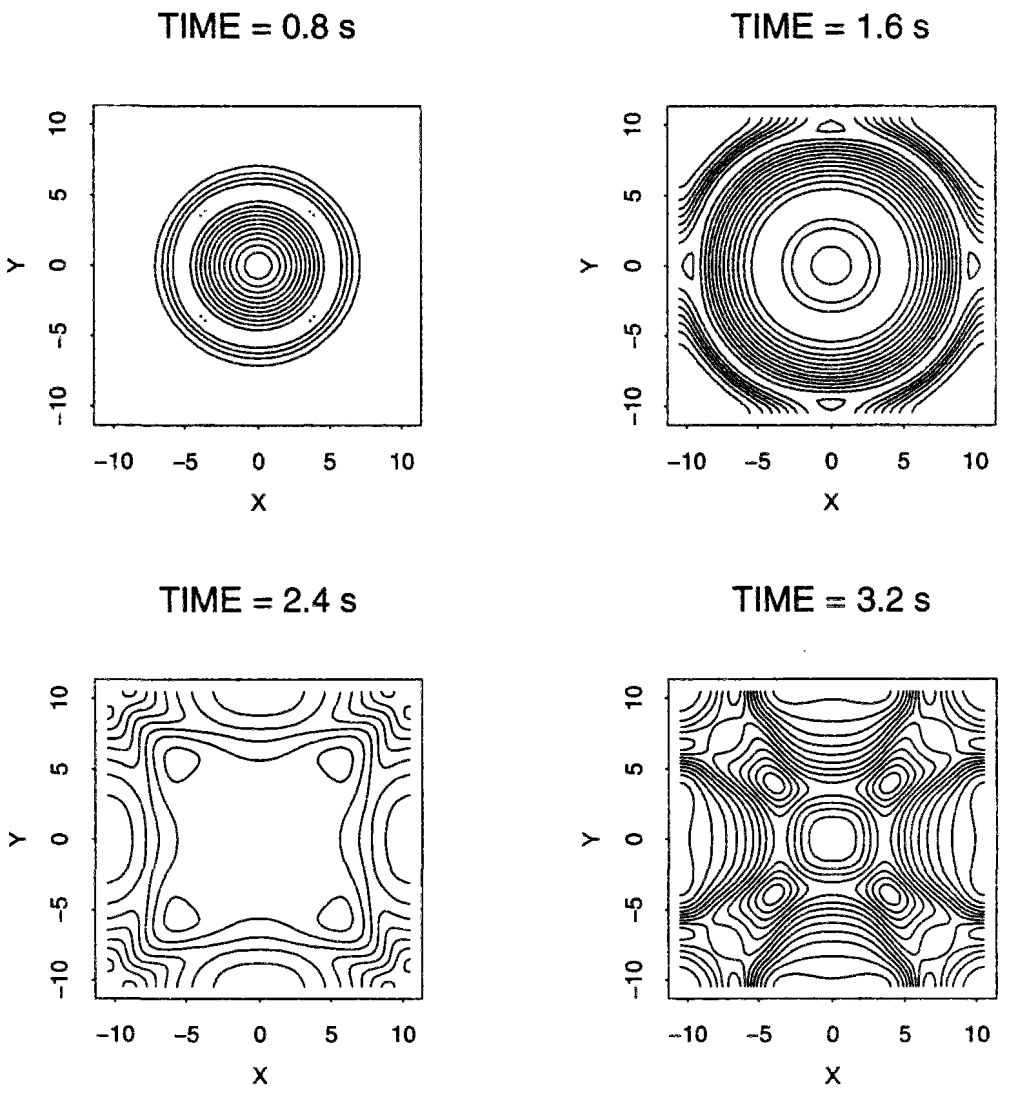


Figure G.9: Water level isolines - "plain" model QH ($\Delta t = 0.04, \gamma = 0.6$)

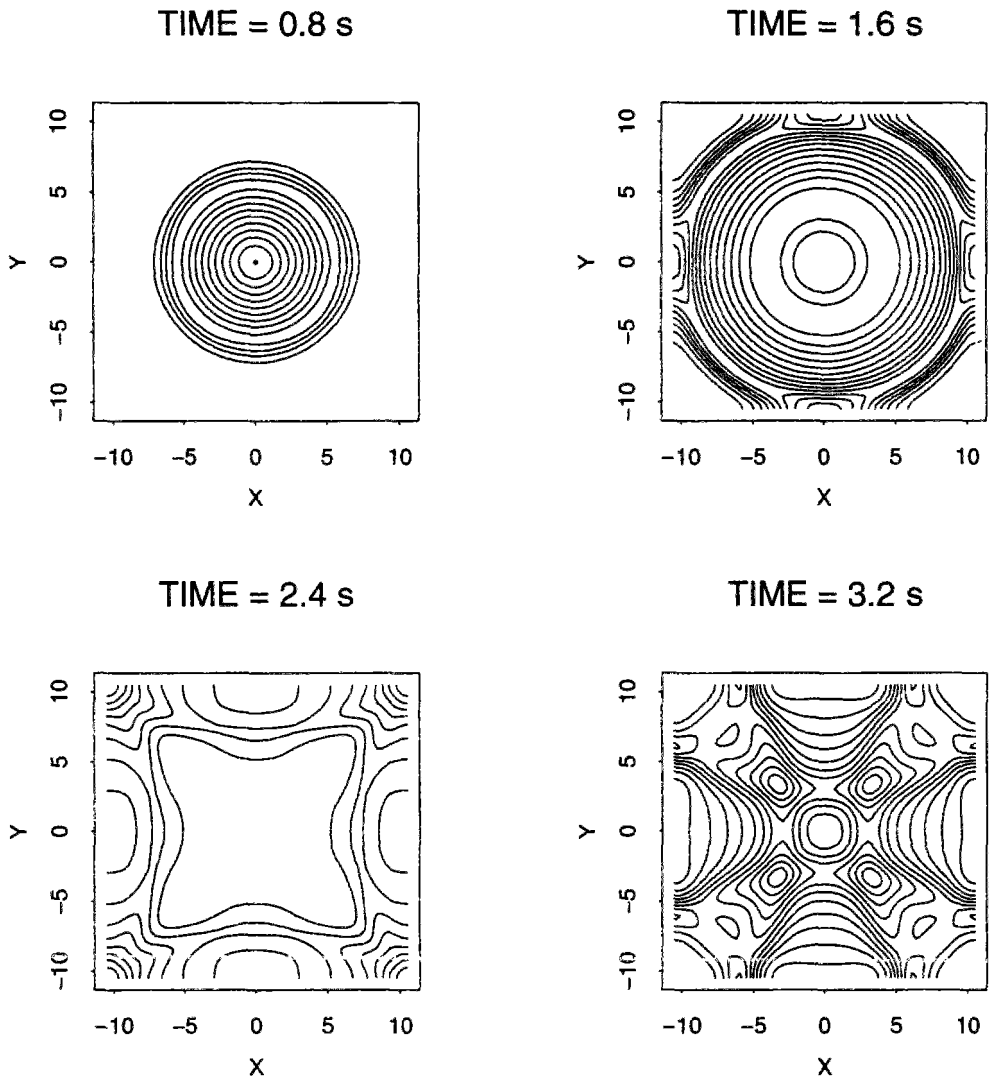


Figure G.10: Water level isolines - "corrected" model QH ($\Delta t = 0.04, \gamma = 0.6$)

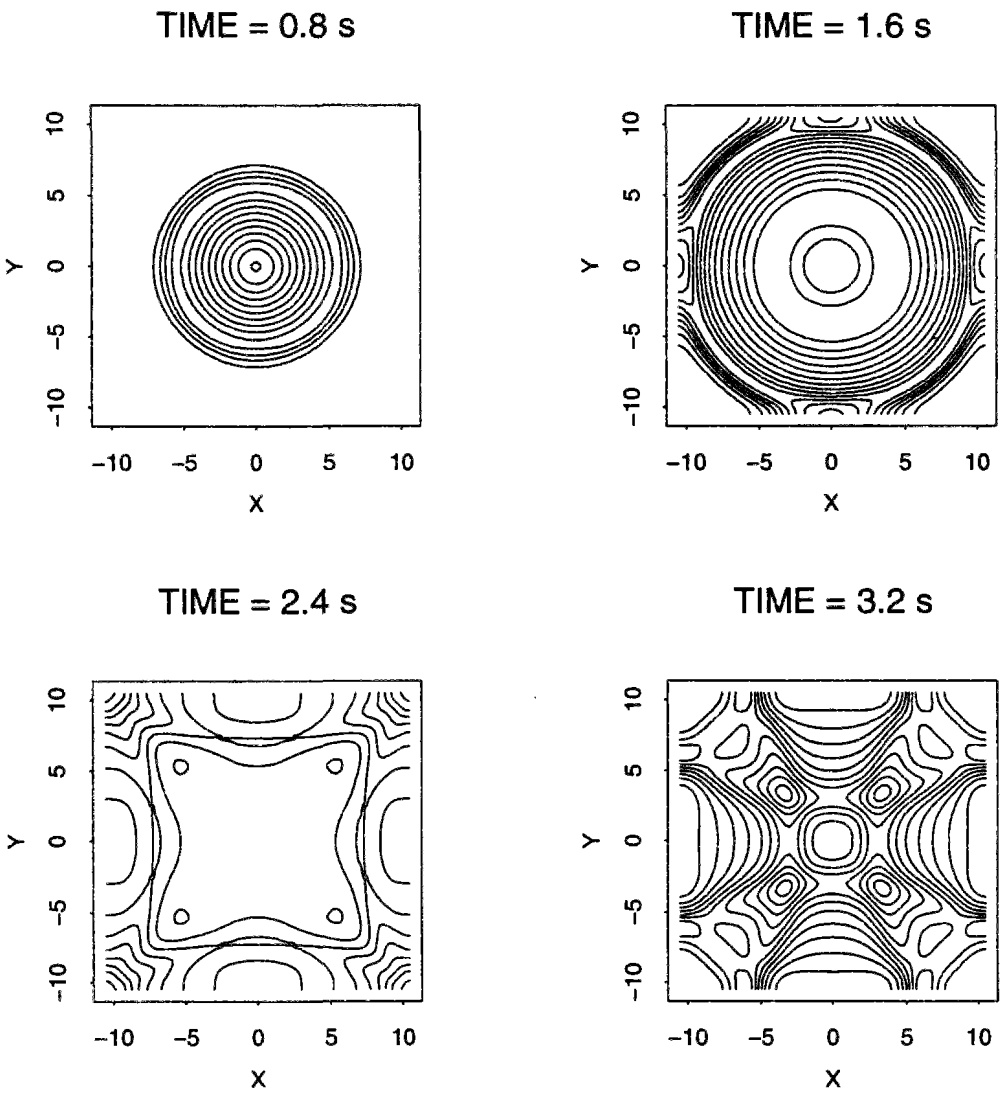


Figure G.11: Water level isolines - model UH ($\Delta t = 0.04$, $\gamma = 0.6$)

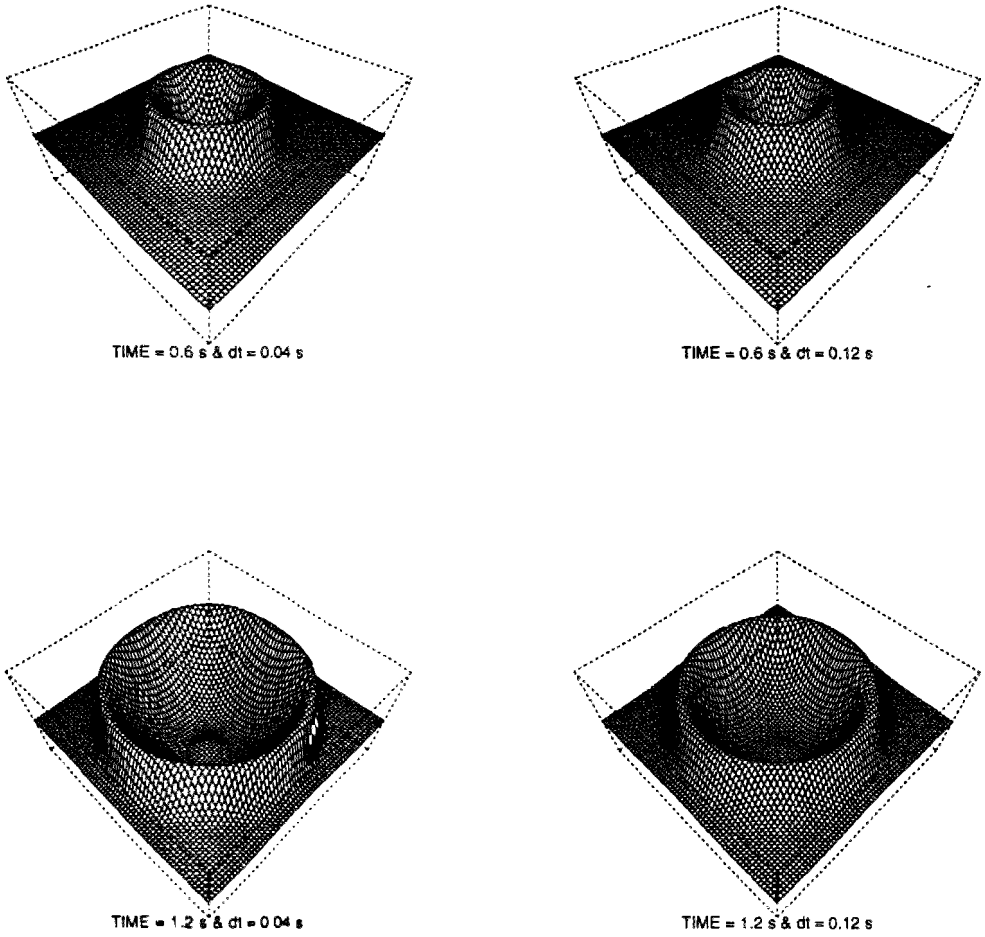


Figure G.12: Comparison of surface profiles computed with $\Delta t = 0.04$ and 0.12 s at $t = 0.6$ & 1.2 s ($\gamma = 0.6$)

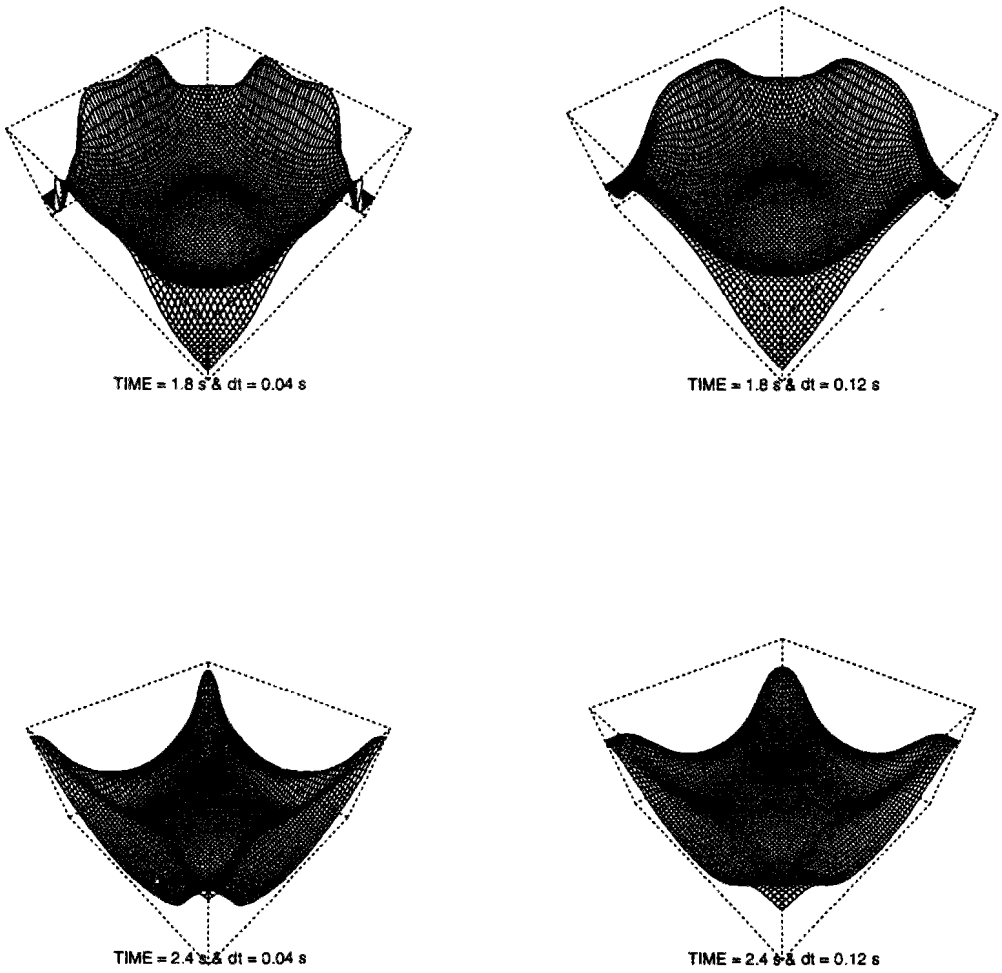


Figure G.13: Comparison of surface profiles computed with $\Delta t = 0.04$ and 0.12 s at $t = 1.8$ & 2.4 s ($\gamma = 0.6$)

