Towards an intelligent control of wastewater treatment process : Development of a hybrid model combining a knowledge-based biofiltration model with a data-driven model to improve simulation performance and optimise process control.

Marcello Serrao

École des Ponts ParisTech

THÈSE DE DOCTORAT
de l'École des Ponts ParisTech

Vers un contrôle intelligent des procédés de traitement des eaux usées urbaines : développement d'un modèle hybride afin d'améliorer la performance de simulation et optimiser le contrôle des processus.

École doctorale N° 531, Sciences, Ingénierie et Environnement SIE

Spécialité : Sciences et Techniques de l'Environnement

Thèse préparée au sein du Laboratoire Eau, Environnement et Systèmes Urbains (Leesu)

Thèse soutenue le 4 juillet 2023, par **Marcello SERRAO**

Composition du jury :

Claire, ALBASI Directrice de recherche, CNRS	*Présidente*
Sylvie, GILLOT Directrice de recherche, INRAE	*Rapporteuse*
Ignasi, RODRIGUEZ RODA LAYRET Professeur, Universitat de Girona	*Rapporteur*
Elena, TORFS Docteure, Université Laval	*Examinatrice*
Bruno, TASSIN Directeur de recherche, ENPC	*Directeur de thèse*
Peter, VANROLLEGHEM Professeur, Université Laval	*Co-directeur de thèse*
Ilan, JURAN Professeur, New York University	*Co-encadrant de thèse*
Vincent, ROCHER Docteur, SIAAP	*Co-encadrant de thèse*

Vers un contrôle intelligent des procédés de traitement des eaux usées urbaines : développement d'un modèle hybride ayant pour but d'améliorer la performance de simulation et l'optimisation du contrôle des processus.

École des Ponts / ParisTech

École doctorale N° 531, Sciences, Ingénierie et Environnement SIE

Spécialité : Sciences et Techniques de l'Environnement

Thèse soutenue le 4 juillet 2023 par Marcello SERRAO

Thèse préparée au sein du
Laboratoire Eau, Environnement et Systèmes Urbains LEESU (ENPC)
en codirection avec model*EAU* (ULaval)
dans le cadre du programme de recherche Mocopée phase II
en collaboration avec le SIAAP et l'association W-SMART.

Bruno Tassin

Peter Vanrolleghem

Vincent Jauzein - Sam Azimi - Vincent Rocher

Ilan Juran

Towards intelligent Process Control of Municipal Wastewater Treatment: the Development of a Hybrid Model that aims to improve Simulation Performance and Process Optimization.

École des Ponts / ParisTech

École doctorale N° 531, Sciences, Ingénierie et Environnement SIE

Spécialité : Sciences et Techniques de l'Environnement

PhD dissertation defended on 4th of July, 2023 by Marcello SERRAO

This thesis was prepared at the

Laboratoire Eau, Environnement et Systèmes Urbains LEESU (ENPC)

in co-direction with model*EAU* (ULaval)

within the framework of the Mocopée phase II research program

in collaboration with SIAAP and W-SMART.

Bruno Tassin

Peter Vanrolleghem

Vincent Jauzein - Eloïse DeTredern

Sam Azimi - Vincent Rocher

Ilan Juran

"If I were given one hour to save the planet,
I would spend 55 minutes defining the problem,
and five minutes resolving it!"
(Albert Einstein)

This thesis is dedicated to Eng**. Alberto Nunes Serrao** (1909 – 1985), who, first generation coming from Europe, overcame immense obstacles when raising his family in the New World. He became a mathematician and a civil engineer. He was highly recognized for his pedagogical talents as a professor at the *Escola Nacional de Engenharia* in Rio de Janeiro (Brazil).

Above all, he was my great-uncle.

For more information contact: Marcello.Serrao@enpc.fr / Marcello.Serrao@outlook.com

Abstract

Wastewater treatment models are recognized as important decision support tools for process design, scenario analysis, impact assessment, optimization of treatment operation, and for educational purposes. Several biofiltration models have been developed so far based on modified Activated Sludge Models. They allow simulating the removal of organic matter and nutrients, as well as greenhouse gas emissions. Generally, these models have been calibrated with short-term experimental data in pilot set-ups and have reached reasonable performance. The greater complexity of mechanistic models of biofiltration makes their parameterization, calibration and validation much more laborious, requiring more dedicated measurement data. This can become a bottleneck to model application in an operational context.

Alternatively, hybridization of mechanistic models with data-driven models allows taking advantage of the strong points of each model type: calculations of water quality variables are based on the fundamental understanding of the physical-chemical-biological processes, which are supported by a fast computing data-driven model to correct for missed dynamics. The data-driven model learns unknown relationships from the data and helps improving simulation quality, especially for interpolation purposes. In recent literature, hybrid models describing activated sludge processes have improved model performance and supported process monitoring and control. As far as known to the authors, in this study, the approach is applied for the first time to biofilm systems. Hybrid models have a higher benefit/cost ratio for solving complex problems, which is a key factor for process systems engineering.

In this study, a hybrid biofiltration model was developed that improves model prediction power compared to the pure mechanistic one. A previously developed biofiltration model was recalibrated with recent data from the biofiltration lane of the 6 million PE, 1.5 million m^3/day wastewater treatment plant that uses submerged upflow biofilters for denitrification with controlled methanol dosing. The mechanistic model is a 1-dimensional dynamic biofilm reactor model describing a 'fixed-culture' in an upward flow biological filter. The model simulates in total 22 biological and physical-chemical conversions of carbon, nitrogen and phosphorus by describing the evolution of all components with mass balances. The recalibrated model effluent nitrate concentrations are however overestimated by a relative mean error of up to 25% (calibration for 1 month).

Data-driven models were selected among linear regression models, regression decision tree models; support vector machines and artificial neural networks with up to three hidden layers. The hybrid model architecture adopted is a cooperative parallel hybrid structure with a 3-layer feedforward neural network with 12, 25 and 1 neurons for the input, middle and output layer and a ReLU activation function. It was trained to calculate the error between the simulations of the mechanistic model and the observations to allow correcting the simulated effluent quality variables. The results obtained show that the output of the hybrid model is much closer to the observations for the effluent nitrate concentrations with a mean error of 2% (training) and 8% (testing). Low Janus coefficients in the range of 1 to 2 confirm the validity of the hybrid model.

The hybrid model was subsequently applied in a process control system for methanol dosing in a predenitrification biofilter to reach a certain set point of the effluent nitrate concentration. The

proposed hybrid model predictive controller first searches for an optimal dosing rate by running multiple scenarios through the mechanistic biofilter model and the data-driven model described earlier for a two-hour prediction horizon. The HMPC then selects the scenario with the closest agreement with the nitrate set point. The proposed action corresponding to the first 15 minutes time step is then applied to a second biofilter model that represents the real physical system. This allows studying the effect of the proposed optimal dosing rate. Feedback to the HMPC optimizer is provided for so that the optimizer can correct of any divergence between the model included in the HMPC and the model playing the role of reality. The results show that the HMPC control strategy leads to an improvement in effluent quality compared to the current feedforward controller implemented at the plant, albeit with an increased operational cost of methanol dosing. Further research is needed to study the effect of a multi-objective optimization strategy that searches for optimal dosing rates, while at the same time considering the economic costs and associated greenhouse gas emissions.

Résumé

En contexte industriel, les modèles mécanistes de traitement des eaux usées décrivant les procédés de boues activées sont reconnus comme des outils importants d'aide à la décision. Ils intègrent un grand nombre de paramètres et nécessitent une phase complexe de calage préalable. Dans le cas du traitement par biofiltration, un procédé compact et largement appliqué en région urbanisée parisienne, la nécessité de modéliser les processus bio-physico-chimiques au sein du biofilm complexifie ces modèles mécanistes. Ces modèles pour la biofiltration nécessitent une puissance de calcul relativement élevée et ne sont pas encore bien établis.

Des modèles de *Machine Learning (ML)* basés sur les données décrivent quant à eux le système uniquement sur les informations extraites des données. Ils présentent une force d'interpolation et sont beaucoup plus rapides en calcul, ce qui les rend très intéressants pour les applications en temps réel.

L'hybridation combine un modèle mécaniste qui intègre des connaissances pertinentes sur les processus, avec un modèle basé sur les données qui augmente la précision des estimations en incluant des informations sur des sous-processus mal décrits.

Dans cette étude, un modèle mécaniste existant est recalibré avec des données de mesure haute fréquence provenant de l'étage de dénitrification amont de la filière de biofiltration d'une usine à grande échelle. Un modèle hybride, qui intègre un modèle de *Machine Learning* d'estimation des résidus, est décrit. Il permet une réduction significative des erreurs d'estimation.

1 Introduction

Le programme de recherche Mocopée phase I (2014 – 2017) a permis de développer des modèles mathématiques de prédiction des variables de qualité d'eau pour les procédés de traitement appliqués aux stations d'épuration de la région parisienne (SIAAP, 2018). Cet effort a abouti à la conception de modèles mécanistes de décantation primaire physico-chimique (Bernier et al., 2016), de traitement secondaire par biofiltration (Zhu, 2020) et par bioréacteur à membrane (Nadeem, 2021). L'étape suivante a consisté à développer un modèle filière qui intègre les trois phases de biofiltration : la dénitrification amont (DAM), la nitrification (NIT) et la dénitrification aval (DAV).

Au contraire des modèles mécanistes ASM (Activated Sludge Models, Henze et al., 2000) qui décrivent les processus de croissance de la biomasse en suspension (présents dans les procédés de boues activées) et qui ont une base solide au sein de la communauté scientifique, les modèles mécanistes de biofiltration, qui est un procédé à base de culture fixée de microorganismes, sont encore en plein développement. Les équations biologiques prises en compte par l'ASM ne peuvent pas être directement appliquées dans un modèle de biofiltration en raison de l'importance des processus de transport par diffusion, des cinétiques biologiques qui se déroulent dans le biofilm et du rétrolavage. De plus, il n'existe pas encore de protocoles largement acceptés pour développer des modèles de biofilm (Rittman et al., 2018). Les travaux de thèse de Jialu Zhu (2020) ont permis de développer un premier modèle filière, qui intègre les trois phases de biofiltration, à partir de données collectées au mois de mai 2018. Les

résultats montrent que le modèle est généralement capable de prédire les valeurs moyennes quotidiennes des variables en sortie.

Cependant, la complexité des modèles mécanistes rend leur paramétrage et leur calage laborieux, consommateurs en temps et en données ; cela peut devenir une contrainte limitant leur application (Vanrolleghem et al., 2005). Ils nécessitent une puissance de calcul élevée, ce qui les rend moins adaptés pour les applications de contrôle en temps réel avec un court horizon temporel d'action (Schneider et al., 2022).

A l'opposé, les modèles de ML basés sur les données utilisent des méthodes d'apprentissage automatique pour trouver des relations inconnues entre les données, et sont très rapides en temps de calcul. Ces modèles décrivent le système uniquement sur la base d'informations extraites des données et ont de fortes capacités d'interpolation (Newhart et al., 2019). A cause des caractéristiques non-stationnaires et de la variabilité temporelle potentiellement importante des liens existants entre les données, un réentraînement fréquent de ce type de modèle reste nécessaire (Torfs et al., 2022).