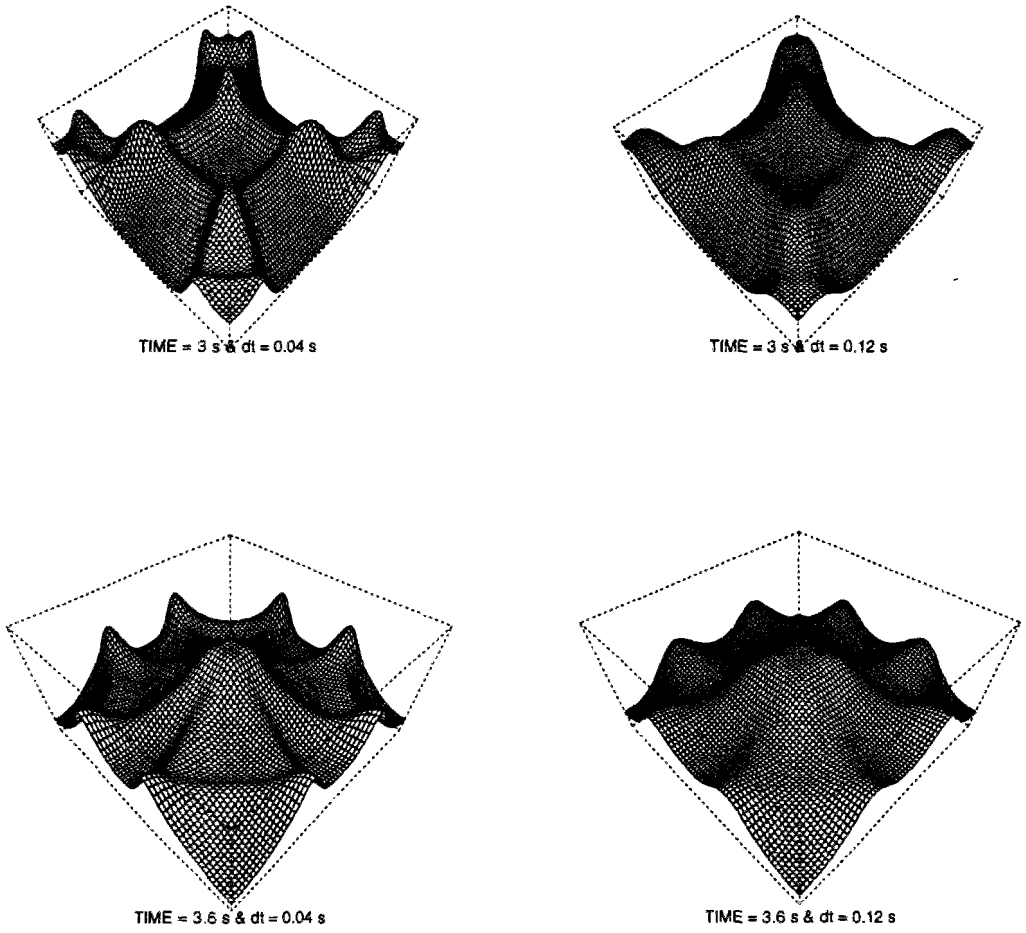


Figure G.14: Comparison of surface profiles computed with $\Delta t = 0.04$ and 0.12 s at $t = 3.0$ & 3.6 s ($\gamma = 0.6$)

G.4 Tide in a rectangular harbour

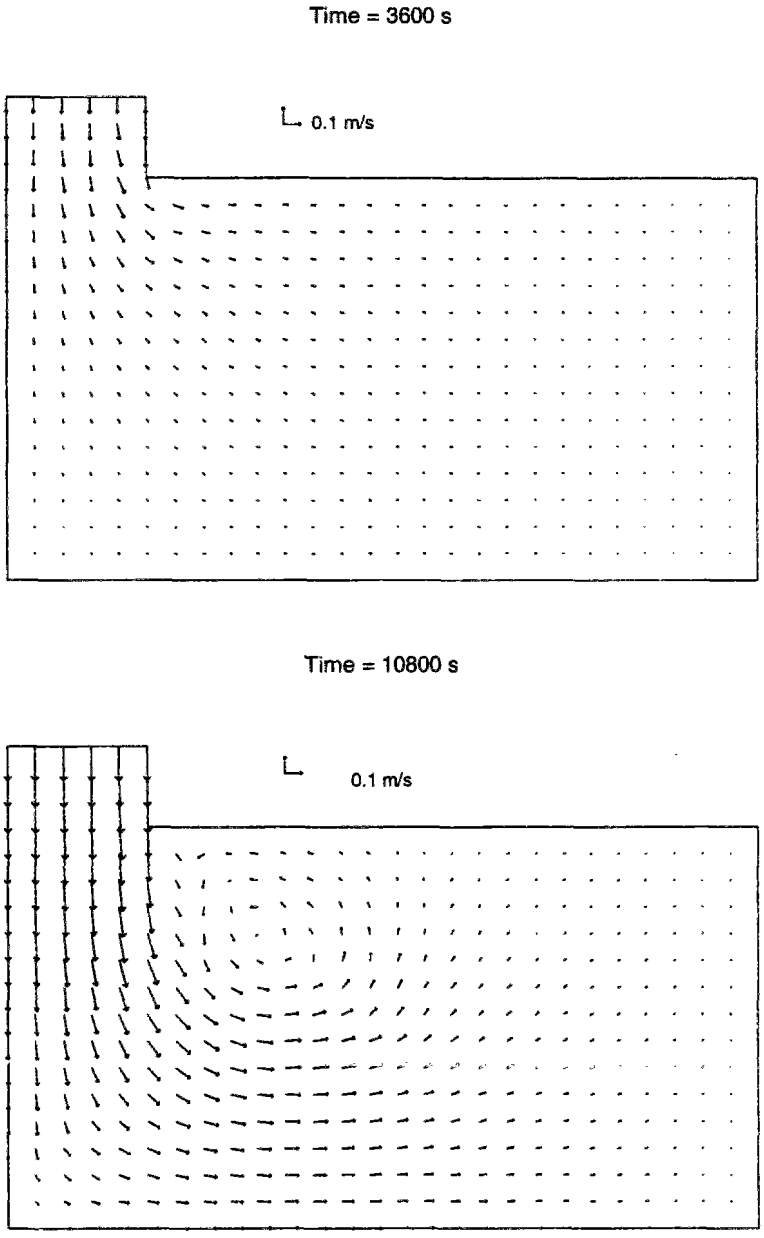


Figure G.15: Velocity field in the harbour after $T = 1$ and 3 h

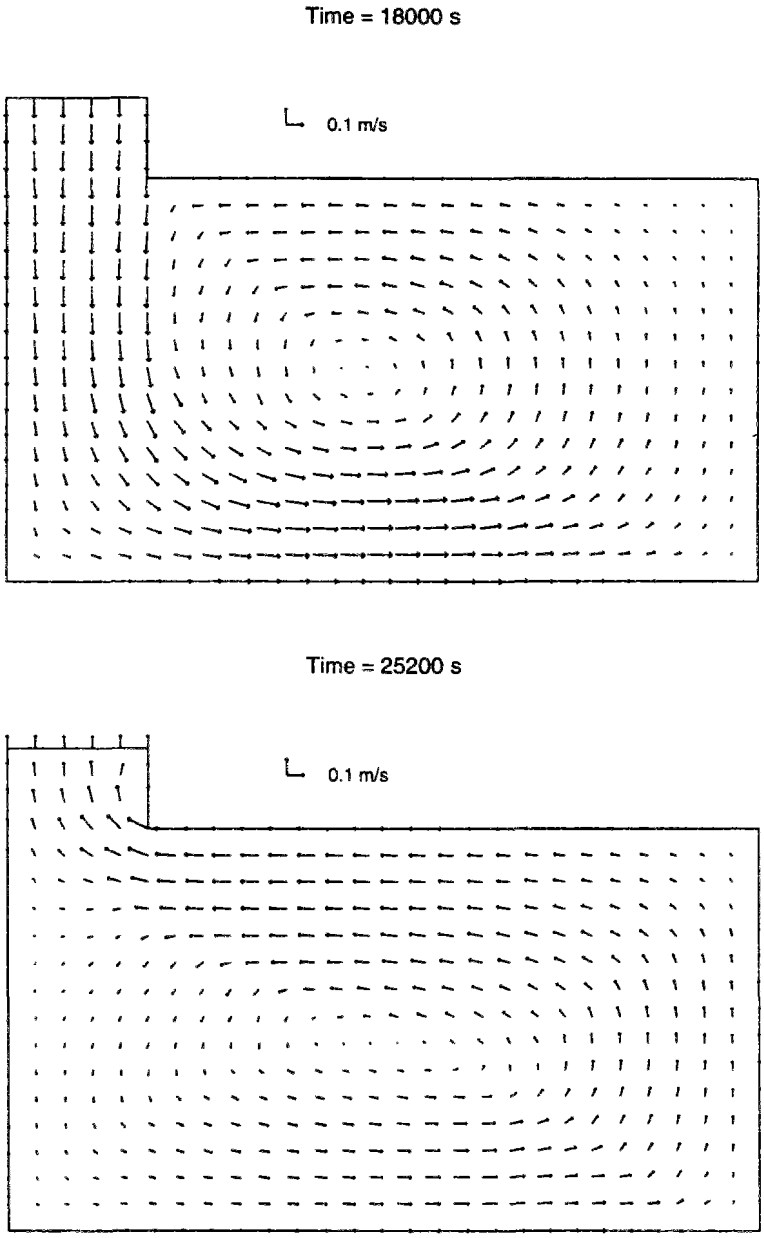


Figure G.16: Velocity field in the harbour after $T = 5$ and 7 h

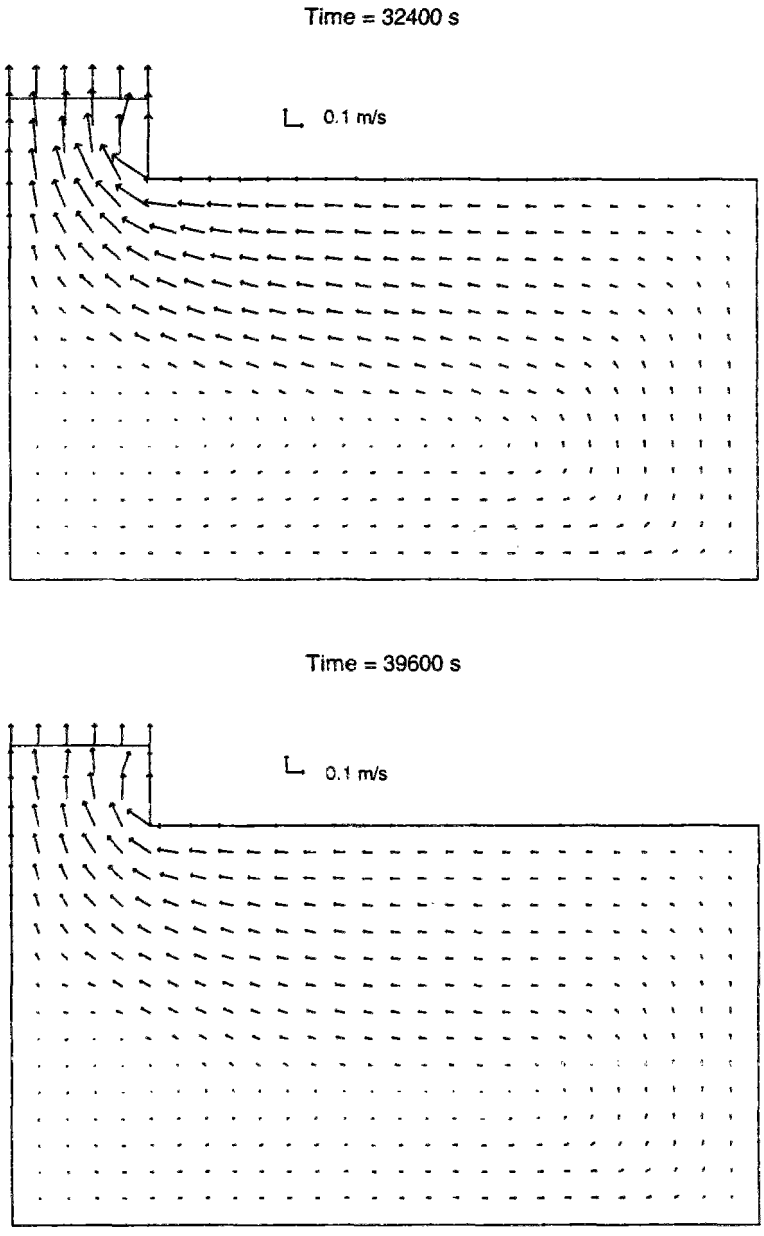


Figure G.17: Velocity field in the harbour after $T = 9$ and 11 h

G.5 Separating flow in an expanding flume

G.5.1 Supplementary informations about the experiment

Eddy viscosity The following table, extracted from (Stelling & Wang, 1984), gives the detailed estimation of eddy viscosities performed by the authors.

Table G.5: Eddy viscosities in the mixing layer ($y = 0.4$m) (unit : 10^{-4} m^2.s^{-1})

	X coordinate (m)							
	1.03	1.20	1.40	1.65	1.90	2.15	2.65	3.65
10 s	2.0	1.4						
20 s	2.1	2.4	3.4					
30 s	1.4	1.4	5.4	1.3	5.0			
40 s	1.6	1.9	4.1	9.3	6.8	6.0	3.0	2.0
50 s	1.3	1.9	2.8	4.9	9.1	5.9	2.6	
60 s	1.2	1.4	3.1	4.4	4.4	7.5	5.0	1.7
70 s	0.8	1.4	2.5	2.7	2.6	2.5	4.3	1.3
80 s			1.4			2.0	2.0	1.2

Experimental scheme Points where the velocity was measured are located at the intersection of dashed lines on figure G.18. Velocities were monitored at 5 to 7 depths per vertical using a laser-doppler velocitymeter. After a few trials, consisting of the study of vertical velocity profiles at different locations within the flume, depth-averaged velocities were calculated by taking the average of velocities recorded at 1.6 and 7.6 cm above the flume bottom. (Stelling & Wang, 1984) consider that this method allows to approximate the exact depth-averaged velocity with a ten percent error margin.

Bold lines denote the two cross sections where water depth was surveyed, using a wave-height meter. They are located respectively 0.225 m and 0.825 m downstream the flume widening.

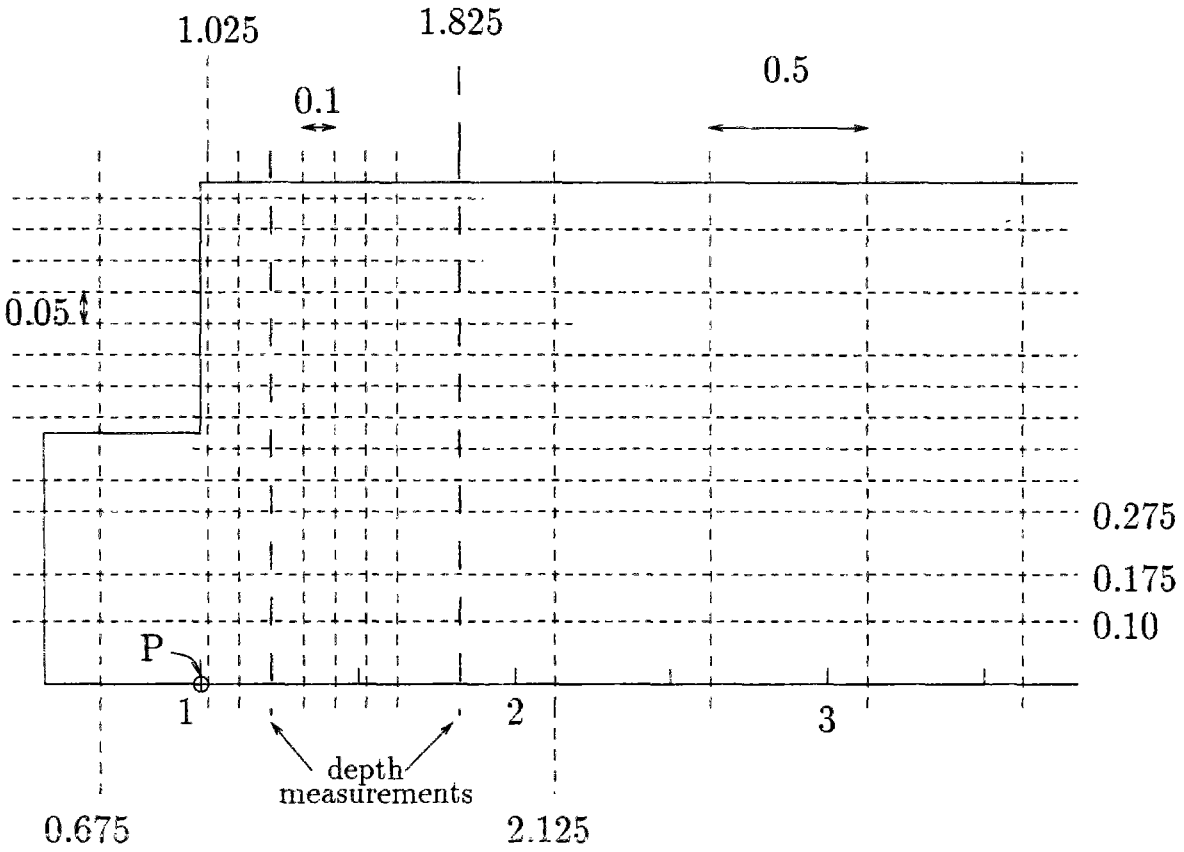


Figure G.18: Location of measuring points in experimental flume

Stelling & Wang results The following table gives the results of numerical experiments Stelling & Wang reported as being the most satisfying.

Table G.6: Eddy characteristics in expanding flume : experimental and Stelling & Wang results

Time (s)	15	25	35	45	55	65
Measurements						
1^{st} eddy (x_c, y_c)	(0.33,0.58)	(0.58,0.58)	(0.96,0.58)	(1.38,0.58)	(1.88,0.58)	(1.90,0.56)
2^{nd} eddy (x_c, y_c)				(0.53,0.50)	(0.68,0.48)	(0.75,0.45)
L_s (m)	0.70	1.40	1.90	2.30	2.40	2.60
Main flow vel. (m/s)	0.29	0.39	0.37	0.33	0.26	0.12
Max. recirculation vel.	0.17	0.24	0.22	0.15	0.15	0.16
Partial-slip condition $\alpha = 0.75$ - Stelling & Wang model						
1^{st} eddy (x_c, y_c)	(0.36,0.59)	(0.66,0.67)	(0.96,0.61)	(1.15,0.58)	(1.25,0.56)	(1.47,0.54)
2^{nd} eddy (x_c, y_c)				(0.50,0.45)	(0.16,0.45)	(0.60,0.50)
L_s (m)	0.71	1.33	1.72	1.92	2.10	2.12
Main flow vel. (m/s)	0.30	0.40	0.37	0.35	0.25	0.17
Max. recirculation vel.	0.11	0.23	0.25	0.25	0.25	0.24
Partial-slip condition $\alpha = 0.90$ - Stelling & Wang model						
1^{st} eddy (x_c, y_c)	(0.36,0.59)	(0.72,0.61)	(0.96,0.59)	(1.17,0.57)	(1.17,0.54)	(1.27,0.49)
2^{nd} eddy (x_c, y_c)					(0.15,0.47)	(0.17,0.47)
L_s (m)	0.67	1.27	1.85	2.22	2.50	2.57
Main flow vel. (m/s)	0.30	0.39	0.37	0.35	0.25	0.17
Max. recirculation vel.	0.11	0.24	0.25	0.28	0.26	0.24

G.5.2 Influence of the treatment of advective terms

This section provides an example of the sensitivity of model forecasts to the solution of the advection step. The boundary condition applied on side walls is no slip and the assumed eddy viscosity amounts to $2.3\,10^{-4}$ m^2.s^{-1}.