La modélisation hybride, combinant l'utilisation de modèles mécanistes, basés sur la connaissance du système, avec des modèles de ML, basés sur les données, est capable de cumuler les avantages de chaque approche (Von Stosch et al., 2014). Le principal avantage est un rapport bénéfice/coût plus élevé pour résoudre des problèmes complexes, ce qui est un facteur clé pour l'ingénierie des systèmes de procédés (Schneider et al., 2021). Ce travail s'intègre dans une démarche de développement de jumeaux numériques dans le domaine de l'assainissement.

L'objectif du travail présenté est donc quadruple ; premièrement, mettre à jour le modèle de biofiltration développé précédemment (Zhu, 2020) en calant ses paramètres grâce à des données acquises sur site industriel en période 2019 – 2020 et préalablement qualifiées et identifiées grâce à une analyse de sensibilité; deuxièmement, développer un modèle de *ML* permettant l'identification des erreurs du modèle mécaniste; et troisièmement, intégrer ces deux composantes dans un modèle hybride afin d'augmenter la précision des estimations ; et quatrièmement, l'application d'un modèle hybride comme étude de cas de contrôle de processus dans un cadre opérationnel avec des données à haute résolution.

Le modèle hybride développé (Figure 1) est une structure hybride parallèle coopérative (Schneider et al., 2022) où le modèle de ML prédit la différence entre les valeurs des simulations du modèle mécaniste et les observations (l'erreur résiduelle, Δ). Ceci nécessite que le modèle mécaniste fasse d'abord des estimations des valeurs de variables en sortie du procédé (ligne bleue dans la figure), qui sont ensuite analysées par le modèle de *ML* (boîte bleue), qui reçoit également les informations sur la qualité d'eau à l'entrée du procédé (ligne noire). Enfin, les résultats du modèle mécaniste et du modèle de *ML* sont combinés pour en déduire une valeur plus précise. Dans le cas de simulations prédictives, en l'absence de mesure en sortie du procédé, le modèle hybride nous permet de prédire les valeurs de l'eau traitée.

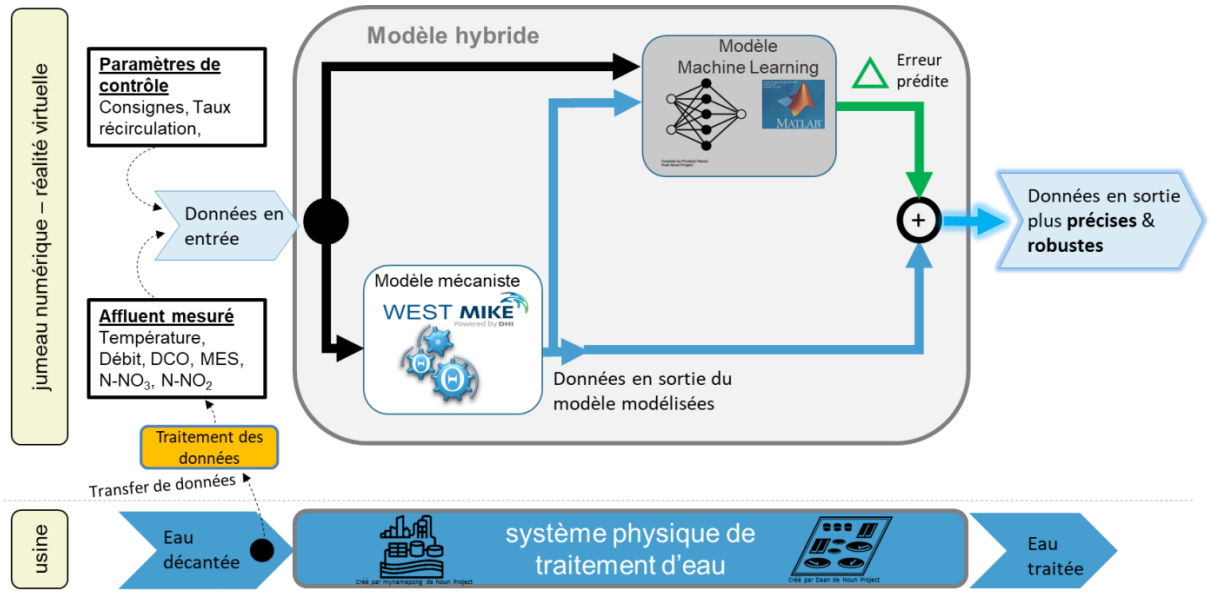


Figure 1 : Structure de modèle hybride (boîte grise) combinant un modèle de Machine Learning basé sur les données (boîte noire) et un modèle mécaniste basé sur la connaissance du système bio-physico-chimique (boîte blanche).

1. Matériels et méthodes

1.1 Description de la station d'épuration Seine aval et de la biofiltration

Le projet de recherche a été mené au sein de la station d'épuration de Seine aval (SAV) à Achères (Yvelines) gérée par le SIAAP. Cette usine épure un volume de 1 250 000 m^3/j en temps sec (14,5 m^3/s ; 6 000 000 eh).

La filière de traitement des eaux est constituée d'un traitement primaire physico-chimique (dégrillage, dessablage-déshuilage et décantation primaire favorisée par l'ajout de réactifs coagulant-floculant), suivi d'un traitement biologique secondaire soit par biofiltration, soit par bioréacteurs à membrane. Pendant la période étudiée, le débit moyen par temps sec vers la filière de biofiltration était de 1 090 000 m^3/j.

Sur la filière de biofiltration, l'élimination de l'azote est un processus biologique en deux étapes : l'oxydation de l'ammonium en nitrites puis nitrates par les bactéries autotrophes (nitrification - NIT), puis la conversion des nitrates en diazote gazeux par les bactéries hétérotrophes oxydant la matière organique dans des conditions anoxiques (dénitrification - D). Dans le cas de la filière de biofiltration du site industriel de Seine aval présentant une configuration dénitrification amont (DAM) – nitrification (NIT) – dénitrification aval (DAV), ceci nécessite une recirculation de l'eau nitrifiée issue de l'étage NIT vers la DAM, qui a pour atout de consommer la matière organique naturellement présente en sortie de la décantation primaire. Les concentrations en nitrates et en demande chimique en oxygène (DCO) en sortie de DAM sont des indicateurs importants de l'activité biologique dans la DAM, procédé sur lequel porte cette étude.

1.2 Jeux de données utilisées

Les jeux de données utilisés lors de cette étude couvrent la période de décembre 2019 à octobre 2020. Ils sont constitués de mesures provenant des capteurs, sondes et analyseurs en ligne de débit et de variables de qualité de l'eau pour la matière organique, l'azote et les orthophosphates toutes les 15 minutes. S'y ajoutent les données issues d'analyses réalisées

au laboratoire accrédité du SIAAP sur des échantillons quotidiens récoltés aux points clés des filières de traitement.

La période de calibration s'étale de décembre 2019 au 15 mars 2020. La période de la validation va de juin à septembre 2020, période estivale caractérisée par de très faibles débits pendant le mois d'août, des conditions très différentes de la période de calibration.

Les données collectées en format brut ont d'abord été prétraitées en utilisant la méthode décrite par Alferes et Vanrolleghem (2016). Cette méthode inclut trois étapes de traitement : un lissage exponentiel des séries temporelles ; un lissage utilisant la méthode des noyaux ; et une détection et une suppression des données aberrantes. Ces données de capteurs ont ensuite été corrigées en utilisant les données issues des analyses en laboratoire qui sont considérées comme plus fiables.

La méthode de caractérisation des variables d'entrée du modèle de biofiltration est basée sur le fractionnement des entrées du modèle ASM1 (Henze et al., 2000). Le modèle de fractionnement de la matière organique de l'affluent a été construit à partir des mesures de DCO totale et DCO soluble. Les valeurs des paramètres utilisés pour quantifier ce fractionnement sont basées sur les valeurs par default d'ASM1 et du modèle calibré par Zhu (2020). Le fractionnement de l'azote a été basé sur des variables mesurées directement.

1.3 Description du modèle mécaniste

Le modèle mécaniste utilisé est un modèle dynamique en une dimension d'un filtre biologique immergé à flux ascendant à culture fixée, sur lequel les micro-organismes se développent au sein d'un biofilm, précédemment développé par Zhu (2020) dans le cadre de sa thèse (2020, UTC). Il se compose de sous-modèles de représentation hydraulique du média filtrant, de transfert et de transport de substrats solubles et particulaires entre eau et biofilm et au sein du biofilm, de la cinétique de croissance et de mortalité des bactéries et du rétrolavage pour éviter le colmatage du filtre. Le modèle simule la conversion biologique et physicochimique du carbone, de l'azote et du phosphore en décrivant l'évolution des substrats par des bilans de masse basés sur le modèle ASM1. Les principales modifications se concentrent sur le transfert de masse dans les biofilms, la cinétique des micro-organismes et l'influence des cycles réguliers de lavage.

Le modèle simule la filière de biofiltration en trois étapes (DAM, NIT, DAV) et est capable de calculer la qualité des eaux dans les différents réacteurs modélisés. Dans la suite de ce travail, nous nous focaliserons sur le modèle de qualité des eaux en sortie de DAM.

1.4 Méthodes de calage et de validation du modèle mécaniste

Des scores statistiques sont calculés pour évaluer la performance et la précision du modèle de manière qualitative et quantitative. Chaque simulation dynamique a été précédée d'une simulation de régime stationnaire de 30 jours afin d'initialiser le modèle.

Dans le cas de la calibration du modèle, les erreurs statistiques sont déduites de la comparaison entre les variables de l'effluent modélisées et les données mesurées. La pratique standard dans la modélisation des boues activées examine (Rieger et al., 2012) : l'erreur moyenne (EM, équation 1) qui permet de mesurer le biais du modèle (une valeur optimale proche de 0 indique un biais faible) ; l'erreur moyenne absolue (EMA, équation 2) et la racine de l'erreur quadratique moyenne (REQM, équation 3) divisées par la moyenne des observations ($\hat{y}$) qui permettent d'estimer la précision du modèle. Plus les valeurs EMA/$\hat{y}$ et

REQM/ŷ sont proches de 0, meilleure est la précision du modèle. Enfin, le coefficient de Janus (équation 4), avec une valeur optimale de 1 et une valeur acceptable de 2, permet de comparer les résultats de la validation avec ceux de la calibration (Hauduc et al., 2015).

$$EM = \frac{1}{n}\sum_{i=1}^{n}(O_i - E_i) \qquad Equation\ 1$$

$$EMA = \frac{1}{n}\sum_{i=1}^{n}|O_i - E_i| \qquad Equation\ 2$$

$$REQM = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - E_i)^2} \qquad Equation\ 3$$

$$Coefficient\ de\ Janus = \frac{RMSE_{validation}}{RMSE_{calibration}} \quad Equation\ 4$$

sachant O Observation, E Estimation.

La calibration du modèle dynamique de biofiltration suit la méthode décrite par Mannina et al. (2011), qui approfondit les lignes directrices de modélisation pour les modèles de boues activées de l'IWA (Rieger et al., 2012). La méthode consiste en : 1) la sélection des variables de sortie ; 2) l'identification des paramètres qui ont un effet important sur chacune des variables sélectionnées en effectuant une analyse de sensibilité ; 3) l'évaluation des scores statistiques qui indiquent la performance du modèle selon le jeu de paramètres testés.

L'analyse de sensibilité a permis de quantifier l'influence des paramètres sur des variables de qualité de l'eau en aval de la dénitrification, considérées comme indépendantes, dans l'ordre suivant : 1) matières en suspension (MES) ; 2) DCO soluble et $N\text{-}NO_3$; 3) $N\text{-}NH_4$; 4) $N\text{-}NO_2$; 5) $P\text{-}PO_4$.

Compte tenu du grand nombre de paramètres (142) et de variables (20) impliqués dans ce modèle, ainsi que de l'incertitude liée à la plupart d'entre eux, une analyse de sensibilité globale a été effectuée en appliquant des coefficients de régression standardisés (Cosenza et al., 2013), avec une présélection des paramètres par une approche experte et issue de la littérature. Au total, 89 paramètres ont été retenus.

L'analyse de sensibilité a consisté en 10 sessions parallèles de 90 simulations en régime permanent pour une période de 30,5 jours en utilisant la méthode probabiliste de Monte-Carlo avec l'échantillonnage hypercube latin (*Latin Hypercube Sampling*), qui permet un échantillonnage aléatoire stratifié (Saltelli et al., 2008). Au-delà de 30,5 jours, la variation des valeurs des variables est supposée être entièrement causée par les valeurs de paramètres choisies.

De façon plus détaillée, une analyse par régression linéaire multiple a permis de définir les relations entre paramètres et variables. Les données ont d'abord été normalisées grâce au calcul de z-scores (pour chaque point, soustraction de la moyenne puis division par l'écart type de chaque paramètre), étape nécessaire pour apporter uniformité et exactitude. Une gamme

d'erreur de 10 à 15 % selon la variable est considérée comme une valeur acceptable (Rieger et al., 2012).

1.5 Description du modèle hybride

Un modèle hybride implique une interaction entre un modèle mécaniste, qui sert de base fondamentale pour les estimations, et un modèle de ML qui applique des corrections sur ce dernier. Le modèle mécaniste assure la préservation de la connaissance du procédé épuratoire et offre de bonnes capacités d'extrapolation. Le modèle de ML offre des capacités d'estimation améliorées grâce à l'apprentissage implicite de relations inconnues entre les variables d'entrée et de sortie du modèle mécaniste. Ces capacités peuvent compenser les dynamiques manquées par le modèle mécaniste (Schneider et al., 2022).

1.6 Méthodes d'entrainement et de test du modèle hybride

Dans une structure de modèle hybride, plusieurs modèles de ML peuvent être appliqués, tels que des réseaux de neurones ou des structures d'arbres de décision. Pour l'entrainement de ces modèles de ML, les sorties modélisées par le modèle mécaniste sont comparées aux variables d'effluent mesurées de la DAM afin de calculer les résidus du modèle (les écarts entre valeurs observées et valeurs prédites). L'entrainement d'un modèle de ML suit un schéma général d'importation des données, d'organisation et de prétraitement des données, d'exploration des caractéristiques des données, de création du modèle et d'évaluation des performances. Les mêmes indicateurs statistiques que ceux appliqués au modèle mécaniste ont été calculés pour évaluer la performance du modèle hybride.

Plusieurs configurations de modèles ont été créées sur un ensemble de données couvrant la période allant du 1er décembre 2019 au 28 février 2020 (8 000 points de données par variable). Les variables normalisées qui ont été fournies au modèle incluent le type de jour (semaine ou weekend), le type de temps (sec ou pluvieux), la température, les débits, les concentrations observées de $N\text{-}NO_3$, MES, DCO totale et DCO soluble dans l'affluent, ainsi que des paramètres opérationnels comme la consigne de $N\text{-}NO_3$ et le dosage de méthanol. Les valeurs estimées par le modèle mécaniste à la sortie de la DAM sont ajoutées. En phase d'entrainement, l'erreur résiduelle du modèle mécaniste est également transmise afin de permettre un apprentissage supervisé. Trois modèles de réseaux neurones ont été entrainés pour chaque variable en sortie univariée afin de calculer les résidus des concentrations en NOx ($N\text{-}NO_3$ + $N\text{-}NO_2$), en DCO totale et en MES. En phase de test, le modèle sélectionné a été testé avec un nouveau jeu de données (1 300 points) couvrant la période du 1er mars 2020 au 15 mars 2020. La structure algorithmique optimale a été sélectionnée en comparant les performances de chaque modèle.

La modélisation mécaniste a été faite à l'aide du logiciel WEST (DHI) ; le traitement des données brutes, le développement et l'entrainement des algorithmes de ML ont été réalisés dans l'environnement MATLAB R2022a (MathWorks) à l'aide des boîtes à outils *Machine Learning* et *Deep Network Designer*.

1.7 Développement du contrôle prédictif du modèle hybride (HMPC)

Le concept du contrôle prédictif du modèle hybride (HMPC) développé dans cette étude pour le dosage du méthanol dans un biofiltre de dénitrification est illustré à la Figure 2. Le HMPC fonctionne en recherchant d'abord un taux de dosage optimal du méthanol en exécutant plusieurs scénarios à travers le modèle de biofiltre hybride, qui contient le modèle mécaniste calibré (le modèle nommé 'MPC', boîte bleue sur la figure) simulant le comportement d'un filtre

comme décrit précédemment, et le *modèle de ML* (ANN Model) développé (boîte jaune). Il compare ensuite l'effluent de nitrate simulé avec la concentration de NO_3-N de l'effluent souhaitée (la consigne).

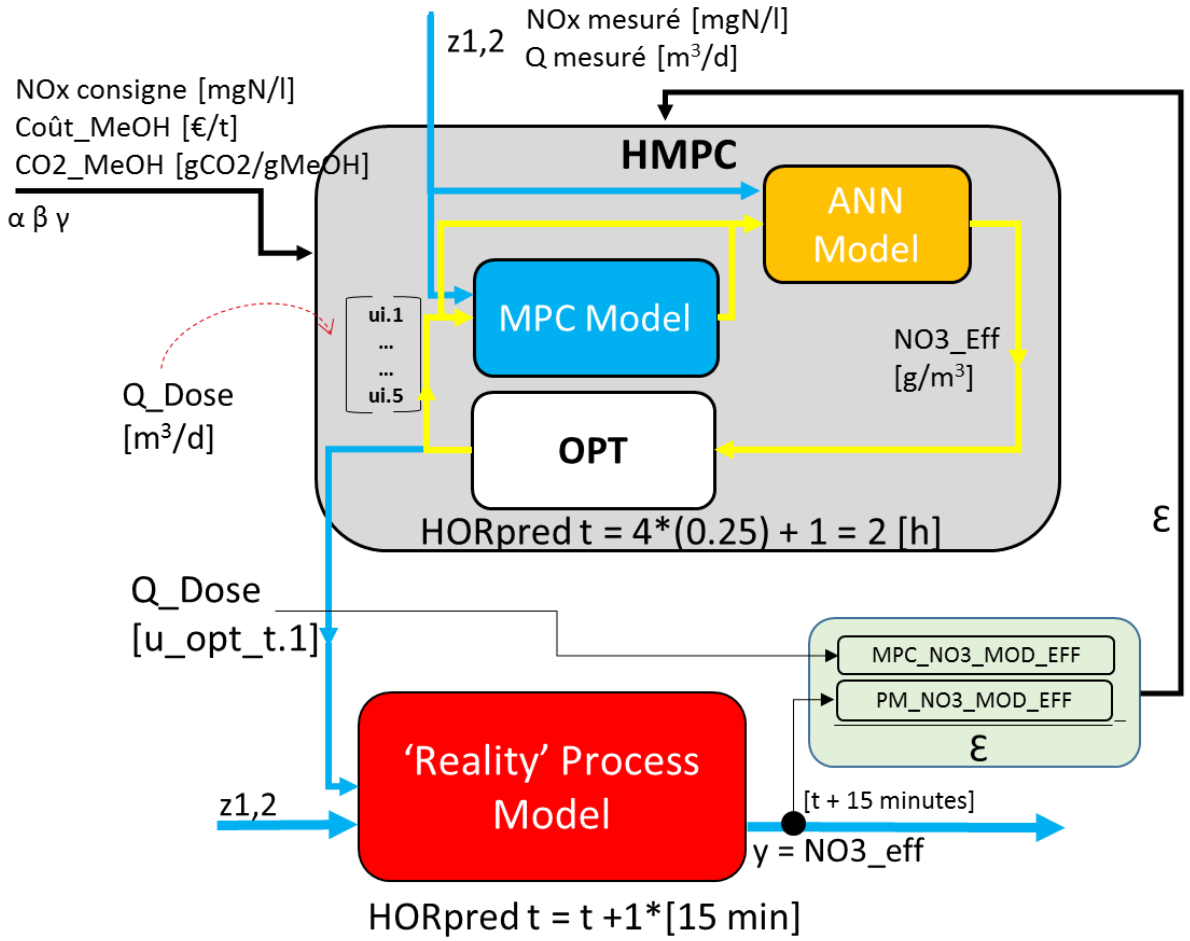


Figure 2: Schéma conceptuel du contrôleur prédictif de modèle hybride développé (boîte grise) qui implique un modèle mécaniste (boîte bleue) interagissant avec un DDM de correction (boîte jaune) pour trouver la valeur d'action de contrôle optimale qui est implémentée dans le modèle de processus "Reality" (boîte rouge boîte). Ce modèle fait ensuite une prédiction à une étape de la qualité de l'effluent, qui est utilisée comme effluent « observé » pour la prochaine itération du HMPC.

Le modèle HMPC reçoit les données d'influent mesurées et des informations pertinentes sur les paramètres opérationnels (par exemple, le nombre de biofiltres actifs). Ces informations pourraient facilement être étendues pour inclure le coût du dosage du méthanol et les coûts environnementaux (par exemple, les émissions de GES) associés au dosage du méthanol.

L'optimiseur recherche la plus petite erreur entre les concentrations de nitrate simulées et la consigne souhaitée. Le débit de dosage optimal correspondant est alors identifié et transmis à un second modèle de biofiltre mécaniste (le *« Reality » Process Model* dans la boîte rouge), qui représente la réalité à laquelle ce contrôleur HMPC serait éventuellement connecté. Il est important de réaliser que ce modèle de processus de « réalité » est différent du modèle MPC d'une manière qui se rapproche de la différence qui se produirait entre le modèle MPC et la réalité à laquelle le HMPC doit être appliqué. Ce modèle représente le processus physique se déroulant dans le monde physique et permet de quantifier l'impact de l'action proposée.