Figures G.19 and G.20 display the flow patterns computed when the advection step is solved by the characteristic method used in combination with the bilinear interpolator. At first glance, these velocity fields appear immediatly to be drastically different from those obtained when a bicubic interpolator is applied (figures 10.21 and 10.22, section 10.3.2.1).

Table G.7 sums up the main features of both forecasts. We recall that the location of eddy

centers is given with respect to point P, whose coordinates are $(x, y) = (1, 0)$. The reference point for comparing velocity forecasts in the main flow has coordinates $(x, y) = (1.025, 0.275)$ in the original reference system, that is to say it is located 2.5 cm downstream and 27.5 cm on the left of point P.
Due to the scarcity of measuring points in the recirculation area as soon as $t = 35$ and especially for $t \geq 45$ s, the "maximum" velocities listed on line "Experimental" are probably not the actual maximum velocities. They nevertheless provide an order of magnitude of the velocities within the main eddy. At all times during the experiment, these velocities appear to be symmetrically distributed. At the contrary, in the numerical applications, during the early stages of eddy development, velocities calculated in the upper half of the eddy, next to the wall, are significantly lower than within the eddy lower half, bordering the main flow. This is probably due to the influence of the no-slip condition. Then, for $t \geq 45$, the distribution of velocity within the eddy becomes more symmetric.

Neither simulation is in close agreement with the observations.

(a) bicubic interpolator As discussed in section 10.3.2.1, the model tends to overestimate the circulations within the concave corner of the flume, their extent, the speed at which they develop. These secondary circulations "push" the main eddy downstream, so that its center is generally misplaced. This also induces that, while the size of the main eddy is underestimated (about 1.2 m instead of 1.7 m), the total length of the recirculation region is greater than observed.

The model slightly overestimates the velocity in the main flow. On the other hand, it tends to underestimate the velocities in the recirculation area during the accelerating phase of the flow. But, on the whole, computed and observed velocities appear to be in the same range.

(a) bilinear interpolator While the velocities estimated by one or another interpolator in the main flow are similar, the use of the bilinear interpolator results in a sharp decrease of the estimated velocities within the recirculation region, notably in the decelerating phase of the flow (30 to 50 % slower). This underestimation may explain why the main eddy fails to develop (it is slightly shorter than with the bicubic interpolator, and thinner) and does not drift downstream sufficiently (so that the recirculation area is underpredicted). Similarly, the secondary eddy is less marked, and less distinguishable from the main eddy.

The weakness of velocities witnessed when using the bilinear interpolator leads to suspect that in that case the solution of the advection step is plagued by numerical diffusivity (nb : when increasing the physical eddy diffusivity, we similarly observ that the recirculation becomes slower). **This numerical diffusion "dampens" the model sensitivity to changes in the physical parameters of the simulation** : for instance, as soon as the slip along the wall is increased

to $\alpha = 0.2$, the secondary eddy fails to develop (cf figure G.21), when, on the other hand, secondary circulations are still observed for $\alpha = 0.75$, the bicubic interpolator being applied (cf next section G.5.3).

This limited numerical experiment underlines the importance of the advective terms in the studied flow and the need for computing them as accurately as possible. As it appears to mask the influence of physical parameters, the spurious diffusivity induced by the use of the bilinear interpolator is plainly unacceptable.

Table G.7: Influence of advection solution on the eddies characteristics ($\varepsilon = 2.3\,10^{-4}$, $\alpha = 0$)

Time (s)	15	25	35	45	55	65
Location of main eddy center (x_c, y_c) (from P)						
Experimental	(0.33,0.58)	(0.58,0.58)	(0.96,0.58)	(1.38,0.58)	(1.88,0.58)	(1.90,0.56)
Linear interpolator	(0.35,0.58)	(0.75,0.62)	(1.02,0.62)	(1.38,0.62)	(1.45,0.60)	(1.58,0.58)
Cubic interpolator	(0.43,0.58)	(0.82,0.62)	(1.25,0.60)	(1.70,0.62)	(1.85,0.60)	(2.15,0.60)
Location of secondary eddy center (x_c, y_c)						
Experimental				(0.53,0.50)	(0.68,0.48)	(0.75,0.45)
Linear interpolator				(0.70,0.52)	(0.60,0.40)	(0.85,0.50)
Cubic interpolator				(0.82,0.48)	(0.70,0.48)	(1.15,0.48)
Length (L_s) of recirculation area (m)						
Experimental	0.70	1.40	1.90	2.30	2.40	2.60
Linear interpolator	0.72	1.38	1.72	2.00	2.15	2.38
Cubic interpolator	0.80	1.55	2.10	2.25	2.50	2.70
Velocity (m/s) in the main flow at point (0.025, 0.275) (from P)						
Experimental	0.290	0.385	0.369	0.331	0.260	0.124
Linear interpolator	0.307	0.414	0.422	0.354	0.261	0.157
Cubic interpolator	0.311	0.414	0.421	0.358	0.259	0.168
Maximum velocity (m/s) in recirculation region						
Experimental	0.17	0.24	0.22	0.15	0.15	0.16
Linear interpolator				0.12	0.12	0.10
eddy upper half	0.06	0.11	0.13			
eddy lower half	0.12 - 0.16	0.18 - 0.22	0.18 - 0.22			
Cubic interpolator				0.18	0.16	0.14
eddy upper half	0.10	0.14	0.16			
eddy lower half	0.17 - 0.18	0.22 - 0.24	0.20 - 0.22			

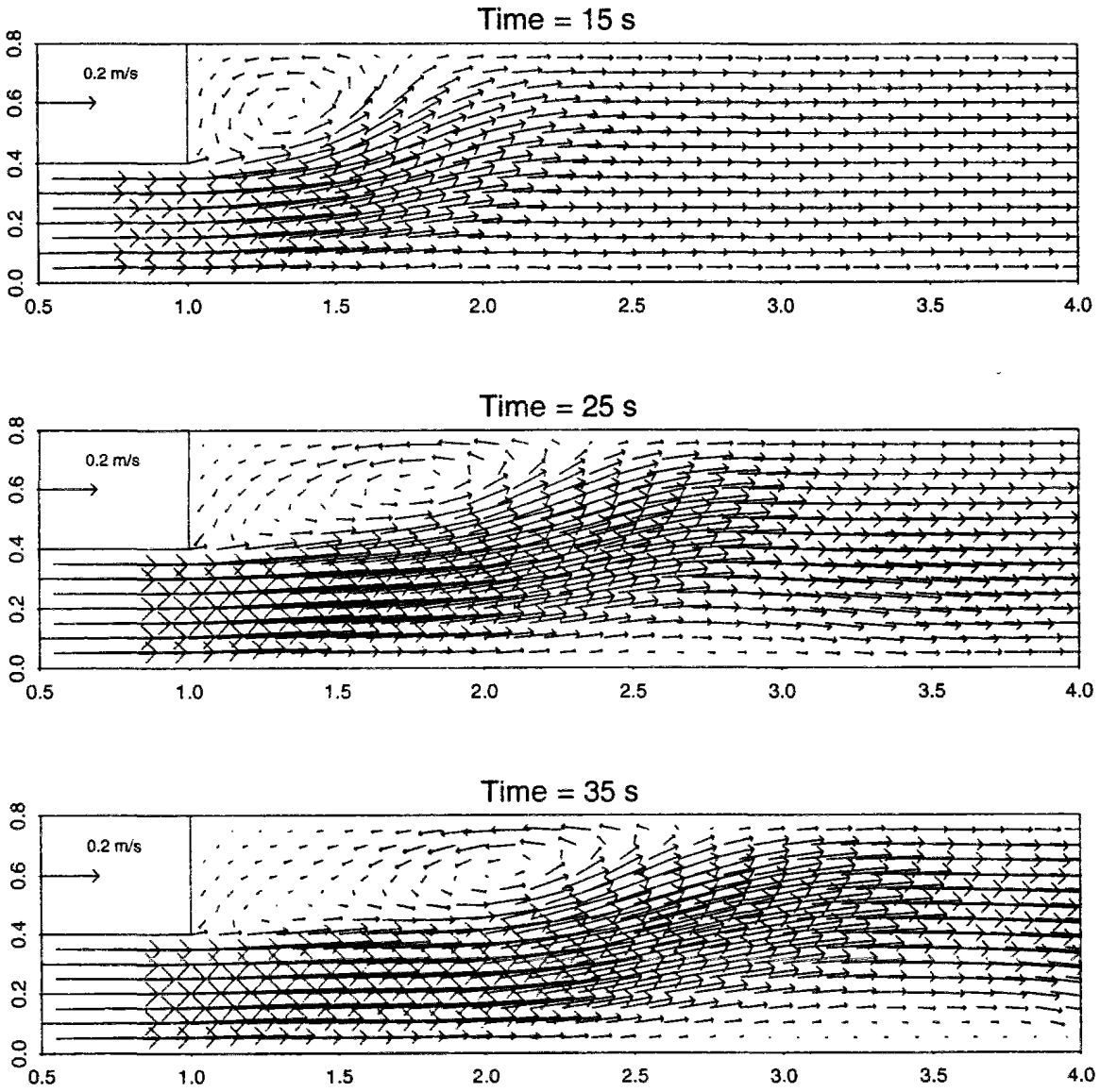


Figure G.19: Velocity field calculated at $t = 15$, 25 and 35 s (no-slip / $\varepsilon = 2.3\,10^{-4}$ / linear interpolator)

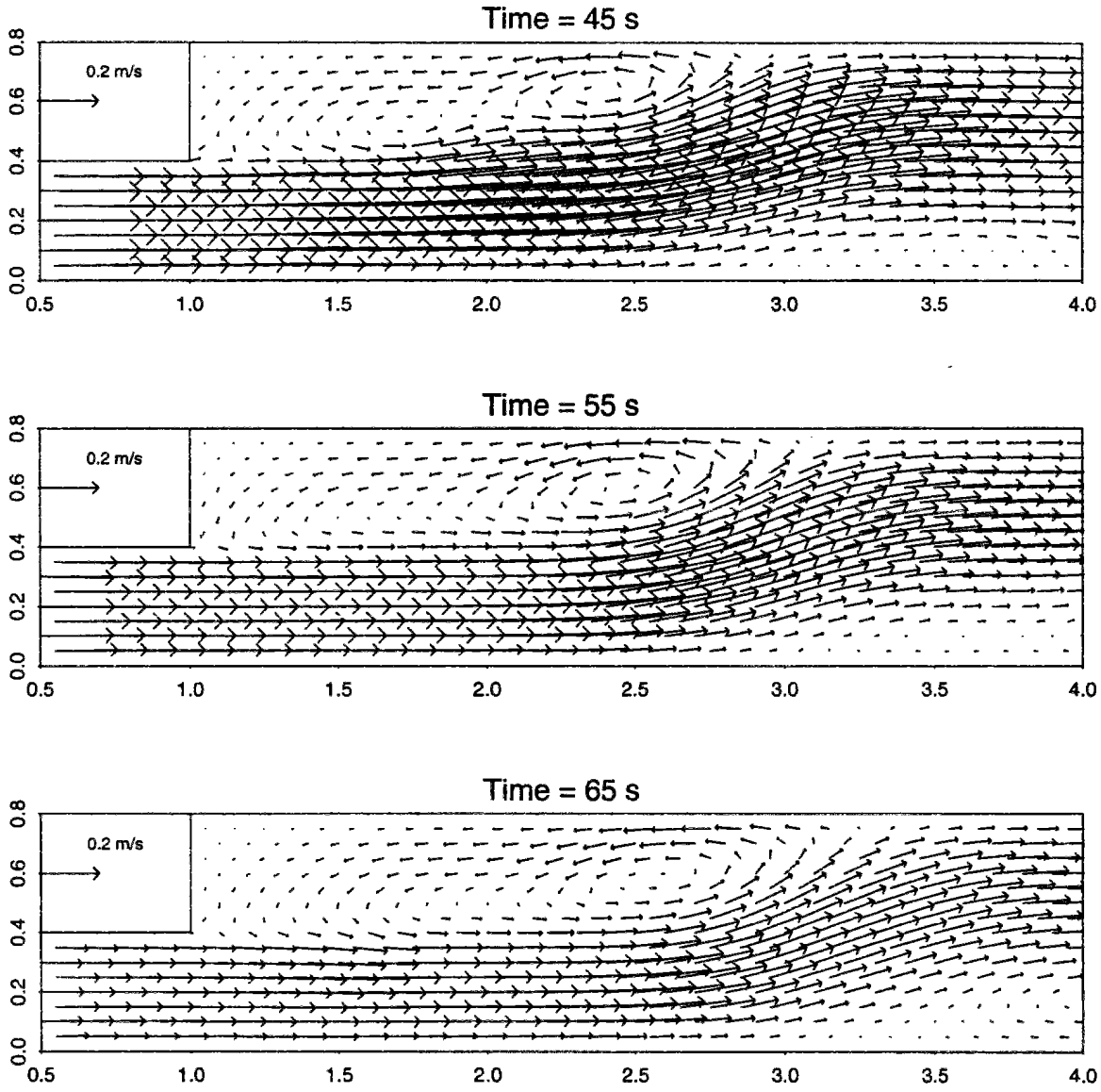


Figure G.20: Velocity field calculated at $t = 45$, 55 and 65 s (no-slip / $\varepsilon = 2.3\,10^{-4}$ / linear interpolator)

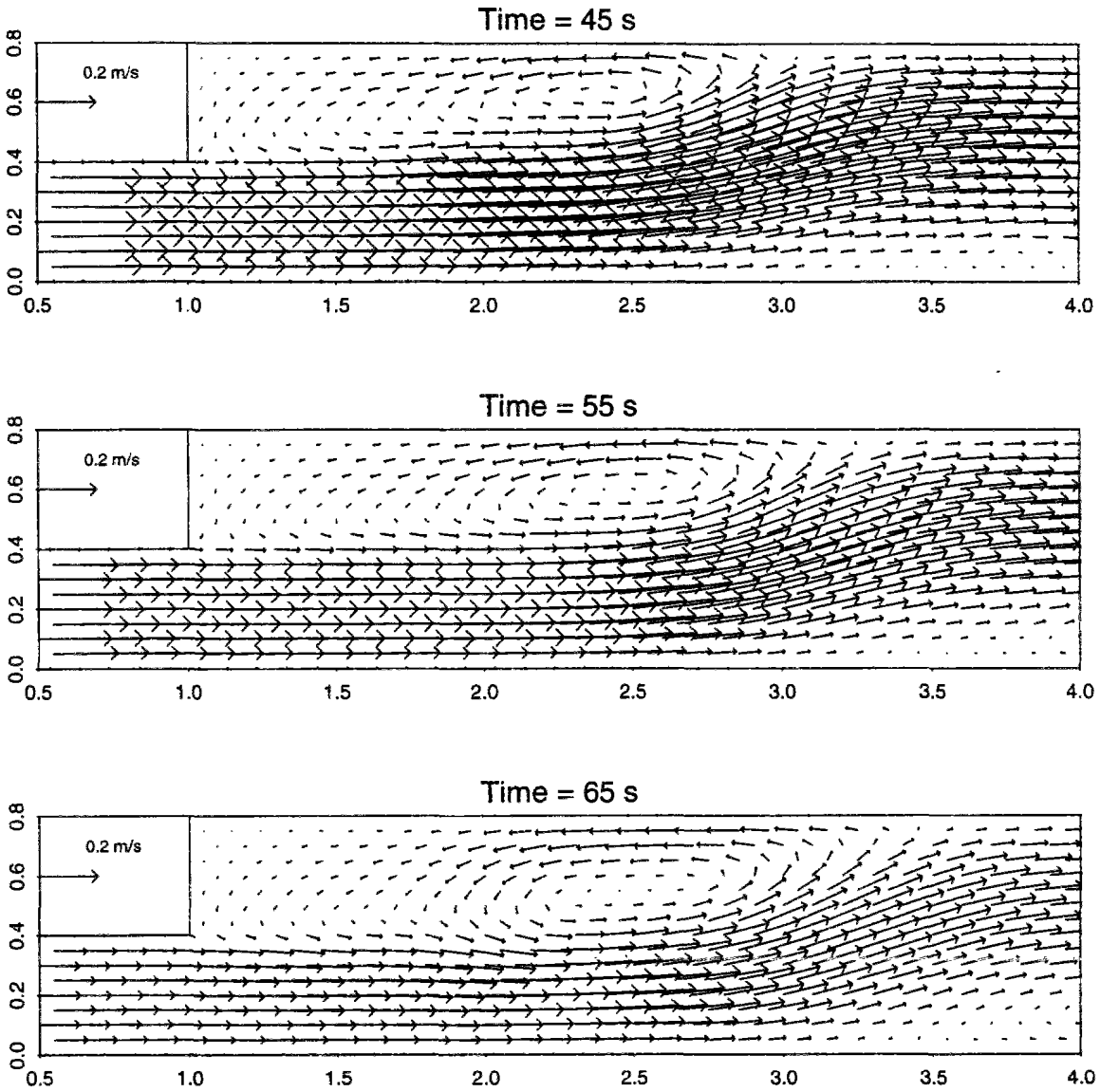


Figure G.21: Partial slip condition ($\alpha = 0.2$) with bilinear interpolator : velocity field at $t = 45$, 55 and 65 s

G.5.3 Influence of slip condition at side walls

Table G.8: Influence of slip conditions on velocities ($\varepsilon = 2.3\,10^{-4}$)