Les fichiers influents et les paramètres opérationnels (par exemple, les informations sur le nombre de filtres actifs dans une étape de traitement) sont identiques pour le modèle MPC et le modèle de processus de réalité. L'initialisation des deux modèles est donc identique. Cependant, en raison de certaines différences de paramètres de modèle (voir section 6.3.3), la condition initiale de chaque modèle dans la simulation dynamique sera différente, de même que leur comportement pendant la simulation dynamique.

La fonction objective de l'optimiseur vise à minimiser l'erreur entre la concentration en nitrate souhaitée, qui est la consigne, et l'effluent en nitrate modélisé. Il cherchera à manipuler le dosage du carbone externe dans le biofiltre simulé dans le modèle MPC pour atteindre l'objectif défini, malgré les perturbations survenant dans le débit et les concentrations de l'influent. Dans l'ouvrage de contrôle développé, la variable procédé (PV) est la concentration de l'effluent nitrate et la variable manipulée (MV) est le débit de dosage du méthanol (Q_Dose) exprimé en m^3/j ; la sortie souhaitée est la concentration de nitrate de consigne ($mgNO_3$-N/L) dans l'effluent (SP_NO3).

Comme indiqué sur la Figure 3, l'horizon de prédiction du MPC est fixé à 120 minutes, ce qui laisse suffisamment de temps pour que l'effet d'une action de dosage de méthanol sur le biofiltre passe complètement à travers le biofiltre. La taille du pas de temps (k) a été choisie à 15 minutes car le temps de rétention hydraulique de l'eau dans un biofiltre est de l'ordre de 20 à 30 minutes. Toute dose de réactif injectée à l'instant k traversera le biofiltre en 20 à 30 minutes, et son effet sur l'activité biologique est supposé être du même ordre. L'horizon de contrôle est défini comme les quatre premiers pas de temps (donc de k à k+4), dont en réalité seule la première action proposée est effectivement transmise au système réel (ici le modèle de processus "Réalité"). Après la première itération, le HMPC remonte d'un pas de temps (k+1) dans un horizon de prédiction fuyant, et recommence l'optimisation. Dans cette étude de cas, cette fenêtre de simulation mobile est répétée pendant une durée totale de 3 heures.

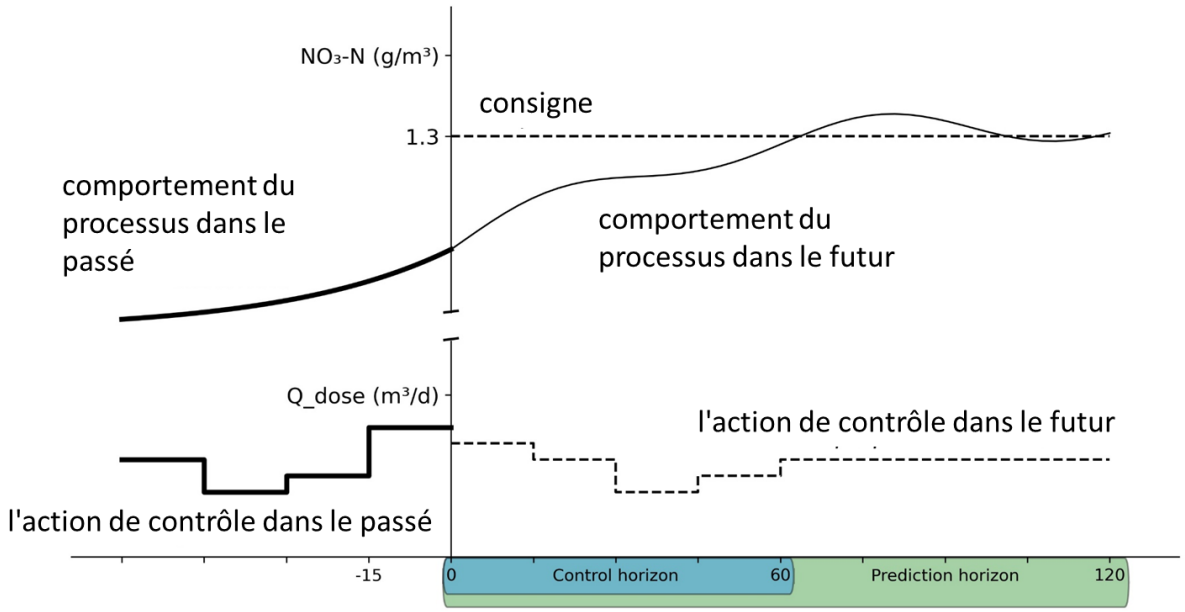


Figure 3: Stratégie de contrôle (H)MPC appliquée montrant l'horizon de contrôle de 60 minutes et l'horizon de prédiction de 120 minutes sur l'axe horizontal ; ainsi que la valeur de consigne souhaitée et le débit de dosage (Q_Dose) sur l'axe vertical.

Outre le débit de dosage variable (Q_Dose) et le point de consigne souhaité, toutes les autres entrées du contrôleur sont considérées comme des perturbations dans des conditions

« idéales » sans tenir compte du bruit de mesure, des retards ou d'autres écarts de surveillance. Dans une extension future de ce projet de démonstration, le *Reality Process Model* peut facilement être remplacé par des mesures en temps réel des concentrations d'effluents de nitrate.

Le contrôleur HMPC était codé en langage python (3.9) afin de permettre l'exécution et le couplage automatique du *modèle de ML* (développé en python) et des modèles de biofiltre, qui ont été développés dans le logiciel WEST (DHI).

2. Résultats et discussion

2.1. Résultats du traitement des données

Les résultats du traitement des données brutes collectées sont présentés dans le tableau 1 pour la période de calibration (colonne de gauche) et pour la période de validation (colonne de droite). En moyenne, les données mesurées pendant la période de calibration contiennent environ 10 500 mesures par variable, dont 4 % ont été classées comme étant aberrantes. Les capteurs mesurant les MES (16 %) et l'oxygène dissous (OD) (28 %) en particulier présentent un pourcentage élevé de données aberrantes. La colonne « Données supprimées (%) » combine les valeurs considérées comme aberrantes par l'étape de lissage exponentiel, ainsi que les données supprimées car considérées comme invalides (Alferes et Vanrolleghem, 2016). Sur l'ensemble des variables traitées, en moyenne 16% des données brutes ne sont pas retenues, avec des pourcentages de perte élevés pour l'OD (52%) et les orthophosphates (48%) mesurés respectivement à l'entrée et à la sortie du DAM. Alferes et Vanrolleghem (2016) confirment des pertes typiques de 5 à 50% des données provenant des sondes immergées dans des eaux usées.

Durant la période de validation, en moyenne 1 % de valeurs aberrantes ont été détectées et 12 % de données ont été supprimées par le lissage (tableau 1). La qualité des mesures de MES en sortie de DAM et en sortie de traitement des boues est particulièrement faible (13 % et 33 % de suppression, respectivement). Cependant, la période de validation présente des pourcentages de perte plus faibles que lors de la calibration.

Tableau 1 : Résultats de la méthode de nettoyage des données appliquée aux données haute résolution pour la période d'étalonnage et de validation.

		Calibration			**Validation**	
Variable	N	Valeurs aberrantes %)	Données supprimées (%)	N	Valeurs aberrantes (%)	Données supprimées (%)
Débit d'eau recirculée	10657	0,5	8,1	8832	0,1	8,4
Débit d'eau décantée	10657	0,2	9	8832	0,1	5,3
DCO entrée DAM	10036	1,9	11,1	8832	0,32	11,6
NOx entrée DAM	10231	1,3	3,5	8832	0,02	6,7
NOx recirculation	10654	0,2	7,6	8832	0,1	7,1
NH4 sortie DAM	10573	0,3	6,9	8832	0,2	9,2
MeS sortie DAM	10654	16,1	18,6	8832	4,9	12,9
PO4 sortie DAM	10626	1,4	48,3	8796	2,2	9,1
Débit traitement des boues	10652	0,06	13	8832	0,4	15,3

Débit refroidissement	10652	2,4	18,6	8832	0,7	9,3
NH4 recirculation	10627	0,22	10,9	8831	1,2	19
MeS sortie traitement des boues	10656	0,04	33,4	8832	6,7	33,2
DCO sortie DAM BAT C	-	-	-	8825	0,2	14,4
NOx sortie DAM BAT C	-	-	-	8832	0,1	4,4
Moyenne	10544	3,8	16,1	8829	1,2	11,9

2.2. Résultats de l'analyse de sensibilité des paramètres du modèle mécaniste

L'analyse par régression linéaire multiple a identifié des paramètres comme étant influents avec des coefficients de régression standardisée absolus supérieurs à 0,1. Les paramètres retenus et recalibrés (voir tableau 2) caractérisent la filtration et le rétrolavage, le transfert de matière au sein du biofilm par diffusion, la nature du biofilm et les cinétiques biologiques.

Par rapport aux valeurs du modèle d'origine, les paramètres ont été modifiés de plus de 20 % en moyenne. Cet ajustement peut s'expliquer par des modifications des conditions d'exploitation des biofiltres et des infrastructures amont ou par une évolution des populations bactériennes et de leur métabolisme entre la calibration actuelle et celle précédemment réalisée.

Des modifications importantes des paramètres qui décrivent l'activité de la biomasse hétérotrophe en anoxie sont proposées. Parmi ceux-ci, le facteur de réduction de la vitesse de croissance en réduisant le nitrite (n_h3 : -34%), et le monoxyde d'azote (n_h2 ; +36%) et la constante de demi-saturation en DCO soluble biodégradable (K_S_NO3 : +45%). Au total, 9 paramètres caractérisant les cinétiques biologiques et 2 paramètres caractérisant la filtration ont été modifiés.

Tableau 2: Résultats de l'analyse de sensibilité des paramètres du modèle mécaniste modifiés

Description des paramètres	Symbole	Unité	Valeur modèle JZ (Zhu, 2020)	Espace attribuée	Valeur recalibrée	Différence
Composition						
Quantité d'azote dans la biomasse	i_X_B	gN/gCOD	0,098	20%	0,109	+11 %
Quantité de phosphore dans la biomasse	i_X_BP	gP/gCOD	0,0055	50%	0,0046	-16 %
Quantité de phosphore dans les produits de la mort bactérienne	i_X_UP	gP/gCOD	0,006	50%	0,008	+33 %
Filtration						
Coefficient de filtration sur le filtre propre	lami	1/m	0,4	50%	0,43	+8 %
Constante empirique y dans l'équation de coefficient de filtration	yy	-	0,83	50%	1,06	+28 %
Cinétique						

Constante de demi-saturation en DCO soluble biodégradable pour la biomasse hétérotrophe consommant le nitrate en anoxie	K_S_NO2	gCOD/m^3	13,8	20%	16,06	+16 %
Constante de demi-saturation en DCO soluble biodégradable pour la biomasse hétérotrophe consommant le nitrite en anoxie	K_S_NO3	gCOD/m^3	28,6	20%	41,37	+45 %
Taux de mortalité de la biomasse hétérotrophe	b_H	1/j	0,7	20%	0,73	+4 %
Taux d'ammonification	k_a	m^3/(gCOD*j)	0,075	20%	0,065	-13 %
Taux d'hydrolyse	k_h	gCOD/(gCOD*j)	2,5	50%	2,6	+4 %
Taux spécifique maximal de croissance de la biomasse hétérotrophe	mu_H	1/j	5,18	20%	4,74	-8 %
Facteur de réduction de la vitesse de croissance en anoxie de la biomasse hétérotrophe en réduisant le nitrite	n_h2	-	0,14	50%	0,19	+36 %
Facteur de réduction de la vitesse de croissance en anoxie de la biomasse hétérotrophe en réduisant le nitrate	n_h3	-	0,47	50%	0,31	-34 %
Facteur de correction du rendement hétérotrophe en anoxie	n_hy	-	0,78	50%	0,72	-8 %

2.3. Résultats de calibration du modèle mécaniste

Le tableau 3 présente les valeurs des critères statistiques calculés à partir des simulations de l'effluent de DAM en période de calibration et validation. Les valeurs de EM indiquent que le biais du modèle est relativement faible par rapport aux valeurs moyennes observées en particulier pour les variables DCO totale, DCO soluble et MES. Dans le cas de N-NO_3 et du N-NO_2, les valeurs négatives d'EM indiquent que le modèle surestime la cinétique de dénitrification. En fait, les mesures de dioxygène dissous utilisées pour alimenter l'entrée du modèle montrent un pourcentage élevé de valeurs aberrantes et de données supprimées (28%). Ces lacunes dans les données sont ensuite comblées par interpolation, ce qui peut entraîner des valeurs non précises prises en compte par le modèle.

Tableau 3: Résultats en termes d'évaluation des simulations de l'effluent du DAM pour la période de calibration et de validation

		Calibration					Validation					
Variable	Unité	N	Moyenne observée	ME	MAE/ ȳ	RMSE/ ȳ	N	Moyenne observée	ME	MAE/ȳ	RMSE /ȳ	Coëff. Janus
NO_3	gN/m^3	10656	2,89	-1,35	0,67	0,86	8015	2,62	-3,22	1,29	1,46	1,53
NO_2	gN/m^3	10656	0,31	0,10	0,45	0,71	8015	0,38	0,13	0,50	1,27	1,36
DCO	gO_2/m^3	10656	89,07	-1,05	0,14	0,20	7218	64,84	-34,38	0,55	0,60	2,20
DCOs	gO_2/m3	10656	43,30	1,73	0,14	0,21	7218	43,92	-15,49	0,39	0,45	2,13
MES	g/m^3	10656	21,20	-2,27	0,23	0,39	7399	20,12	-7,18	0,46	0,63	1,54

Le coefficient de Janus indique que les estimations du modèle sont fiables et que le modèle peut être considéré comme valide pour les composants azotés, avec des valeurs considérablement en dessous de la valeur critique de 2. En revanche, les coefficients de Janus calculés pour la DCO totale et la DCO soluble (2,20 et 2,13 respectivement) sont légèrement supérieurs à 2, ce qui est expliqué par un REQM faible sur la période de calibration (20% et 21% respectivement) et significativement plus élevé sur la période de validation (60% et 45% respectivement).

La Figure 4 présente les résultats de calibration des nitrates en sortie de DAM durant le mois de décembre 2019. Les valeurs journalières moyennes modélisées s'élèvent à 4,67 mg N-NO_3/L, tandis que la concentration moyenne en nitrate observée est de 5,34 mgN-NO_3/l. Les scores statistiques calculés sont : EM à 0,67 mgN-NO_3/L, EMA à 1,08 mgN-NO_3/L et REQM à 1,36 mgN-NO_3/L. Le diagramme de parité (figure b), qui permet une comparaison visuelle des valeurs prédites en fonction des observations avec les valeurs optimales se trouvant sur la ligne diagonale, montre que le modèle est capable de simuler la conversion de N-NO_3 en N_2 par les bactéries.

La concentration en nitrites (N-NO_2) en sortie du système est quant à elle mal estimée par le modèle sur cette période, puisque la valeur modélisée moyenne est deux fois plus basse (0,12 mgN/L) que la moyenne des observations (0,26 mgN/L). Cet écart se reflète dans les scores statistiques : EM à 0,14 mgN/L, EMA à 0,14 mgN/L et REQM à 0,18 mgN/L. L'estimation des concentrations en nitrites est difficile car le nitrite est un composé instable qui s'oxyde rapidement en nitrates. De plus, les faibles niveaux normalement observés se rapprochent souvent des limites de détection des sondes. Cette difficulté métrologique se reflète donc dans la qualité d'estimation du modèle. Les dynamiques des variations de DCO totale sur la figure 4c et DCO soluble (Figure 4d) sont correctement évaluées par le modèle, même si l'amplitude de celles-ci est sous-estimée.

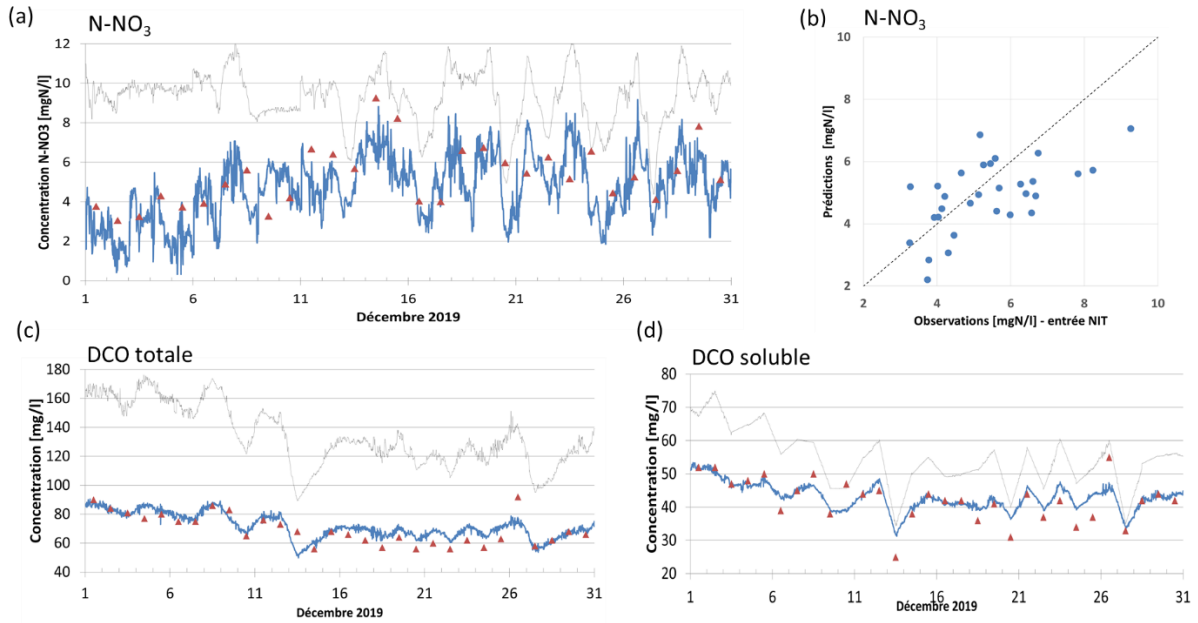


Figure 4: Résultats du calage du modèle de qualité des eaux en sortie de DAM. a) Résultats de calibration de la concentration en N-NO3 en sortie de DAM sur une période de 30 jours en décembre 2019 (triangles bruns : observations laboratoires, ligne bleue : résultats de simulation, ligne noire : concentration en entrée du procédé de DAM). b) Concentration moyenne journalière prédite par le modèle en fonction de la concentration observée. La ligne noire pointillée représente l'estimation

optimale. c) Résultats de calibration de la concentration en DCO totale. d) Résultats de calibration de la concentration en DCO soluble.

2.4. Résultats de validation du modèle mécaniste

La figure 5 récapitule les estimations des concentrations en $N\text{-}NO_3$ (a), DCO totale (c) et DCO soluble (e) en sortie de la DAM sur la période de validation. La concentration en $N\text{-}NO_3$ prédite présente des écarts significatifs aux observations, en particulier à partir de début août, ce qui se traduit par un diagramme de parité assez faible (figure 5b). Le débit arrivant à l'usine et la charge en matière organique diminue drastiquement en août en raison d'une réduction de population en période estivale.

Les estimations de DCO totale et soluble sont globalement assez satisfaisantes, mais sont systématiquement surévaluées (figure 5c à f). De même, l'effluent réel est moins concentré en nitrates que ne le prévoit le modèle, en particulier durant la période de fin juillet à mi-août 2020. Cette surestimation de DCO et de nitrates en sortie de DAM peut s'expliquer par une activité métabolique insuffisante du biofilm dans le modèle sur la période de validation. Ceci pourrait traduire une prise en compte partielle des effets de la température sur les cinétiques biologiques.

De nombreux autres facteurs peuvent introduire les erreurs d'estimation, comme la précision des valeurs mesurées (en particulier pour les capteurs en ligne qui montrent de grandes variabilités), le fractionnement des données mesurées de DCO totale et DCO soluble en variables d'état utilisées par le modèle.