Time (s)	15	25	35	45	55	65
Maximum velocity (m/s) in recirculation region						
Experimental	0.17	0.24	0.22	0.15	0.15	0.16
$\alpha = 0.00$				0.18	0.16	0.14
eddy upper half	0.10	0.14	0.16			
eddy lower half	0.17 - 0.18	0.22 - 0.24	0.20 - 0.22			
$\alpha = 0.50$				0.18	0.19	0.17
eddy upper half	0.09	0.15	0.18			
eddy lower half	0.17 - 0.19	0.23 - 0.26	0.20 - 0.23			
$\alpha = 0.75$				0.21	0.19	0.21
eddy upper half	0.10	0.16	0.20			
eddy lower half	0.20 - 0.21	0.22 - 0.24	0.21 - 0.22			
$\alpha = 1.00$			0.24		0.28	0.27
eddy upper half	0.13	0.19		0.25		
eddy lower half	0.20 - 0.22	0.24 - 0.28		0.30		
Velocity (m/s) in the main flow at point $(0.025, 0.275)$ (from P)						
Experimental	0.290	0.385	0.369	0.331	0.260	0.124
$\alpha = 0.00$	0.311	0.414	0.421	0.358	0.259	0.168
$\alpha = 0.50$	0.295	0.388	0.396	0.326	0.251	0.170
$\alpha = 0.75$	0.287	0.379	0.385	0.314	0.262	0.146
$\alpha = 1.00$	0.280	0.370	0.376	0.301	0.260	0.133

Table G.9: Influence of slip conditions on total recirculation length ($\varepsilon = 2.3\,10^{-4}$)

Time (s)	15	25	35	45	55	65
Experimental	0.70	1.40	1.90	2.30	2.40	2.60
$\alpha = 0.00$	0.80	1.55	2.10	2.25	2.50	2.70
$\alpha = 0.50$	0.78	1.50	2.05	2.25	2.40	2.55
$\alpha = 0.75$	0.78	1.45	2.15	2.55	3.00	3.00
$\alpha = 1.00$	0.75	1.35	1.90	2.20	3.05	3.00

Table G.10: Influence of slip conditions on main eddy characteristics ($\varepsilon = 2.3\,10^{-4}$)

Time (s)	15	25	35	45	55	65
Coordinates (x_c, y_c) with respect to P						
Experimental	(0.33,0.58)	(0.58,0.58)	(0.96,0.58)	(1.38,0.58)	(1.88,0.58)	(1.90,0.56)
$\alpha = 0.00$	(0.43,0.58)	(0.82,0.62)	(1.25,0.60)	(1.70,0.62)	(1.85,0.60)	(2.15,0.60)
$\alpha = 0.50$	(0.42,0.60)	(0.72,0.62)	(1.15,0.62)	(1.50,0.62) at $t = 44$, (1.42,0.60)	(1.60,0.62) at $t = 52$, (1.60,0.58)	(1.70,0.58)
$\alpha = 0.75$	(0.40,0.62)	(0.68,0.62)	(1.00,0.60)	2 poles (0.80,0.58) & (1.80,0.60)	2 poles (1.25,0.55) & (1.95,0.58)	(1.50,0.52)
$\alpha = 1.00$	(0.35,0.60)	(0.55,0.62)	(1.00,0.60)	(0.93,0.58)	(1.27,0.55)	(1.08,0.52)
Eddy length						
Experimental	0.70	1.40	1.80	1.70	1.80	1.60
$\alpha = 0.00$	0.80	1.30	1.50	1.10	1.20	1.10
$\alpha = 0.50$	0.78	1.40	1.55	1.45	1.75	1.50
$\alpha = 0.75$	0.78	1.45	1.85	2.20	2.20	2.20
$\alpha = 1.00$	0.75	1.35	1.90	2.20	2.70	2.70

Table G.11: Influence of slip conditions on secondary eddy characteristics ($\varepsilon = 2.3\,10^{-4}$)

Time (s)	45	55	65
Coordinates (x_c, y_c) with respect to P			
Experimental	(0.53,0.50)	(0.68,0.48)	(0.75,0.45)
$\alpha = 0.00$	(0.82,0.48)	(0.70,0.48)	(1.15,0.48)
$\alpha = 0.50$	(0.60,0.48) at $t = 44$, (maximum strength) (0.50,0.48)	(0.22,0.48) at $t = 52$, (maximum strength) (0.58,0.48)	(0.72,0.50)
$\alpha = 0.75$		(0.18,0.48)	(0.35,0.48)
Eddy length			
Experimental	0.7	0.8	1.0
$\alpha = 0.00$	0.8	1.1	1.1
$\alpha = 0.50$	0.55 at $t = 44$, 0.60	0.5 at $t = 52$, 0.70	1.0
$\alpha = 0.75$		0.35	0.5

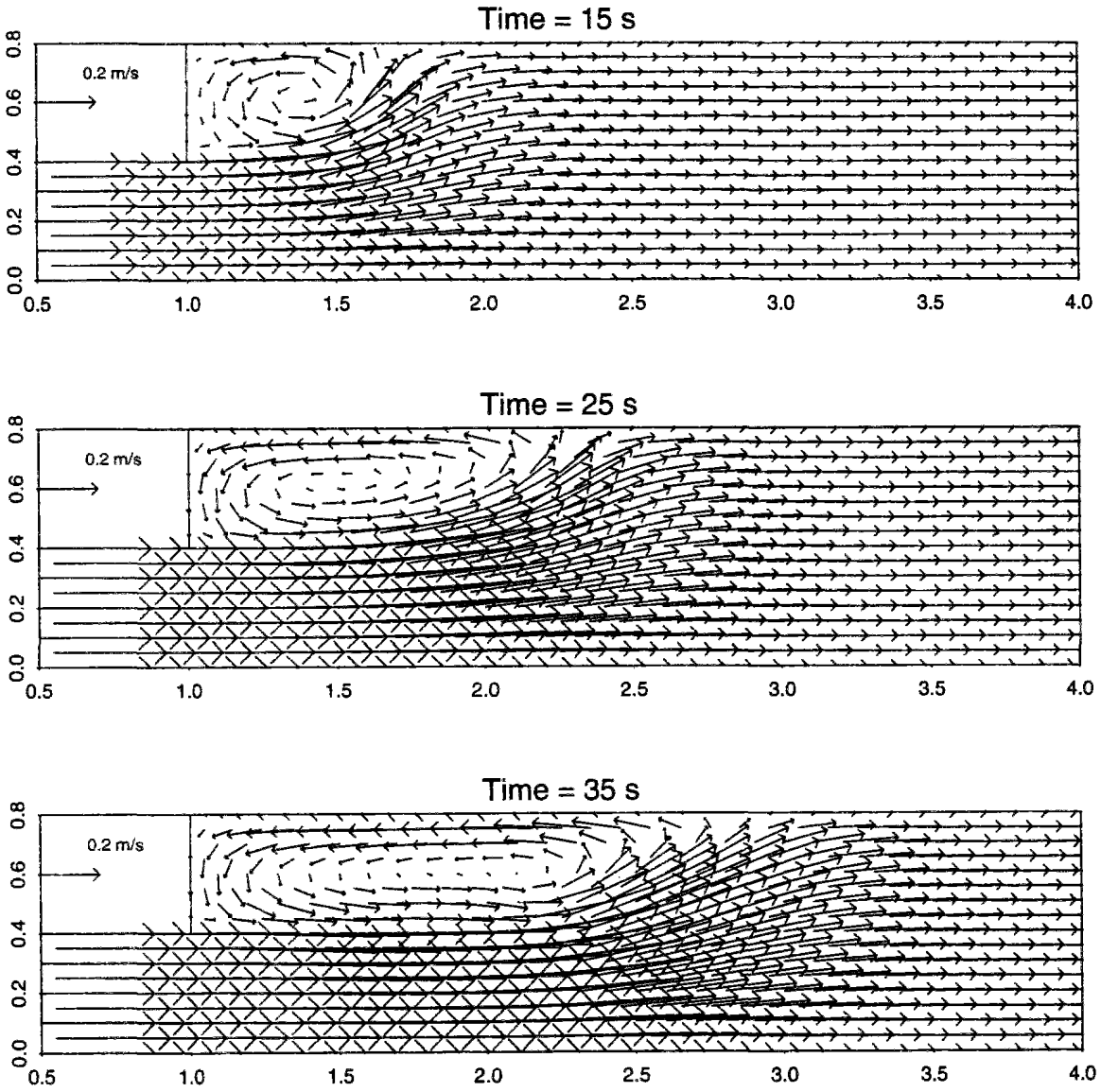


Figure G.22: Perfect slip / $\varepsilon = 2.3\,10^{-4}$ / Velocities at $t = 15, 25, 35$ s

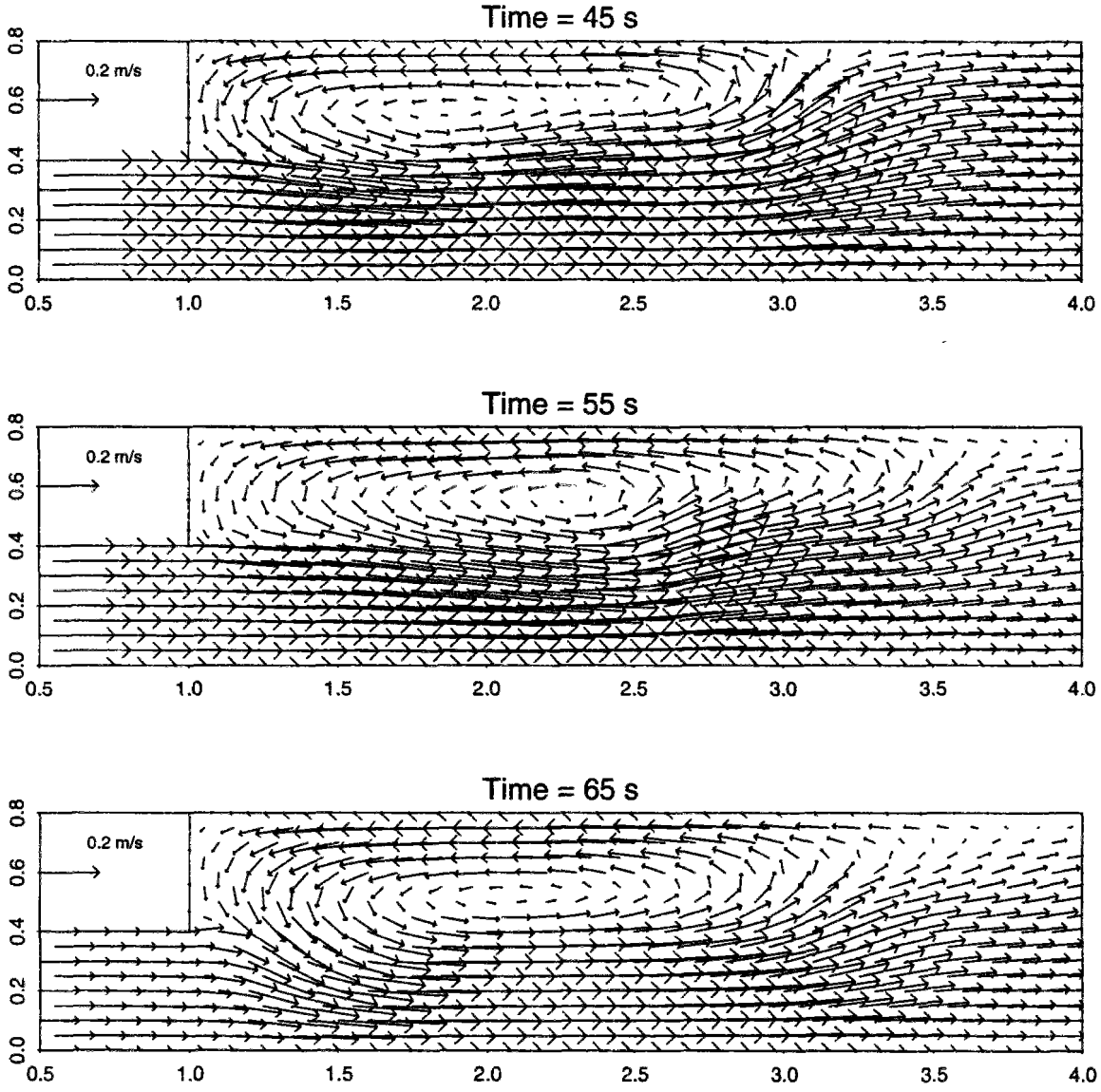


Figure G.23: Perfect slip / $\varepsilon = 2.3\,10^{-4}$ / Velocities at $t = 45, 55, 65$ s

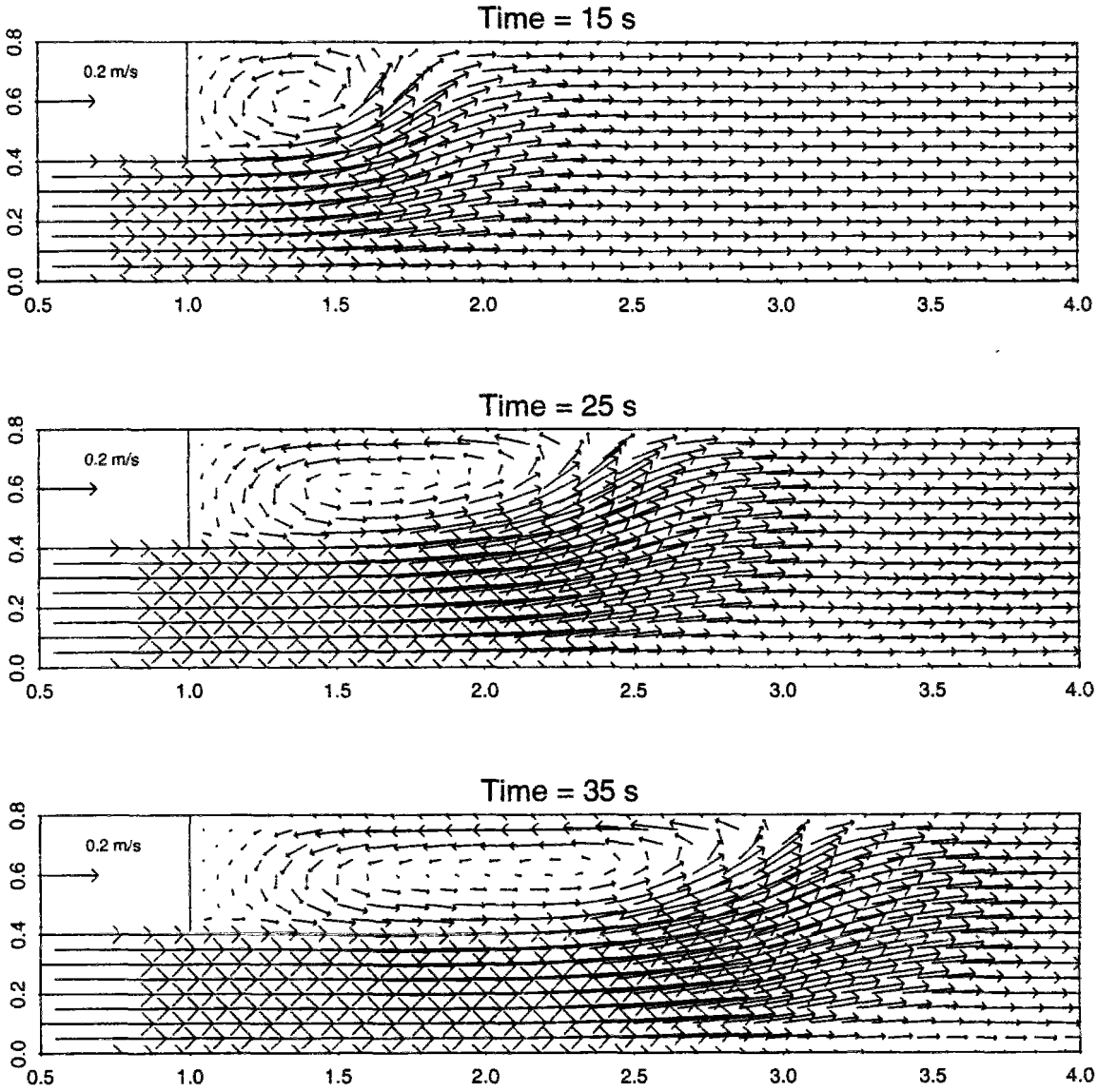


Figure G.24: Partial slip : $\alpha = 0.75$ / $\varepsilon = 2.3\,10^{-4}$ / Velocities at $t = 15, 25, 35$ s

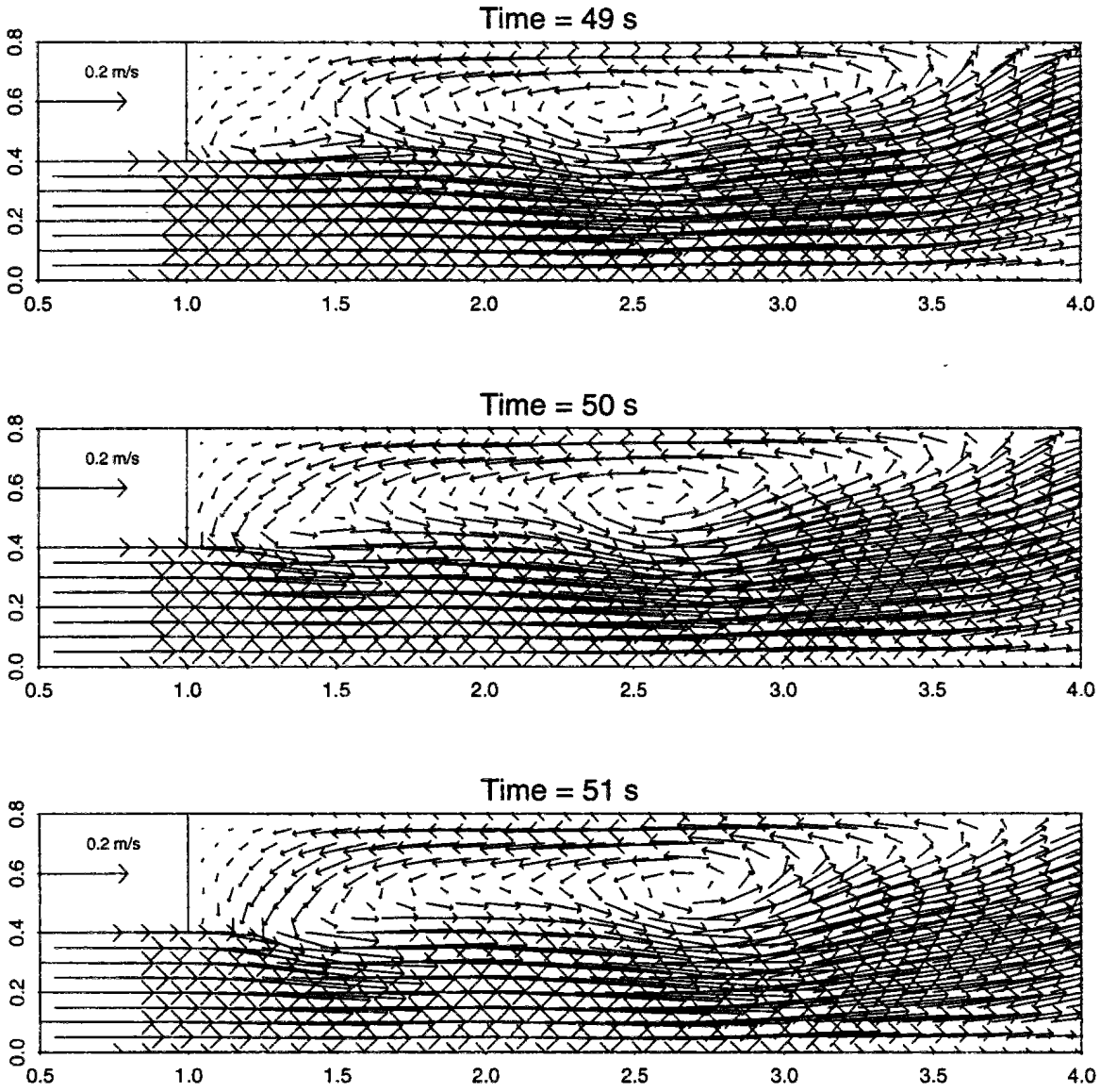


Figure G.25: 1^{st} cycle of secondary circulation development ($\alpha = 0.75$ / $\varepsilon = 2.3\,10^{-4}$) : merging stage

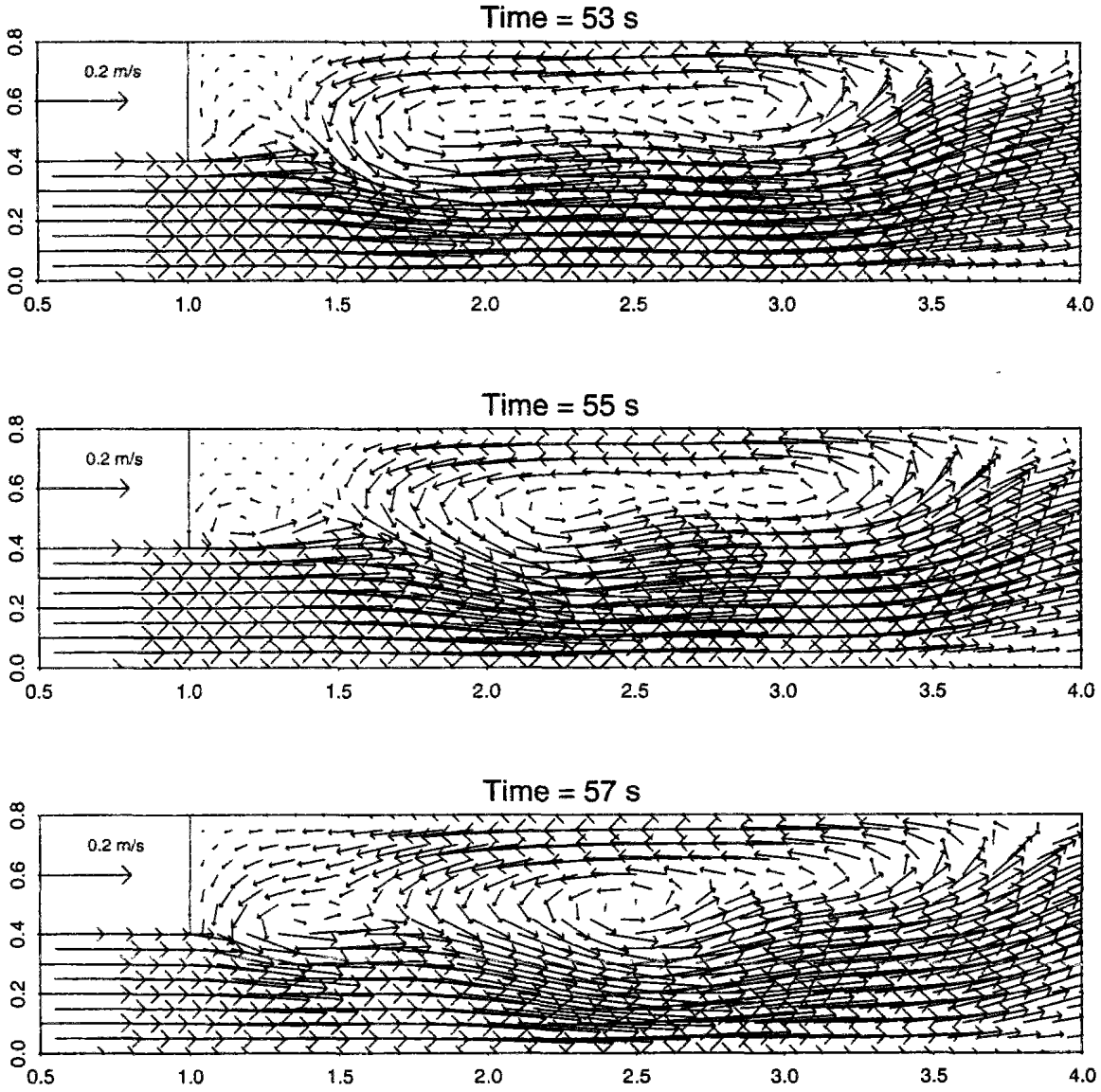


Figure G.26: 2^{nd} cycle of secondary circulation development ($\alpha = 0.75$ / $\varepsilon = 2.3\,10^{-4}$) : growth stage

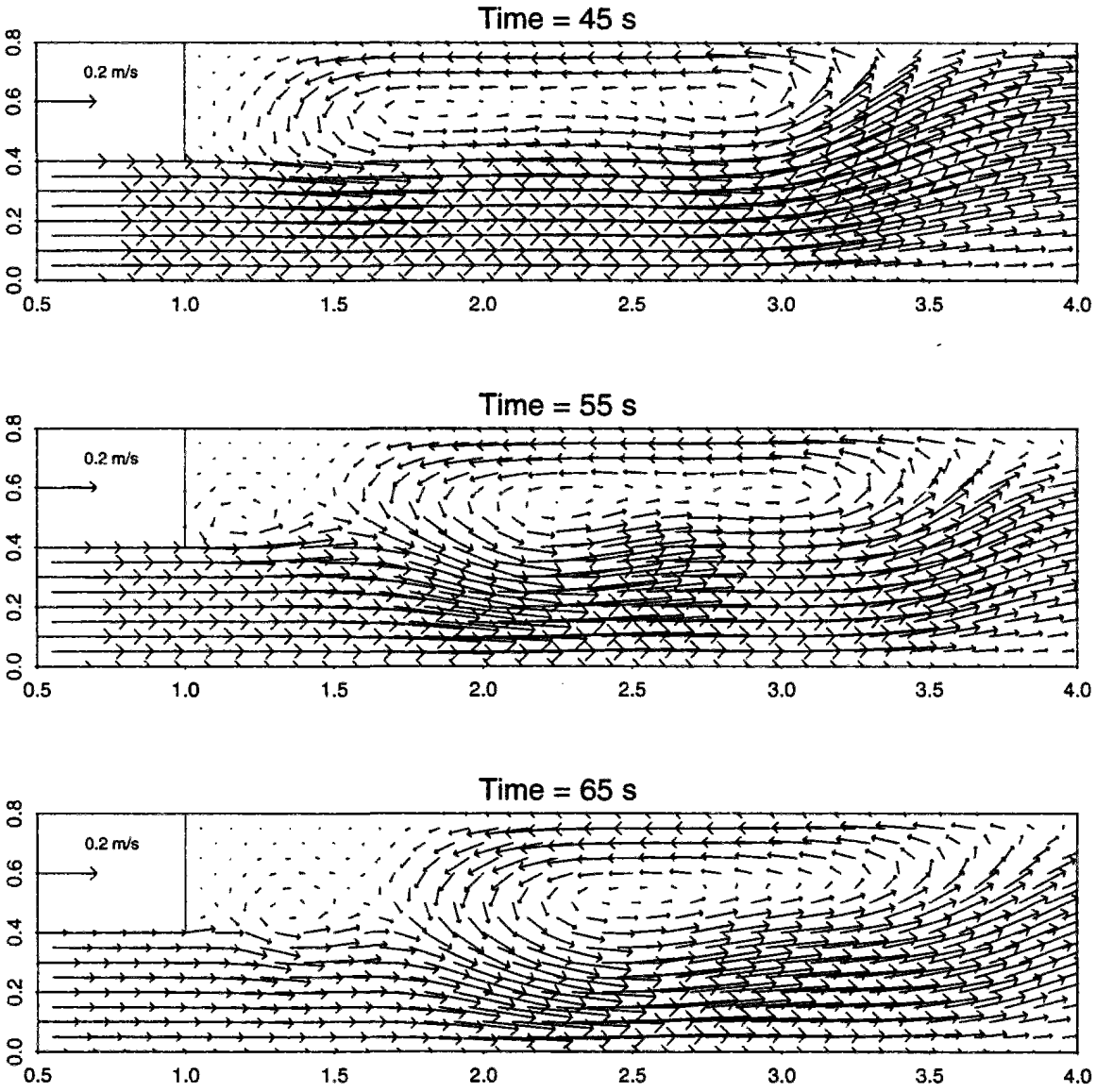


Figure G.27: Partial slip : $\alpha = 0.75$ / $\varepsilon = 2.3\,10^{-4}$ / Velocities at $t = 45$, 55, 65 s

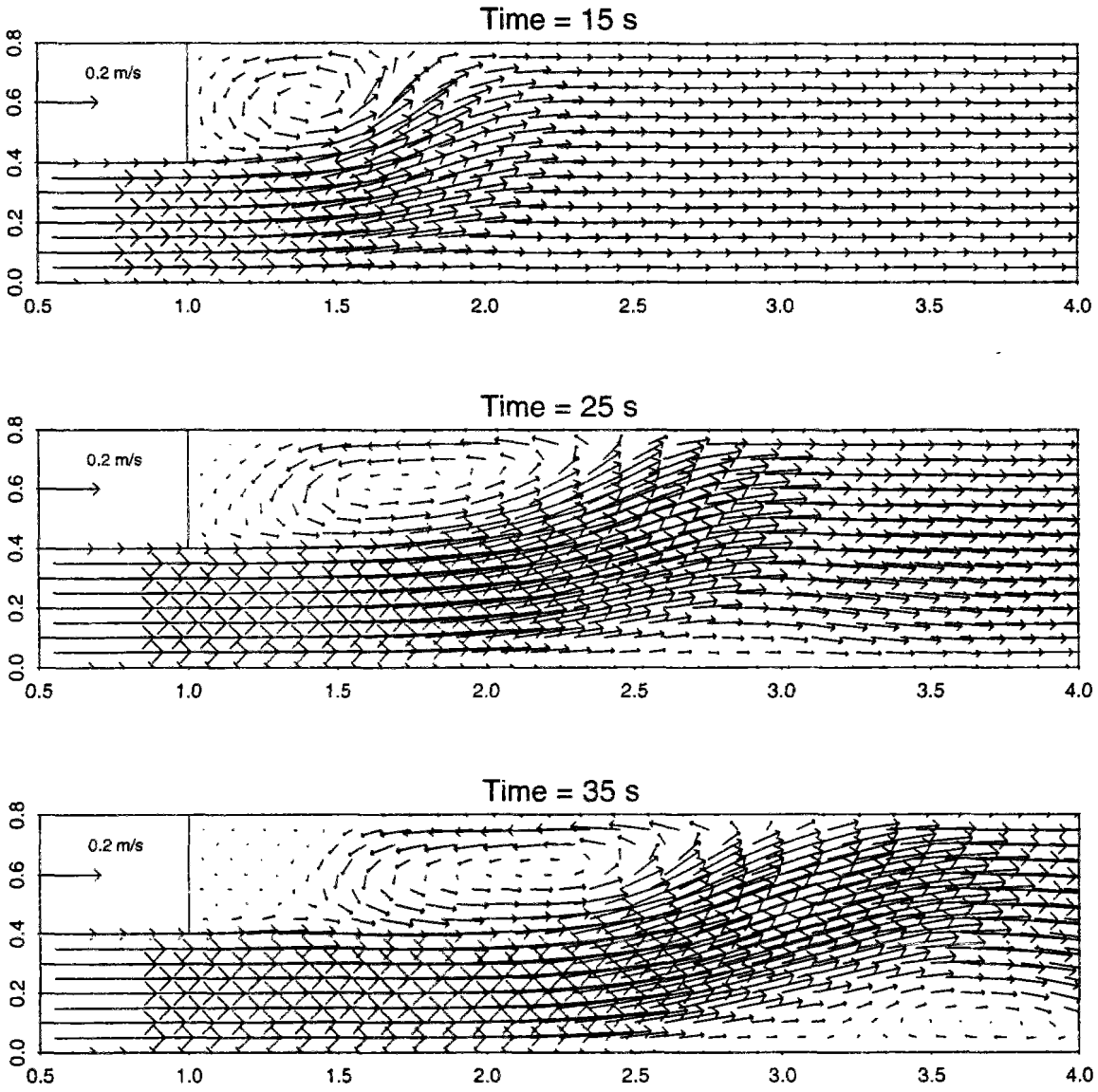


Figure G.28: Partial slip : $\alpha = 0.50$ / $\varepsilon = 2.3\,10^{-4}$ / Velocities at $t = 15, 25, 35$ s

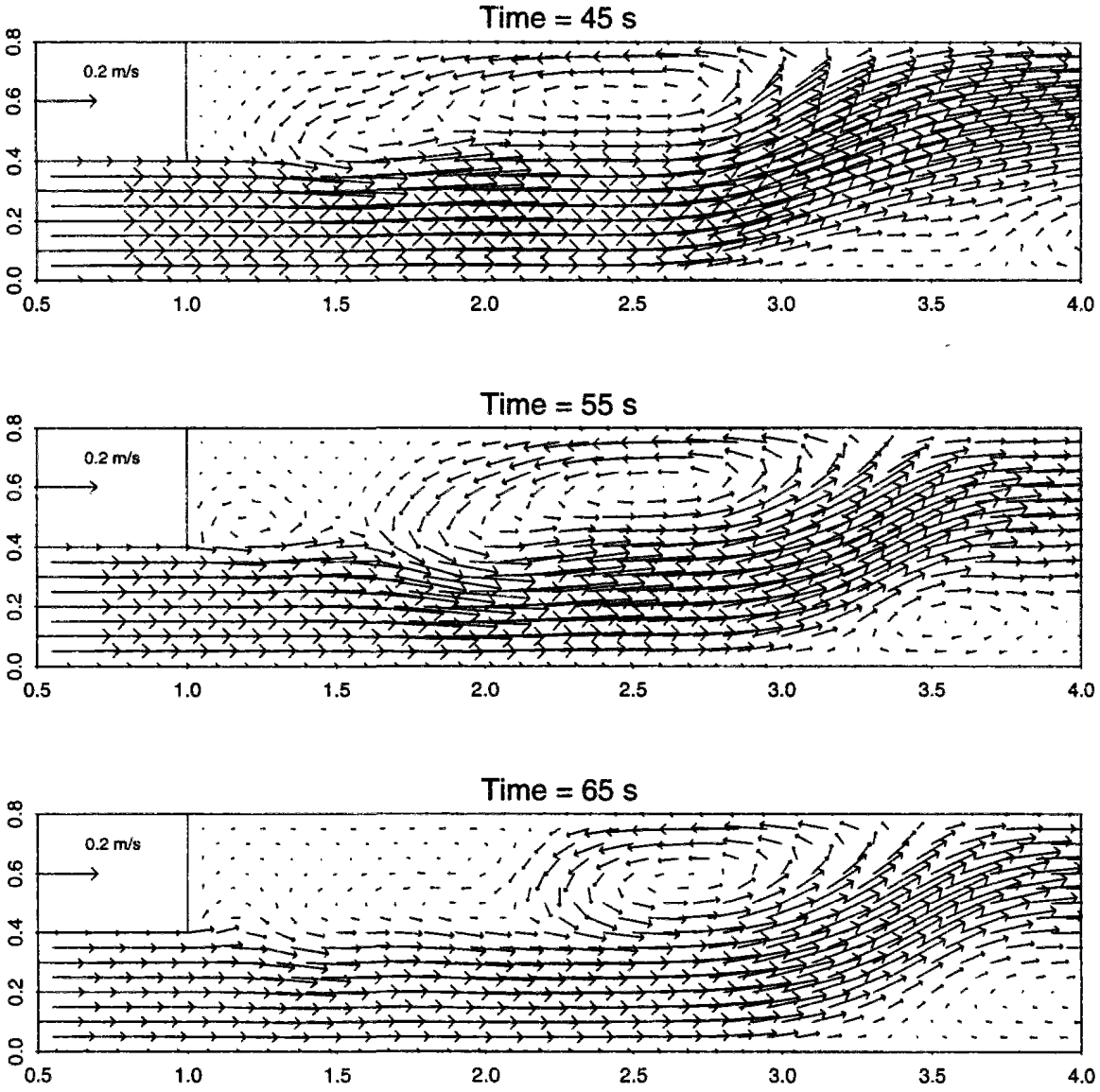


Figure G.29: Partial slip : $\alpha = 0.50$ / $\varepsilon = 2.3\,10^{-4}$ / Velocities at $t = 45$, 55, 65 s

G.5.4 Influence of eddy diffusivity

Table G.12: Influence of diffusivity on eddies characteristics (partial slip $\alpha = 0.5$)