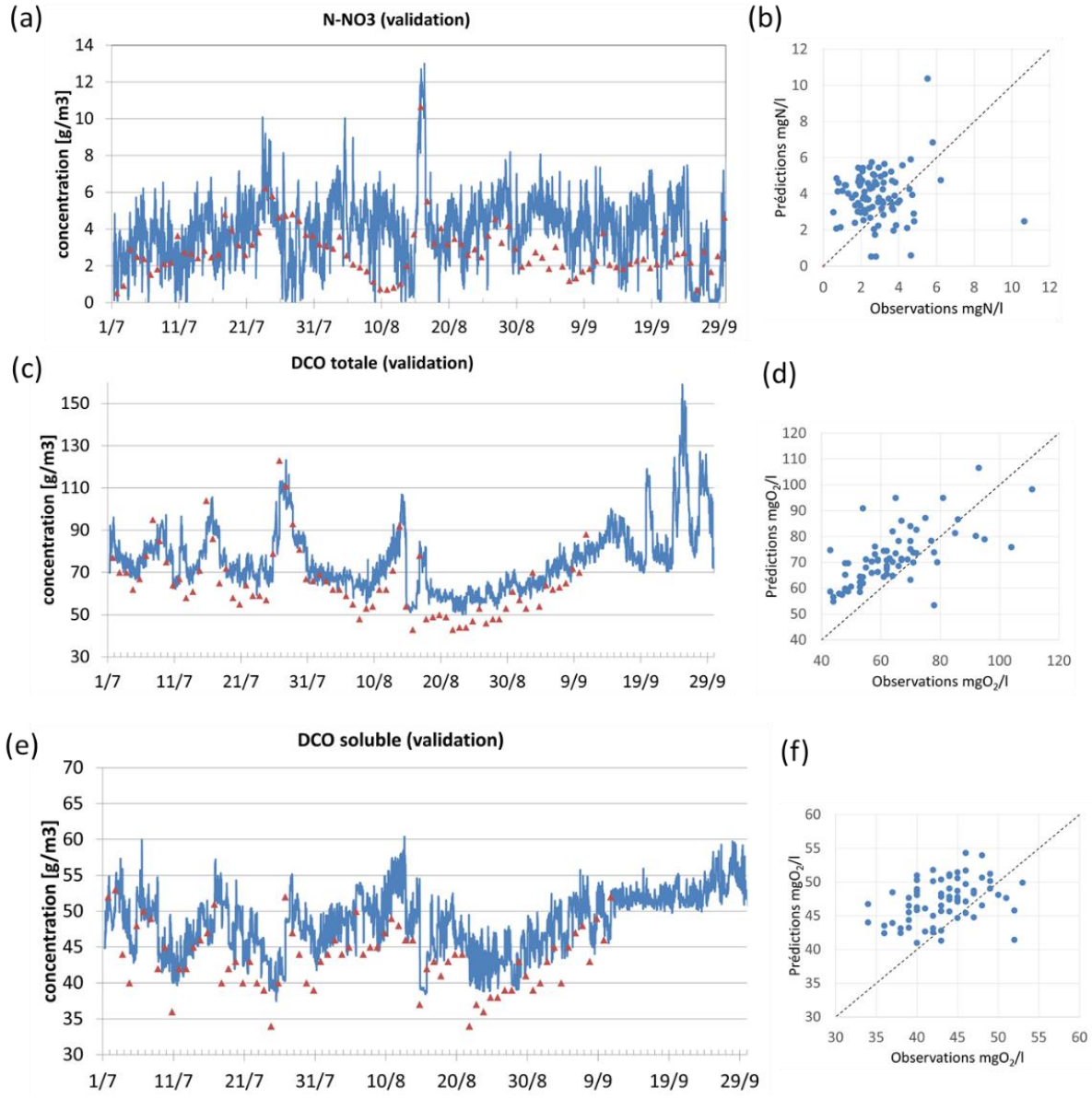


Figure 5: a) Résultats de validation de la concentration en N-NO_3 en sortie de la DAM ; b) diagramme de parité présentant les concentrations prédites en N-NO_3 en fonction des concentrations observées ; c et d) Résultats de validation de la concentration en DCO totale ; e et f) Résultats de validation de la concentration en DCO soluble. a, c et e), triangles bruns : observations laboratoires, ligne bleue : résultats de simulation.

2.5. Résultats d'entrainement et test du modèle hybride

L'algorithme de ML le plus performant a été sélectionné sur le critère de la REQM la plus faible, et se compose d'un réseau de neurones à trois couches comportant 12 neurones dans la couche d'entrée, 25 neurones dans la couche intermédiaire et un neurone dans la couche de sortie en appliquant une fonction d'activation Rectified Linear Unit (ReLU). Un réglage optimal des hyper-paramètres du modèle a été recherché afin d'éviter des conditions de sous-ajustement ou de sur-ajustement sur les deux périodes d'entraînement et de test.

Le tableau 4 présente les valeurs des indicateurs statistiques pour les simulations de N-NO_3 en sortie de DAM en période de calibration et validation pour le modèle hybride. Pendant la phase d'entrainement, les performances du modèle sont très bonnes et les EM, EMA et REQM sont très faibles. Lors de la phase de validation (1 – 15 mars 2020), la qualité d'estimation

diminue : l'EM passe de 2 % à 8 %, l'EMA de 20% à 24% et la REQM de 27% à 32%. Néanmoins, les scores restent globalement acceptables, ce qui montre que le modèle est capable de prédire l'évolution du système sur un jeu de données avec lequel il n'a pas été entrainé.

Tableau 4: Modèle hybride résultats des scores statistiques pour les périodes de calibration (01.12 – 31.12.2019) et validation (01.03 – 15.03.2020) pour les variables simulées en sortie de DAM.

Période	Variable	Unité	N	Moyenne observée	EM	EM/ŷ	EMA	EMA/ŷ	RMSE	RMSE/ŷ
Training	NO_3	gN/m^3	2881	3,17	0,07	2%	0,64	20%	0,85	27%
Testing	NO_3	gN/m^3	1440	3,23	0,27	8%	0,76	24%	1,03	32%

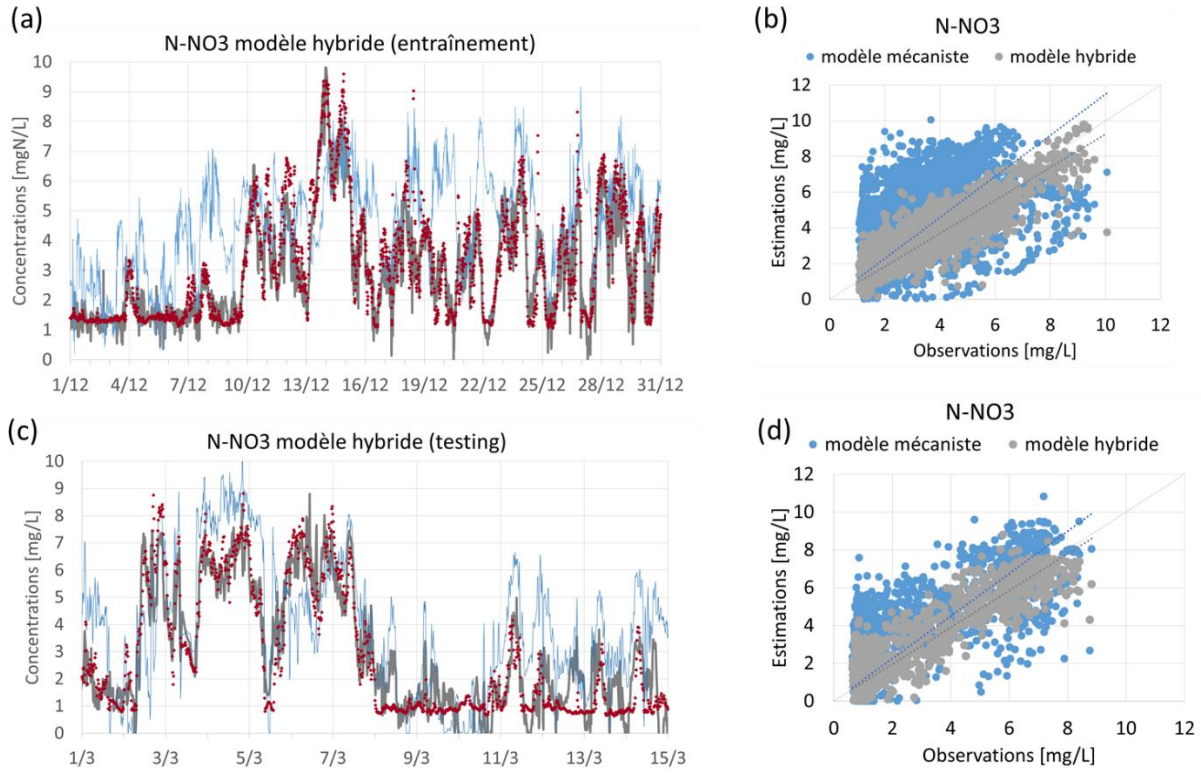


Figure 6: Résultats de l'entraînement (a) et de test (c) du modèle hybride de la concentration en N-NO_3 en sortie de DAM. Ligne bleue : simulation du modèle mécaniste ; ligne grise : résultats du modèle hybride ; points orange : observations haute fréquence. Figures b et d) concentrations prédites par le modèle mécaniste (bleu) et hybride (grise) en fonction de la concentration observée. La ligne noire pointillée représente l'estimation optimale.

La figure 6 récapitule les estimations de concentration en nitrates obtenues pour les périodes de calibration, (a) et (b), et validation, (c) et (d), par les modèles hybride et mécaniste. Les estimations du modèle hybride sont beaucoup plus proches des valeurs observées que celles du modèle mécaniste sur les deux périodes. Cependant, la performance du modèle hybride diminue sur la période de validation par rapport à celle de calibration, comme relevé lors de l'analyse des indicateurs statistiques. Les modèles de ML émettent des estimations correctes dans les conditions de leur apprentissage ; les modifications de débits et de température entre décembre et mars ont dû limiter les capacités de généralisation du modèle hybride. Une période plus longue d'entrainement ou un ré-entraînement régulier du réseau de neurones permettrait d'augmenter la robustesse du modèle.

On observe malgré tout que le modèle hybride prédit l'ensemble des dynamiques de la concentration en nitrates sur les deux périodes et prédit des variations de même amplitude que celles des observations dès le début de la simulation, d'une manière plus précise que le modèle mécaniste.

1.8 Résultats de HMPC

Les simulations ont été effectuées avec l'HMPC contrôleur pour des scénarios qui testent différentes structures d'un contrôleur prédictif de modèle dans différentes conditions. Toutes les simulations ont utilisé les mêmes données d'entrée pour la période du 1er au 8 juin, une période caractérisée par des concentrations élevées de N-NO_3 dans l'affluent et un dosage réel de méthanol dans le système physique.

La figure 7 montre les performances de chaque scénario en comparant la concentration de N-NO_3 simulée dans le *Reality Process Model* avec le point de consigne de nitrate.

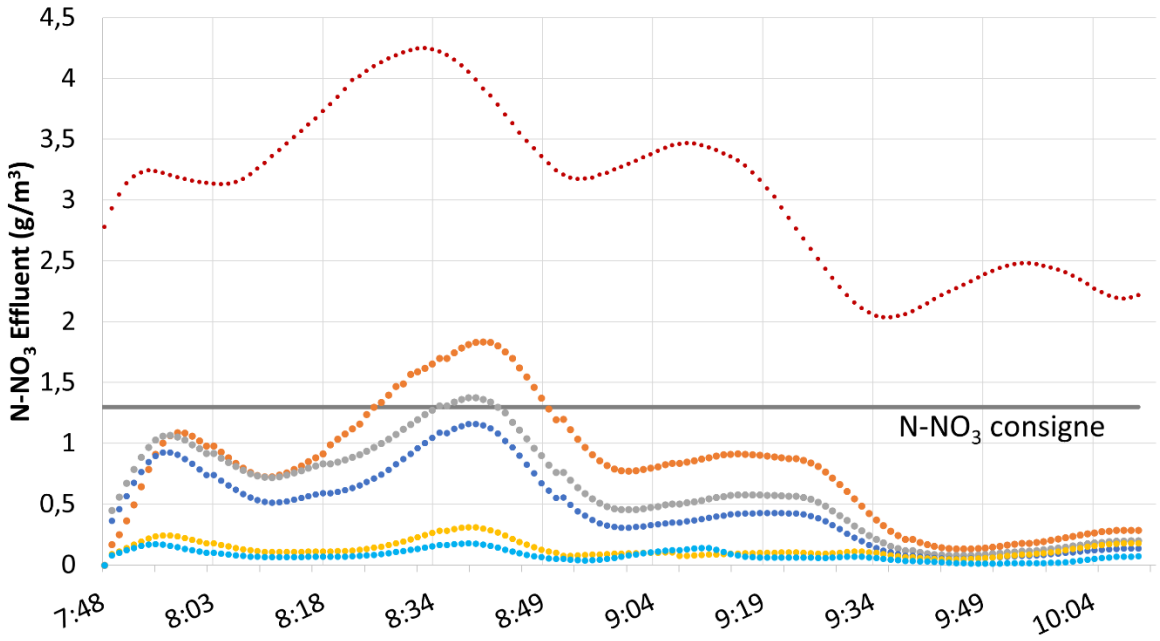


Figure 7 : Concentrations d'effluents N-NO_3 simulées dans le modèle de processus réel des divers scénarios de haut niveau utilisant des variantes MPC et HMPC, ainsi que l'effet des paramètres de contrôle de dosage actuels. Sont indiqués le résultat actuel du contrôleur (couleur rouge), MPC_A (orange), MPC_B (gris), MPC_C (bleu foncé), HMPC_B (jaune) et HMPC_C (bleu clair).