Time (s)	15	25	35	45	55	65
Location of main eddy center (x_c, y_c) (from P)						
Experimental	(0.33,0.58)	(0.58,0.58)	(0.96,0.58)	(1.38,0.58)	(1.88,0.58)	(1.90,0.56)
$\varepsilon = 10^{-4}$	(0.42,0.60)	(0.68,0.62)	(1.02,0.60)	(0.85,0.58)	(1.10,0.55)	(1.25,0.55)
$\varepsilon = 2.3\,10^{-4}$	(0.42,0.60)	(0.72,0.62)	(1.15,0.62)	(1.50,0.62)	(1.60,0.62)	(1.70,0.58)
$\varepsilon = 4.6\,10^{-4}$	(0.40,0.60)	(0.80,0.62)	(1.28,0.60)	(1.58,0.58)	(1.75,0.60)	(1.95,0.58)
$\varepsilon = 10^{-3}$	(0.35,0.60)	(0.65,0.60)	(1.20,0.62)	(1.75,0.62)	(2.20,0.65)	(1.50,0.58)
Location of secondary eddy center (x_c, y_c)						
Experimental				(0.53,0.50)	(0.68,0.48)	(0.75,0.45)
$\varepsilon = 10^{-4}$					(0.15,0.50)	(0.48,0.43)
$\varepsilon = 2.3\,10^{-4}$				(0.60,0.48)	(0.22,0.48)	(0.72,0.50)
$\varepsilon = 4.6\,10^{-4}$				(0.70,0.52)	(0.30,0.50)	(0.65,0.55)
$\varepsilon = 10^{-3}$				(0.78,0.58)	(1.15,0.58)	
Length (L_s) of recirculation area (m)						
Experimental	0.70	1.40	1.90	2.30	2.40	2.60
$\varepsilon = 10^{-4}$	0.80	1.45	1.90	2.10	2.20	2.20
$\varepsilon = 2.3\,10^{-4}$	0.80	1.50	2.05	2.25	2.40	2.55
$\varepsilon = 4.6\,10^{-4}$	0.80	1.55	2.05	2.30	2.60	2.70
$\varepsilon = 10^{-3}$	0.75	1.45	1.95	2.30	2.65	3.00

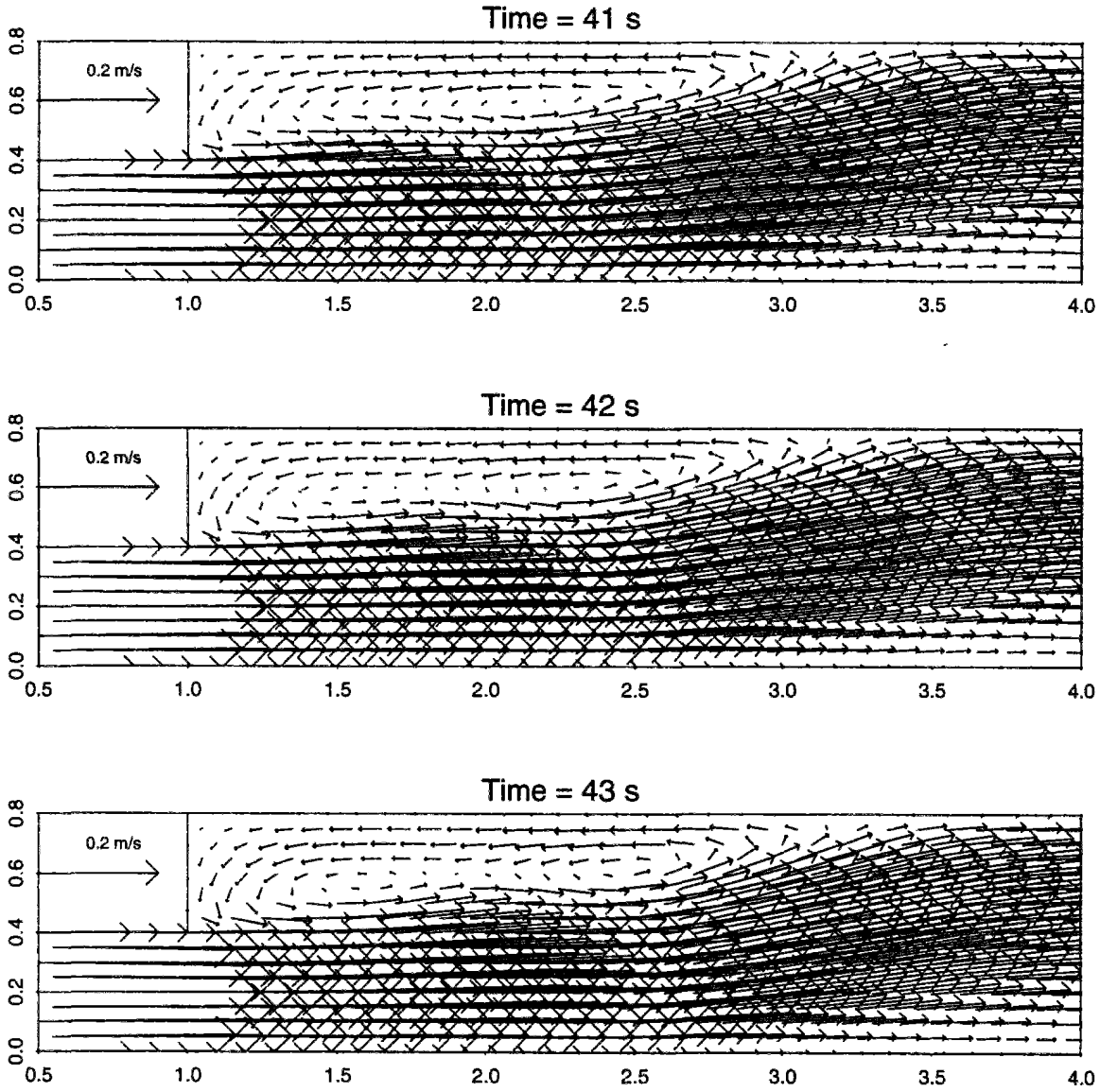


Figure G.30: Main eddy splitting ($\alpha = 0.5, \varepsilon = 10^{-3}$ m^2.s^{-1})

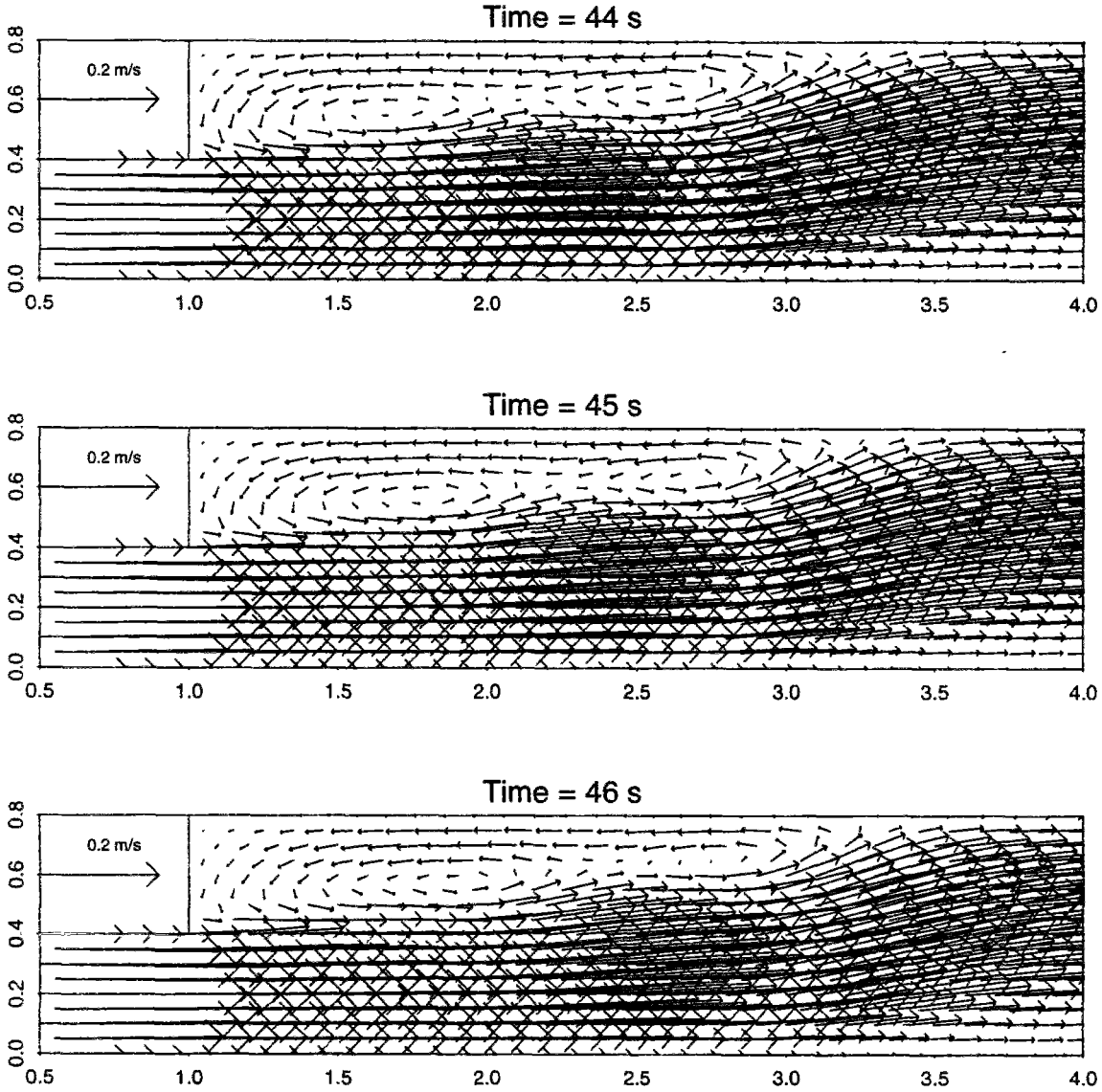


Figure G.31: Evolution of splitted eddy ($\alpha = 0.5, \varepsilon = 10^{-3}$ m^2.s^{-1})

Appendix H

Case study on the Seine River

H.1 Estimation of the Strickler roughness coefficient

H.1.1 Available hydraulic information

In the frame of one-dimensional hydraulic models based on St-Venant equations, the only parameter to calibrate is the Strickler (K_s) or Chezy coefficient (C_h) which controls bed friction. The balance between propagation and friction determines the free surface profile. Consequently, the method followed to estimate Strickler coefficient consists of gathering information about the "typical" surface profiles obtained for different flow rates and of tuning K_s in each case so that model forecasts agree with these observations. Calibrated K_s usually exhibit some dependence with respect to the flow rate. This can be explained by the fact that, as flow changes, the perimeter over which friction applies is modified, both in extent and in nature (e.g. when flow rises and banks covered by vegetation get drowned).

In the case of the Seine River, the first problem is that informations relative to water levels are sparse. Water levels are surveyed with water level recorders located upstream and downstream each navigation lock and under some bridges of Paris. However, complete level hydrographs are not exploited. Elevations are effectively recorded at set times in the day (7 or 8 a.m., eventually also at 5 p.m.) and often after some modification of the dam regulation. These instantaneous evaluations may sometimes be blurred by local perturbations (Even & Poulin, 1993) which have the same order of magnitude (several centimeters) than difference levels between successive dams ! Besides, due the dominant part played by dams in the river hydraulics, there is no unique, one-to-one, relationship between flow rate and free surface slope within a reach.

Nevertheless, S. Even (Even & Poulin, 1993) performed a statistical analysis of 1989 to 1991 data including daily estimations of Seine flow rate and free surface elevations. This allowed both to establish some correlation between free surface slope and discharges and to define more precisely the functioning of the various weirs, spillways and sluice gates on the navigation dams. The data we used for our first round of Strickler calibration (for instance table H.1) are extracted from this study.

H.1.2 Reconstitution of missing bathymetric data

Hydraulic modelling requires the knowledge of river cross sections. Due to the soundings achieved by the Service de la Navigation de la Seine (SNS), the Seine River bed downstream Paris is fairly well known, except for arms which are not open to barge navigation .. as the left arm off St-Denis island.

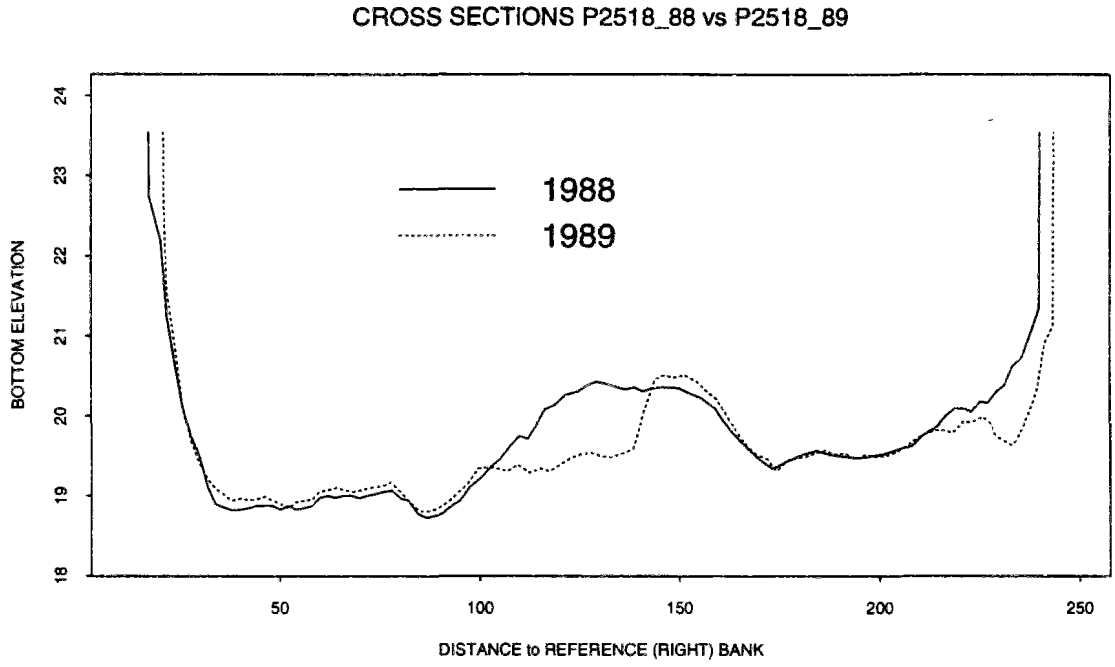


Figure H.1: Seine River transverse profile upstream St-Denis island

Table H.1: Reach Suresnes-Chatou : typical water slopes

Discharge (m^3/s)	"typical" level difference (cm) Suresnes - Chatou
100	5
200	26
300	56
400	87
500	118

We have tried to restore the missing data concerning the left arm according to the following method :

1. The closest *complete* cross sections available at each extremity of the island are respectively profile P33.01 ($\simeq$ 200 m downstream the island) and profile P25.18 ($\simeq$ 50 m upstream). Profile P33.01 has been recorded in 1988, profile P25.18 twice, in 1988 and, after some dredging work in the navigation channel, in 1989. Figure H.1 displays the two soundings. There are but slight differences between them, except in the middle of the section. This corresponds to the aera immediatly upstream the tip of the island, which is subject to rapid silting-up. The average bed elevation in the left half of the cross section (i.e. the beginning of the left arm) is 19.68 m in 1988, 19.60 m in 1989. Profile P33.01 has a nearly trapezoidal shape, average elevation being 18.88 m. As it is located at some distance of the island downstream tip, we can expect this cross section to remain rather stable.

   Assuming that the bed slope in the left arm is uniform, we can determine it from the extreme profiles. We obtain that the slope is $1.2\,10^{-4}$ m/m when refering to the oldest version of profile P25.18 , $1.1\,10^{-4}$ m/m otherwise.

2. Aerial plans of the Seine River allow to determine the river width. The left arm can then be schematized by a succession of reaches with uniform width.

Table H.2: Schematization of the Seine left arm (St-Denis island)

Length (m)	600	300	600	1200	500	600	500	1850	450
Width (m)	75	83	75	90	75	60	75	83	90

3. We assume that the river bathymetry is uniform in each of the above sub-reach. Each will then be characterized by a typical cross-section, defined as follows :

   (a) The cross section is assumed to have either a trapezoidal or a rectangular shape (cf figure H.2). In case we choose a trapezoidal shape, the bank slope is set to 1/2, which is the typical value in the Seine.

   (b) The bottom elevation is defined according to the uniform slope previously evaluated. We have in fact made a reconstitution for three different values of the slope, namely $1.2\,10^{-4} \pm 10\%$.

   (c) Finally, the cross section is such that, when the free surface elevation is at its reference level ("retenue normale", defined by the SNS) in the Suresnes-Chatou reach, namely 23.22 m, the river width equates the width observed on aerial surveys, and reported in table H.2.

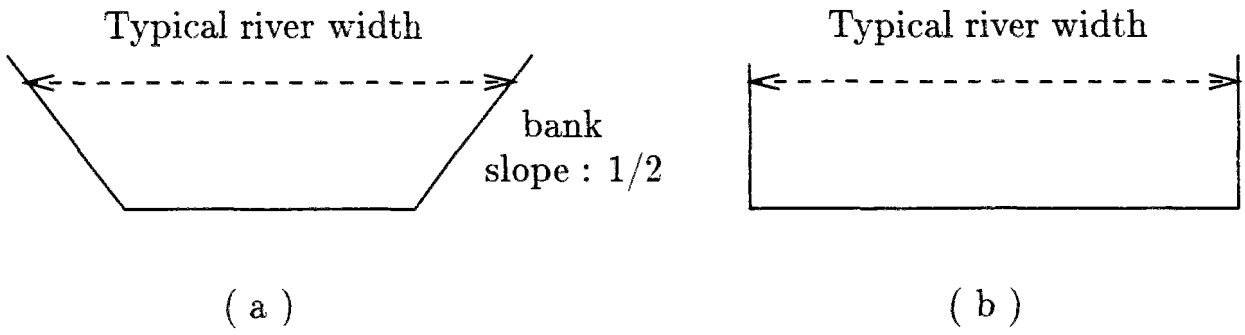


Figure H.2: Cross section shapes chosen for reconstitution

There are other areas where bathymetry is unknown. First, a straight stretch, 750 m long, in the middle of which we find the Clichy outlet. There, the bathymetry has merely been interpolated between the closest upstream and downstream available profiles. Secondly, there is a 500 m area in the right arm, between profiles P2721 and P2801 (cf figure H.3). This is a slightly curving strech, with a narrowing in its central part. A survey of profiles P2721 and P2801 (figure H.4) (the first one recorded in 1988, the second one in 1974) reveals that they have a nearly perfect trapezoidal shape : average bottom elevation is 19 m for the upstream (P2721) profile, 17.92 m for the downstream one. In order to reconstitute the bathymetry, we have introduced to hypothetical profiles, one at profile P2728 (the narrowest section, 74 m wide), the other at profile P2733 (84 m wide). These cross sections are assumed to be trapezoidal, with the same bank slopes as P2801. Their bottom elevation is computed by linear interpolation between P2721 and P2801.

H.1.3 Strickler calibration for reference flow rates

The above reconstitution of St-Denis left arm is of course rather crude, but it seems it's the best we can do so far. The remaining part of the Suresnes-Chatou reach is described as follows (Even & Poulin, 1993). Typical cross sections are selected by comparing the available profiles. The selection takes of course into account the major variations in the river bathymetry (depth or width changes). Each selected cross section is assumed to represent some part of the river course. The average length of these sub-reaches is about 500 m.

We shall see hereafter that arms geometry strongly influences the flow repartition. Consequently, we have been trying to calibrate the model with two different representations of the right arm off St-Denis island. The first one is the representation actually used in PROSE, where the arm is divided into 12 pieces (average length 590 m). The second one is more detailed, based on a closer examination of the profiles : there the arm is divided into 26 pieces (average length

Figure H.3: Area of missing bathymetric data in the Seine right arm

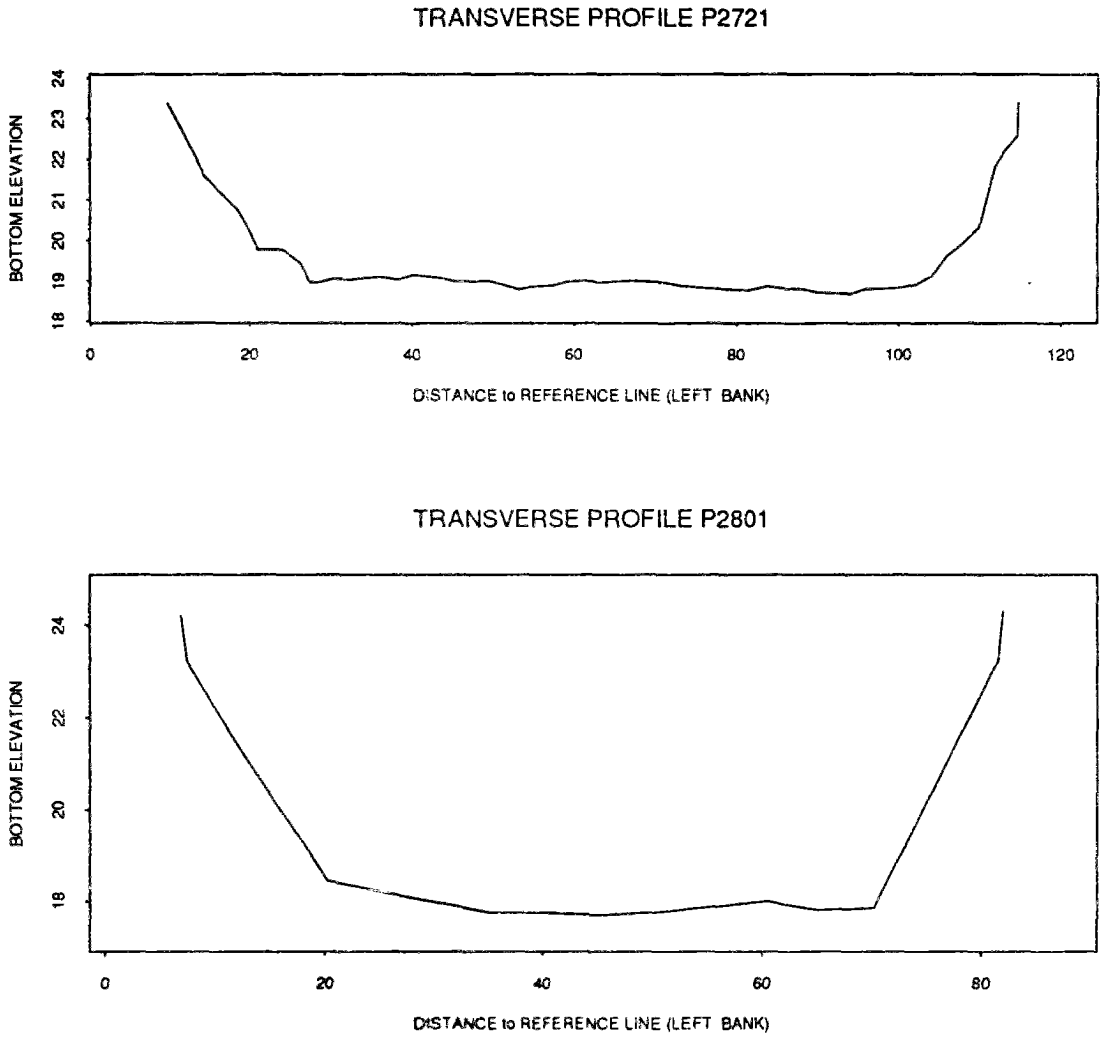


Figure H.4: Seine cross sections on both sides of the unknown area

270 m).

The calibration has been undertaken for these 12 possible combinations : 2 possible descriptions of the St-Denis right arm, 6 possible descriptions of the left arm (3 slope values $\times$ 2 sections shape).

The Strickler coefficient is dependent on the bed roughness, on the kind of materials we find there, and of the bed irregularities (Carlier, 1986). The Seine River features downstream Paris appear to be fairly uniform. Consequently, it seems reasonable to assume, in this first stage of our analysis, that the Strickler coefficient is uniform between Suresnes and Chatou.

Calibration results are given in tables H.3 to H.5. The tables indicate for different K_s (Strickler) values the estimated difference level ΔZ between the upstream and downstream endpoints of the Suresnes-Chatou reach, its absolute (ϵ) and relative error (δz) with respect to the expected difference (as documented in table H.1), as well as the percentage of discharge that flows through the right arm off St-Denis island (δQ).

It turns out that, once the shape of the left arm cross-sections has been chosen, the choice of the slope value has negligible influence over the calibration. This can be explained by the fact that these slope variations induce but minor changes in the arm wet sections (0.5 to 2 %). On the contrary, when assuming the cross sections are rectangular, the wet sections increase by 9 to 15 % with respect to what we obtain with trapezoidal sections with the same bed slope. Yet, even these bigger variations barely modify the Strickler adjustment. What they do change is the flow repartition between the arms around St-Denis island. Refining the right arm description has similarly little impact on the Strickler tuning, more on flow balance.

Table H.3: Calibration of Strickler coefficient for $Q = 100$ m^3/s

K_s	ΔZ (cm)	ϵ (cm)	δz (%)	δQ (%)
Left arm : trapezoidal shape Right arm : as PROSE				
32	5.6	+0.6	+12	
33	5.3	+0.3	+6	
34	**4.97**	**-0.03**	**-0.6**	**57**
35	4.7	-0.3	-6	
36	4.4	-0.6	-12	
Left arm : trapezoidal shape Right arm : refined				
34	5.11	+0.11	+2.2	
34.5	**4.96**	**-0.04**	**-0.8**	**55.3**
Left arm : rectangular shape Right arm : as PROSE				
33	5.1	+0.1	+2	
33.5	**4.94**	**-0.06**	**-1.2**	**54.1**
34	4.8	-0.2	-4	

Table H.4: Calibration of Strickler coefficient for $Q = 200$ and 300 m^3/s

$Q = 200$ m^3/s					$Q = 300$ m^3/s				
K_s	ΔZ (cm)	ϵ (cm)	δz (%)	δQ (%)	K_s	ΔZ (cm)	ϵ (cm)	δz (%)	δQ (%)
Left arm : trapezoidal shape & Right arm : as PROSE									
27	29.9	+3.9	+15		26	62.9	+6.9	+12.2	
28	28.0	+2.0	+6.9		27	59.1	+3.1	+5.5	
29	**26.22**	**+0.22**	**+0.8**	**57.2**	**28**	**55.7**	**-0.3**	**-0.5**	**57.1**
30	24.6	-1.4	-5.4		29	52.5	-3.5	-6.2	
31	23.2	-2.8	-10.8		30	49.5	-6.5	-11.6	
Left arm : trapezoidal shape & Right arm : refined									
29	27.0	+1	+3.8		28	57.2	+1.2	+2.1	
29.5	**26.16**	**+0.16**	**+0.6**	**55.2**	**28.5**	**55.5**	**-0.5**	**-0.9**	**55.1**
Left arm : rectangular shape & Right arm : as PROSE									
28	27.2	+1.2	+4.6		27	57.7	+1.7	+3	
28.5	**26.3**	**+0.3**	**+1.2**	**54.4**	**27.5**	**56.0**	**0**	**0**	**54.6**
29	25.5	-0.5	-1.9		28	54.3	-1.7	-3	

Table H.5: Calibration of Strickler coefficient for Q = 400 and 500 m^3/s

Q = 400 m^3/s					Q = 500 m^3/s				
K_s	ΔZ (cm)	ϵ (cm)	δz (%)	δQ (%)	K_s	ΔZ (cm)	ϵ (cm)	δz (%)	δQ (%)
Left arm : trapezoidal shape & Right arm : as PROSE									
30	95.3	+8.3	+9.5		31	127.6	+9.6	+8.1	
31	90.8	+3.8	+4.4		32	122.2	+4.2	+3.6	
32	**86.6**	**-0.4**	**-0.46**	**57.2**	**33**	**117.1**	**-0.9**	**-0.76**	**57.1**
33	82.7	-4.3	-4.9		34	112.2	-5.8	-4.9	
34	79	-8.0	-9.2		35	107.7	-10.3	-8.7	
Left arm : trapezoidal shape & Right arm : refined									
32	88.8	+1.8	+2.1		33	119.8	+1.9	+1.6	
32.5	**86.8**	**-0.2**	**-0.23**	**55.1**	**33.5**	**117.4**	**-0.6**	**-0.5**	**54.8**
Left arm : rectangular shape & Right arm : as PROSE									
31	88.8	+1.8	+2.1		32	120	+2.0	+1.7	
31.5	**86.7**	**-0.3**	**-0.34**	**54.5**	**32.5**	**117.3**	**-0.7**	**-0.59**	**54.6**
32	84.6	-2.4	-2.8		33	114.8	-3.2	-2.7	

Finally, the calibration results can be summarized as below :

Table H.6: Calibration results for the Suresnes-Chatou reach

Q (m^3/s)	100	200	300	400	500
K_s	34 ± 0.5	29 ± 0.5	28 ± 0.5	32 ± 0.5	33 ± 0.5

According to the assumed geometry of the arms, the right arm off St-Denis island appears to convey between 54 and 57 % of the total discharge, *with the hypothesis that the Strickler coefficient has the same value on every part of the Suresnes-Chatou reach.*

As mentionned earlier, due to the part played by dams, there is no unique relationship between surface slope and flow rate in the reach. Besides, our knowledge of the reach bathymetry is far from being perfect. Consequently, we do not consider that the values listed in table H.6 provide us with a reliable relation $K_s(Q)$ allowing to tune surely the Strickler as a function of the flow rate. We reckon simply that they indicate us the probable range of variation of rugosity

coefficients in this reach. This range (namely [28, 35]) is slightly lower than the usual values (40-45) recorded for rivers similar to the Seine River (Carlier, 1986).

H.1.4 Calibration for the dye-tracing experiments

During the dye-tracing experiment of 8^{th} September 1992, the flow stayed in the range 167 to 178 m^3/s (mean value : 170) according to the SNS gauging stations. This is confirmed by the measurements achieved in each arm of the St-Denis island during the morning : the estimated discharges are 68 and 100 m^3/s respectively in the left and right arms. During the day, the Bougival dam regulation was not modified. Such is not the case for Chatou and Suresnes dams, as can be seen on table H.7. Information available on free surface profile is sparse as usual.

Table H.7: Dam regulation during dye-tracing experiment 8/9/92

Suresnes lock			Chatou lock	
Hour	**Upstream elevation**	**Downstream elevation**	**Hour**	**Upstream elevation**
3h 00	26.58 m			
5h 00	Modification			
6h 30	Modification			
7h 00	26.49 m	23.75 m		
			8h 00	23.63 m
			9h 00	Modification
			11h 00	Modification
			12h 00	23.56 m
17h 00	26.43 m		17h 00	23.48 m

We have looked for the Strickler value which allows to reproduce the free surface profile observed at about 7 a. m., namely 23.75 m downstream Suresnes dam (this one being regulated so that the water level upstream the dam is 26.49 m.) and 23.63 m upstream Chatou dam. The Seine flow is set to 170 m^3/s in Paris, which reduces to 168.8 m^3/s at Suresnes, because of a water intake located immediately upstream the lock. As previously, the Strickler is first assumed to be uniform all over the reach. We selected the finer representation of the right arm and trapezoidal cross sections in the left one. Anyway, this choice should have few influence on the global Strickler calibration, as illustrated in the previous section. We observ the following :

K_s	ΔZ (cm)	ϵ (cm)	δz (%)	δQ (%)
34	13.02	+1.02	+8.5	54.90
35	12.30	+0.30	+2.5	54.86
36	11.66	-0.34	-2.8	54.82
35.5	11.98	-0.02	-0.2	54.84

The adequate Strickler appears thus to be in the range 35-36. However, the forecasted flow repartition does not agree with the observed one, namely $\delta Q = 59.5$ %. Considering that the bathymetry influences flow repartition and that our knowledge of the left arm geometry is poor, such discrepancy is not surprising.

How could we achieve a better agreement ? This could be done either by modifying left arm geometry, either by modifying the respective Strickler values in each arm. We haven't any clues available to guide us in changing the left arm representation, except perhaps that it should be done so that wet sections are reduced. Indeed, the larger the wet section in the left arm, the more even becomes flow repartition (compare forecasts with rectangular vs. trapezoidal shape in tables H.3 to H.5). Consequently, we have decided to act on the Strickler values. There again, we have two options, changing the Strickler value in both arms with respect to the global reach value, or changing it only in the left arm. After several trials, the best agreement is observed with the following combinations :

K_s (global)	K_s Right arm	K_s Left arm	ΔZ (cm)	ϵ (cm)	δz (%)	δQ (%)
35.5	35.5	29.5	12.74	+0.74	+6.2	59.54
35.5	35.5	30	12.66	+0.66	+5.5	59.12
36	36	30	12.39	+0.39	+3.2	59.45
35.5	38.5	32.5	11.93	-0.07	-0.6	59.03
35.5	39	32	11.92	-0.08	-0.6	59.72

It appears that Strickler values must be strongly modified in order to achieve the adequate flow balance. Roughness must be considerably increased in the left arm : there, the Strickler needs to be 20 % lower than in the right one. Considering the nature of the Seine bed, it is hard to find some logical, sound (from a physical point of view), justification for such difference in the rugosities. The modification of the Strickler values appears thus to be a completely artificial way of achieving better forecasts. In that case, we thought it preferable to let all the changes affect the poorly known left arm, and to keep the same Strickler value, namely 36, for other parts of the reach.

An other dye-tracing experiment, on the 5^{th} of October 93, has involved the collection of hydraulic data on the Suresnes-Chatou reach. The flow rate during this experiment stayed fairly

constant too, around 320 m^3/s ($\pm$ 5 %, according to the gauging stations precision). A gauging operation was achieved in the left arm off St-Denis island, with some difficulty because of the strength of the flow which prevented accurate measurements to be realized in the vicinity of the river bed. Consequently the discharge estimation is plagued with some uncertainty. The suggested value is 158 m^3/s (49.4 % of a total flow of 320 m^3/s), but the possible range is 144 to 170 m^3/s (45 % to 53 %). Thus, the flow repartition seems to be different than observed on September 92 (40 %), which is somewhat surprising. The global Strickler value which allows to reproduce the observed level difference between Suresnes and Chatou (namely 60 cm) is approximately 30-31. When the Strickler value is tuned so that the left arm conveys about 160 m^3/s, the calculated transit time between St-Denis and Epinay bridges (in the left arm) is approximately 10 % larger than the transit time observed for the rhodamine cloud. This hints again at the fact that our actual reconstitution of the left arm probably overestimates the size of its cross sections.

H.2 Details about the representation of the studied reach

We recall hereafter how we have been representing the studied domain in a particular area, namely the upstream tip of the St-Denis island. Figure H.5 details the respective locations : of the true banks (dashed lines), of available river profiles (which extent is delimited by the small dotted lines), of the solid and open boundaries according to the chosen computational grid. Profiles P2519 and P2520 are not complete. However, it's most probable that the area located immediately usptream the island tip is a very shallow one. In the absence of any information relative to this zone, we assumed there was a solid boundary there, at the extremities of profiles P2519 and P2520. The open boundary is defined so that it has the same width as the "true" left arm.

As can be observed on figures H.6 to H.8, there are differences between soundings made in 1988 and in 1989, obvious on profiles P2518 to P2520. In order to restore the navigation channel, dredging has been applied. We have no idea of the speed of the silting phenomenon that takes place upstream the island. Consequently, we do not know whether the state of the Seine on the day of experiment (October 1992) was closest to 1988 or 1989 profiles. Fearing that this may have some influence over the flow repartition we have been studying this repartition with the two kind of bathymetric data : 1988 and 1989 soundings.