Comme on peut le voir sur la figure, la performance des paramètres actuels de contrôle du dosage est la référence à laquelle ce travail doit être comparé. L'amélioration significative des contrôleurs MPC et HMPC est clairement visible par rapport au contrôle actuellement appliqué. L'impact des stratégies MPC et HMPC est une réduction de la concentration en nitrate pendant la période étudiée, ce qui se traduit par une amélioration de la qualité des effluents. Nous pouvons constater que les variantes MPC ont une erreur inférieure à celle des variantes HMPC. Un problème probable est que le modèle ML a été formé sur la « mauvaise » erreur résiduelle. Au lieu de définir l'erreur résiduelle comme la différence entre les mesures des nitrates en sortie de DAM et les concentrations modélisées de nitrates, le modèle ML dans cette démonstration de HMPC devrait être entrainé sur la différence entre le modèle représentant le processus réel et le modèle utilisé dans le MPC, qui est un modèle simplifié.

En raison des caractéristiques différentes des deux modèles, l'effluent modélisé des deux modèles n'est pas la même. En entraînant un modèle ML sur cette différence, on s'attend à ce que le HMPC soit capable de corriger les résidus et ainsi de donner l'erreur le plus bas. Ce travail sera finalisé et soumis pour publication ultérieurement.

2 Conclusion

Dans un premier temps, un modèle mécaniste précédemment développé (Zhu, 2020) a été recalibré avec des données de l'étage de DAM de la filière de biofiltration de l'usine Seine aval. L'analyse de sensibilité globale a identifié les paramètres les plus influents sur la réponse du modèle : ceux qui caractérisent la composition, la filtration et les réactions biologiques. Les résultats de simulation avec les paramètres calibrés montrent que le modèle arrive pendant la calibration à prédire les variations de $N\text{-}NO_3$, DCO et MES avec des performances raisonnables. Sa robustesse d'estimation avec le jeu de données de validation de $N\text{-}NO_3$ est cependant insuffisante sans l'apport de l'hybridation.

Ensuite, un modèle hydride parallèle coopératif a été développé dans lequel le modèle mécaniste recalibré est soutenu par un modèle de ML dans le but d'améliorer la précision d'estimation de qualité de l'effluent en sortie des biofiltres. Différentes techniques de ML ont été appliquées pour obtenir une estimation précise de l'erreur du modèle mécaniste sur les estimations de concentrations en $N\text{-}NO_3$, MES et DCO totale en sortie de DAM. Ces modèles sont alimentés par les qualités d'effluent simulées par le modèle mécaniste, ainsi que par les informations de la qualité d'eau en entrée et les paramètres de traitement (tels les consignes). En phase d'entrainement, l'erreur résiduelle est fournie au modèle. Les meilleurs résultats ont été obtenus avec un réseau de neurones. Les estimations obtenues sur la période de calibration et celle de validation montrent que la précision du modèle hybride est largement supérieure à celle du modèle mécaniste seul. L'hybridation est donc une approche extrêmement pertinente dans le cas de la modélisation d'un procédé complexe de traitement des eaux.

L'utilisation de données couvrant des périodes plus longues permettrait probablement d'atteindre une plus grande précision. Dans un contexte opérationnel, un tel modèle hybride doit être ré-entraîné dès que de nouvelles données sont disponibles (Schneider et al., 2022). La fréquence d'entraînement du modèle et la taille du jeu de données à utiliser reste à déterminer.

Cette étude montre qu'un modèle hybride permet de mieux décrire le comportement de dénitrification d'une usine d'épuration de très grande échelle qu'un modèle mécaniste seul, améliorant ainsi sensiblement la qualité des estimations. C'est une première étape vers la conception de jumeaux numériques intégrés aux filières d'assainissement.

Les résultats de simulations avec le contrôleur prédictif du modèle hybride (HMPC) réduit considérablement les concentrations de $N\text{-}NO_3$ dans les effluents et améliore ainsi la qualité des effluents. La demande de calcul de l'approche de modèle hybride utilisant un script Python appelant WEST est beaucoup plus faible par rapport à l'exécution des simulations de modèles mécanistes dans l'interface utilisateur graphique de WEST. Ceci a été réalisé en développant des scripts Python qui exécutent le modèle mécaniste et les modèles d'apprentissage automatique automatiquement et sans la surcharge d'une interface utilisateur graphique.

D'autres études des stratégies de contrôle à l'aide du modèle hybride sont nécessaires pour comprendre les faibles valeurs de nitrate simulées. L'optimiseur HMPC peut certainement

bénéficier d'un réentraînement régulier du réseau de neurones. Cela augmenterait encore la confiance dans les performances du modèle qui permet de l'appliquer dans un cadre de contrôle de processus actif.

Ce résumé a été soumis pour publication dans l'ouvrage de programme de recherche Mocopée Phase II, qui sera publié en novembre 2023 par le SIAAP-Direction Innovation.

List of Contents

List of Figures

List of Tables

List of Abbreviations

ANN	Artificial neural network
ASM	Activated Sludge Model
BOD	Biochemical oxygen demand
BAF	Biological Activated Filtration
BSM1	Benchmark Simulation Model No 1
CAS	Conventional activated sludge
COD	Chemical oxygen demand
DDM	Data-driven Model
DT	Digital Twin
DO	Dissolved oxygen
HM	Hybrid Model
HMPC	Hybrid Model Predictive Controller
HRT	Hydraulic retention time
MAE	Mean Absolute Error
MBR	Membrane bioreactor
ME	Mean Error
MIMO	Multiple In Multiple Out
MM	Mechanistic Model
MPC	Model predictive control
NH_4-N	Ammonia
NN	Neural networks
NO_2-N	Nitrite
NO_3-N	Nitrate
RMSE	Root Mean Square Error
SCADA	Supervisory Control and Data Acquisition
WRRF	Water Resources Recovery Facility
WWTP	Wastewater Treatment Plant

Acknowledgements

For funding and supporting this research through the Mocopée research program, I would like to thank SIAAP, the French public authority responsible for the management and treatment of municipal wastewater in the Greater Paris region, and in particular the personnel at SIAAP-Direction Innovation guided by Dr. Vincent Rocher and Dr. Sam Azimi. Their industry experience and expertise helped me to better understand the modern challenges we are faced with in wastewater treatment. Their '*esprit*' for applied research and openness to share information allowed me to make the most of my research. Thank you to the operators of the Seine Aval wastewater plant for all the data!

Moreover, for supporting my research I would like to thank the Laboratoire Eau, Environnement et Systèmes Urbains LEESU at the Ecole des ponts (ENPC)/ParisTech (France), model*EAU* at the Université Laval (Canada), the W-SMART association (France) and BioMATH at the UGhent (Belgium).

I express my deepest thanks to my mentors Professor Bruno Tassin (ENPC), Director of this research, and Professor Peter Vanrolleghem (ULaval), co-Director of this research, for showing guidance and support over these three intense years in useful decision-making and for giving necessary advice. I could not have accomplished this project without your guidance. I choose this moment to acknowledge their contribution gratefully.

Thank you to Dr. Saba Daneshgar and Loes Verhaeghe from BioMath (UGhent), who have supported the development of the hybrid approach. Dr. Saba Daneshgar has shown great dedication to this research; thank you to be always there to discuss the development of the Hybrid Model Predictive Controller. I'm looking forward to carrying over our cooperation into future research projects.

A special word of thanks I dedicate to Professor Ilan Juran (NYU, US), who has shown to be a visionary leader dedicated to research at the finest level. He recognized the potential for this research early on and played a key role in the initial start-up as the Director of the W-SMART association and has built bridges in the international research arena.

The advisory committee overseeing the progress during regular meetings has been instrumental in my development as a researcher. For that, I would like to thank each member: Dr. Vincent Jauzein, Dr. Eloïse Detredern, Dr. Sam Azimi and Dr. Vincent Rocher (all SIAAP-DI), Dr. Gilles Varrault (U-PEC) and Dr. Yannick Fayolle (INRAE), as well as Dr. Jean Bernier (currently SUEZ). I would like to thank the additional members of the jury for taking time to evaluate my research work: Dr. Sylvie Gillot (INRAE), Dr. Ignasi Rodriguez Roda Layret (Universitat de Girona), Dr. Claire Albasi (CNRS) and Dr. Elena Torfs (ULaval).

There are countless students, faculty members and staff at LEESU and model*EAU* that have supported me throughout this voyage. I want to thank Dr. Regis Moillerion, the Director of LEESU, for his confidence, as well as Dr. Max Dechesne, Dr. Angélique Goffin, Dr. Niels Nicolaï, Dr. Thomas Maere, Dr. Gamze Kirim, Dr. Sovanna Tik, Dr. Feiyi Li, Karen Mesta, Jean-David Therrien, Daniel Mendoza and my digital twin wingman Jeffrey Sparks. I've also had the honor to brainstorm at regular occasions with Sam White and Shahryar Pasha Zadeh Monajjemi, both PhD students at NYU supervised by Professor Ilan Juran. You have all taught me about how to develop a mindset for PhD research and how to push

the boundaries of Digital Twins. A special thanks to Dr. Jialu Zhu, who developed the biofiltration model used here and providing technical support.

My sincere appreciation to DHI for providing access to the wastewater treatment process modelling platform WEST and in particular to Dr. Enrico Regimi for his technical support; as well as to MathWorks for providing the MATLAB software university license.

I am forever grateful to my family and close friends who provided me strength and wisdom whenever I needed support and gave me peace-of-mind to think when that is what was needed. Thank you for your love, energy, and compassion, first to my dearest wife Deborah Leah Serrao and our amazingly supportive children; Gabriel, Ethan and Yoel Serrao, who have adapted smartly to the changing conditions; may they find this research period inspirational and learn that "*The impossible because possible if you really put your mind to it*!"; as well as to my parents Deborah Serrao and Dr. Mikael Snidecor, my sisters Alexandra and Melanie and brother Joshua. Thank you to my friends who showed interest and asked me regularly "*How are things going?*" and showed me support when it wasn't; in particular family Weisman, family Cohen-Salmon, family Walk, Oriel Lubowski, David Belhassen, Michel Bousidan, Albert Cohen, Patrick Boudjnah, Luis Martins Dias and Daan Gomes deMesquita.