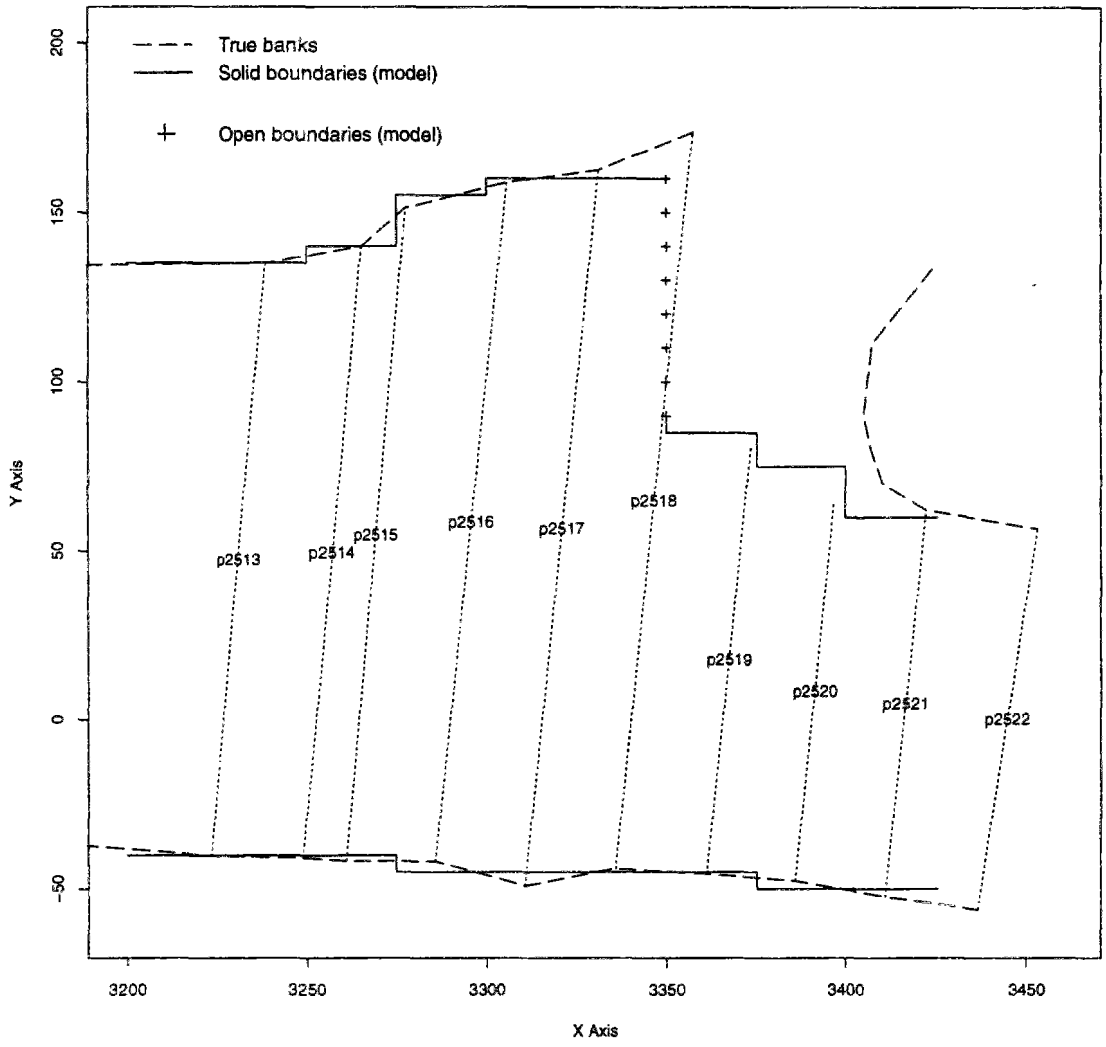


Figure H.5: Details of the Seine discretization at the upstream end of St-Denis island

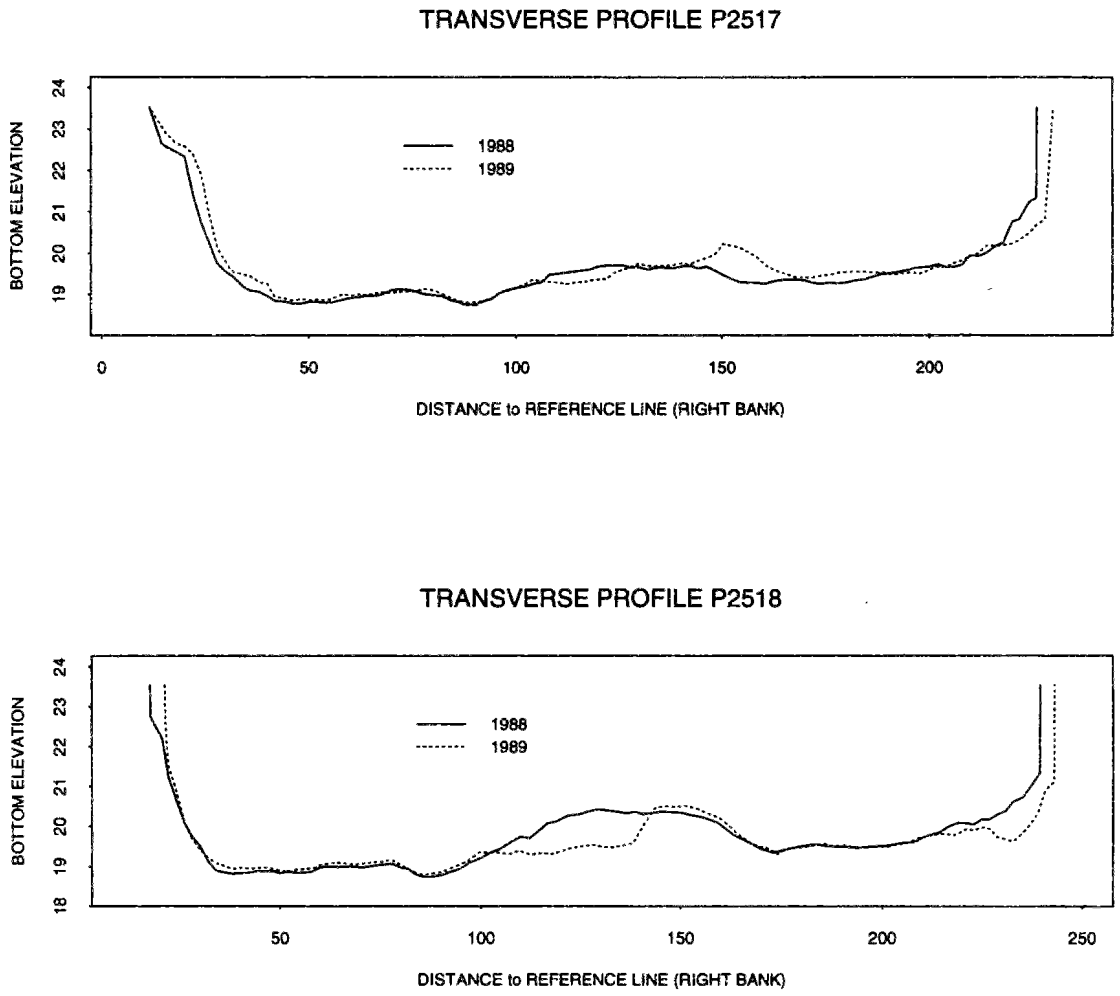


Figure H.6: Changes in Seine bathymetry upstream St-Denis island : profiles P2517 & P2518

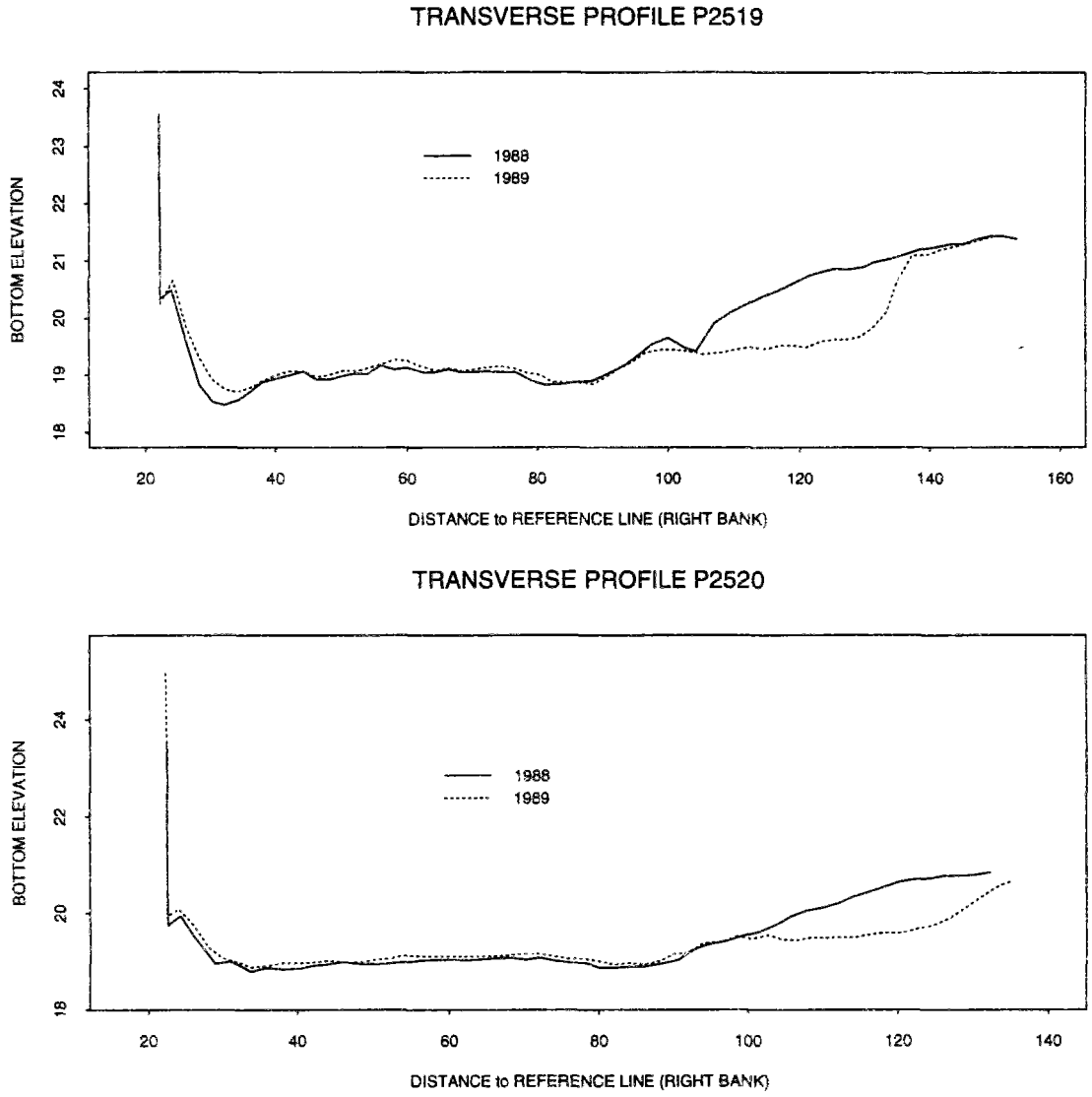


Figure H.7: Changes in Seine bathymetry upstream St-Denis island : profiles P2519 & P2520

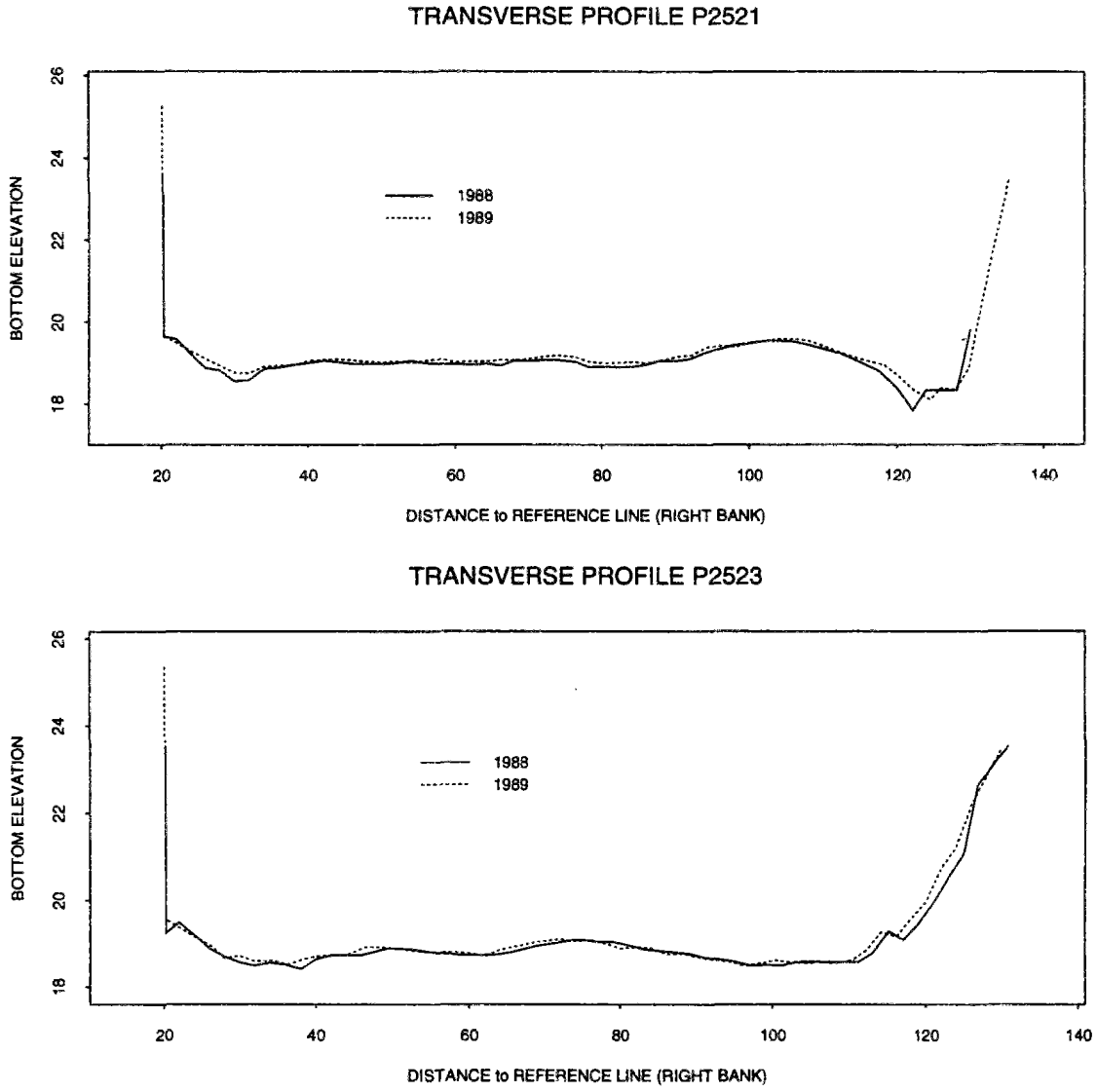


Figure H.8: Changes in Seine bathymetry upstream St-Denis island : profiles P2521 & P2523

H.3 Supplementary visualizations of forecast velocity fields

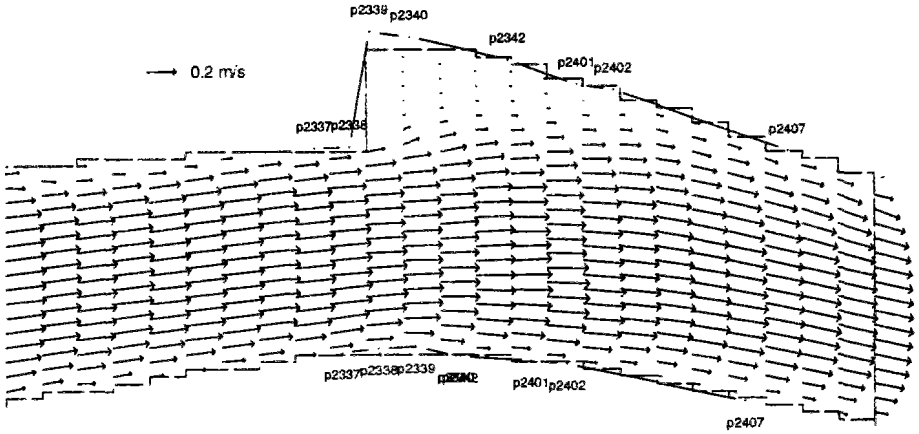


Figure H.9: Flow pattern around local widening of the river (PK 24, 500 m downstream Clichy)

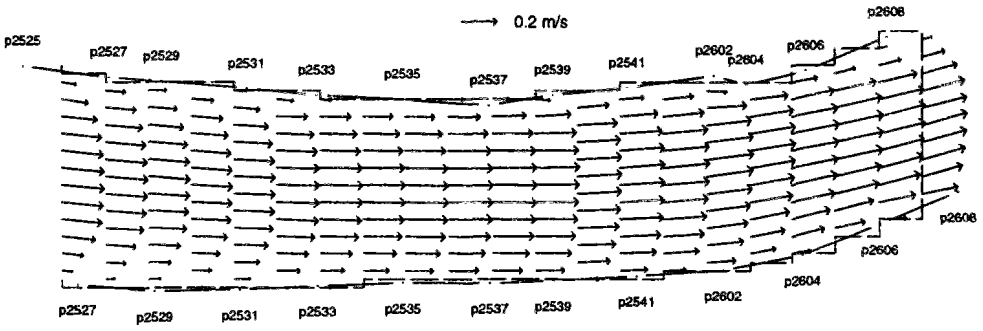


Figure H.10: Flow pattern in the vicinity of St-Ouen road bridge ($\simeq$ profile P2602)

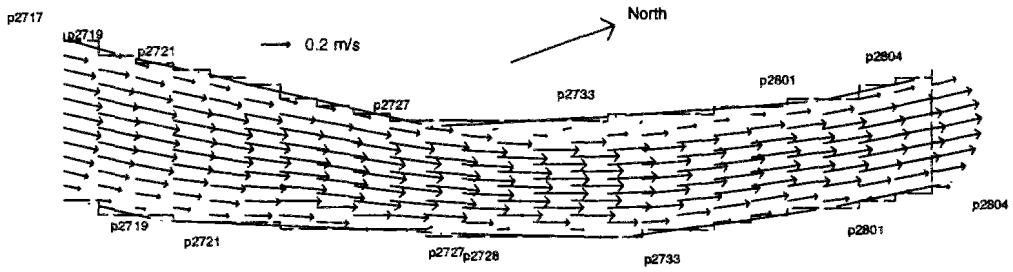


Figure H.11: Flow pattern in the reconstituted narrowing of the right arm (profiles P2721 to P2801)