My sincere appreciation to the Natural Sciences and Engineering Research Council of Canada, who awarded model*EAU* the "2021-2026 Discovery Grant Towards digital twin based control of water resource recovery facilities - Methods supporting the use of adaptive hybrid digital twins".

1 Introduction

Digitalization of the wastewater sector is gaining importance globally with the advancement of data-based monitoring, model-based predictive process control and other tools for data-driven decision-making. Literature points out significant potential benefits thanks to increased collection and utilization of data (Eggimann et al., 2017). The digitalization drivers are many: process efficiency, energy transition, greenhouse gas emission reduction, water quality improvement, recovery of resources and the continuous urbanization, as well as climate change and evolving regulations (e.g. Blumensaat et al., 2019; Garrido-Baserba et al., 2020). Improving the performance and efficiency of wastewater treatment plants (WWTPs) involves strengthening process control (Olsson, 2008). The expected benefits are many: optimization of energy consumption, reduction of reagents usage, lowering of greenhouse gas emissions and overall ecological impact of operations, an increase of the reuse of resources and a stable effluent water quality. Actually, all these aspects are nowadays considered important objectives for the wastewater treatment industry (i.e.: Flores-Alsina et al., 2010; Novak et al., 2015).

1.1 The Power of Models

The use of mathematical modelling in the wastewater sector has been established since the 1990's and is widely recognized as an important tool for research, system design, simulation of processes and operator training (Gernaey et al., 2004). Mechanistic models, whose parameters have a real physical meaning, describe the biological, physical and chemical processes that take place in a treatment reactor based on the domain knowledge of these processes. They are often represented by mass balance equations that change over time and in space. These mechanistic models are also named First-Principles-based or knowledge-based models; they have a strong predictive power and have become accepted tools in the wastewater sector (Rieger et al., 2012).

More recently, the development of digital twins that 'mirror' the behaviour of a physical system or entire plant, and that receive measurement data in (near) real time allow predicting future states of process behaviour (IWA & Xylem, 2019). In addition, an important field of modern research aims to study the potential of data-driven models, including Machine Learning models, to achieve higher performance of wastewater treatment plants (Schneider et al., 2022).

In this study, it will be shown how the hybridization of a data-based model using artificial intelligence techniques with a mechanistic model increases the model's performance. This allows for the application of hybrid models as a decision-making tool for managers and operators in the wastewater industry. A case study then demonstrates the use of Hybrid models in a Model Predictive Controller (HMPC) for optimization of process control in a predenitrification biofilter with near real-time operational data from a super-large WWTP. The HMPC is able to improve process stability and to significantly improve the effluent quality in terms of nitrate. As far as known, the HMPC developed here is one of the first implementations of a hybrid MPC on an operational wastewater treatment facility. The only other known example is described by Sparks et al. (2022) using a similar approach as developed here, but for the application of the control of chlorine dosing and aeration in an activated sludge system.

1.2 Relevance of a Hybrid Approach in Modelling of Biofiltration

In order to design and optimize the operation of wastewater treatment processes, mathematical models are useful tools for engineers and researchers. For the activated sludge process, a simple and

robust model (ASM1; Henze et al., 1987) was proposed more than three decades ago. It describes the removal of organic carbon compounds and nitrogen, with simultaneous consumption of oxygen and nitrate as electron donors, and allows estimating the sludge production (van Loosdrecht et al., 2008). The relative simplicity of this model led to a widespread use by researchers and practitioners; it has become a reference for many scientific and practical modelling projects for biological nitrogen removal. Since then, it has been modified and extended to simulate the dynamic behaviour of phosphorus removal (ASM2, ASM2d, ASM3; Henze et al., 2000).

In contrast to activated sludge system, where the active microorganisms are suspended in the wastewater, in the case of biological treatment by biofiltration, the wastewater flows through a completely submerged porous filter filled with small diameter size (4 – 5 mm) media on which the microorganisms develop as a thin biofilm layer (typical average thickness of 100 µm) attached to these support media. As a consequence of the large bacterial surface over volume ratio in a biofilter, the hydraulic residence time is drastically decreased compared to conventional treatment systems. Biofiltration is considered a compact and reactive treatment process that is typically applied in densely urbanized regions (Metcalf & Eddy, 2014).

Due to the important role of the biofilm, inside which active biomass in large concentrations perform the treatment processes that remove nutrients and pollutants from the wastewater, biofilm models require more complex and additional equations that take into account transport in and transfer between different compartments. The processes occurring inside a biofilter remain difficult to be accurately described mathematically by a mechanistic model due to the many hypotheses and assumptions that underlie the equations (e.g. kinetic equations of biomass growth) and typically large number of model parameters required. The biokinetic equations considered in the ASM cannot be directly applied in a biofiltration model due to the importance of the transfer processes by diffusion into the biofilm and the necessity of regular backwashing of the filter. Washing forms an essential aspect of biofiltration and is typically required once or twice per day for periods of 15 to 30 minutes per session in order to avoid an accumulation of biomass in the biofilm and suspended solids captured between the open spaces in the filter bed (physical separation). Consequently, biofilter models are considered more difficult to apply than the ASM family models and have thus found little application in engineering practice so far (Plattes et al., 2006). Additionally, there are no widely accepted protocols to aid the community in the development of biofilm models (Rittman et al., 2018).

Few biofilm modelling studies applying submerged biofilter systems using operational data are reported in the literature. Most models were developed, calibrated and validated by means of experiments carried out for case studies at different scales: lab-scale (Mann and Stephenson, 1997; Hidaka and Tsuno, 2004), semi-industrial scale (Sanz et al., 1996; Farabegoli et al., 2004), and industrial scale (Viotti et al., 2002). The main differences consist in the mechanisms that are simulated. Most models include a transport model for the particulates along the filter and a conversion model of the soluble compounds (the biological reactions) from the wastewater (‘bulk liquid’) to the biofilm. Few of them include a filtration or a head loss prediction model (Vigne et al., 2010). Although biofilms develop heterogeneously and have a complex constitution of various types of bacteria embedded in a matrix of extracellular polymeric substances and water, almost all biofilm models found in the literature describe a one-dimensional biofilm space, where the concentration gradients of the nutrients, pollutants and biomass are perpendicular to the inert filter bed media (Wanner et al., 2006). Multi-dimensional biofilm models require long computational times and are typically only used to answer

specific questions related to the development of the various heterogeneous structures inside a biofilm (Morgenroth, 2008b).

On the other hand, data-driven models describe a process by finding patterns only based on the information extracted from the provided data. Machine Learning models can rapidly discover and learn 'hidden' patterns in the data and the metadata without assuming any domain knowledge. These data-driven models have strong interpolation strengths and are much faster in computation, which makes them very attractive for real-time data applications (Newhart et al., 2019). DDMs have found application in wastewater treatment as influent generators, where they create dynamic influent data that serve mechanistic models (e.g. Ahnert et al., 2016; Li & Vanrolleghem, 2022); for process monitoring and control, where they predict time series (e.g. Zhu et al., 2018), for fault/anomaly detection in observed data (e.g. Wade et al., 2005) and for estimation of model parameters (e.g. Mesta et al., 2023). These applications support mechanistic modelling and result in improved accuracy of the model's predictions. Another applied example is the deployment of software sensors that output a signal based on an internal model and several other measured input signals. These software sensors form a support for and an alternative to the installation of hardware sensors that is gaining more application in the wastewater industry (e.g. Dürrenmatt & Gujer, 2012).

Finally, model-hybridization is a process that combines a mechanistic model, which integrates relevant knowledge about processes, with a data-driven model that increases the accuracy of estimates by including information about poorly described sub-processes extracted from data. In literature, hybrid models have been successfully applied to mechanistic models in experimental and/or small-scale set-ups (e.g. Anderson et al., 2000; Lee et al., 2005b; Lee et al., 2008; Dalmau et al., 2015; Cheng et al., 2021). The main configurations of hybrid models that have been studied in literature are a serial and parallel structure. In a serial HM, the output of a data-driven model is used as input of the mechanistic model (or the inverse, although less applied), while in a parallel HM both model types cooperate to produce an improved final output. A novel approach is to integrate neural differential equations into the mechanistic ordinary differential equations of the mechanistic model (Quaghebeur et al., 2021 & 2022).

Literature points out significant potential benefits thanks to increased data collection and (real-time) utilization of data (Eggimann et al., 2017). In due course, hybrid models can function as an intelligent digital twin of a process (or entire system) to assess the current state of the system and make trustworthy predictions about future states (Schneider et al., 2022). This then allows having more confidence in the model predictions and applying it directly in process control, which requires a high level of accuracy.

It is expected that the ongoing trends in digitalization of the wastewater treatment industry will lead to a greater acceptance of data-driven approaches (Corominas et al., 2018) and will replace some knowledge-based models in the near future (Hadjimichael et al., 2016). The number of hybrid models being applied in real-world systems will only increase in the near future (Sin & Al, 2021).

1.3 Context

SIAAP, the wastewater utility in the Greater Paris region (France) responsible for the principal transport and treatment of municipal wastewater of approximately nine million inhabitants, manages six wastewater treatment plants with a total daily dry weather flow rate of 2.5 million m^3/d. For the biological removal of organic matter and nitrogen, it largely employs biofiltration in its facilities. In

order to gain a better understanding of the processes taking place and achieve efficient process control, SIAAP has engaged in research to advance the modelling of biofiltration.

In the Mocopée research program phase I (2014 – 2017) actions were taken to develop mathematical models for predicting water quality variables for the treatment processes applied in wastewater treatment plants in the Paris region (SIAAP, 2018). These efforts have resulted in the development of a physical-chemical primary settling model (Bernier et al., 2016) and models of the secondary treatment phase by biofiltration (Bernier, 2014; Zhu, 2020) and by membrane bioreactor (Nadeem, 2021). The next step consisted in developing an integrated model that combines the three phases of biofiltration, namely upstream denitrification, nitrification and downstream denitrification. The results show that the biofilter model developed by Zhu (2020) is generally able to predict the daily mean values of the output variables.

However, these biofilm models require high computing power, which makes them less suitable for real-time process control applications with a short time action horizon (Schneider et al., 2022).

1.4 Problem Statement and Objectives

Based on the description of the challenges and objectives of the wastewater sector, as well as a review of the state-of-the-art of process modelling this study aims to overcome the obstacles faced by mechanistic biofilter models to be applied as support tools in real-time process control. The study aims to achieve this by combining the advantages of the two modelling approaches through linking a mechanistic model, which makes calculated predictions of effluent water quality, greenhouse gas emissions and operating costs based on the knowledge of the processes, with a data-driven model that can rapidly discover unknown patterns in the data and indicate the mechanistic model's error. The development of this hybrid model is expected to improve understanding and control of the processes involved. The main advantage is an anticipated higher benefit/cost ratio for solving complex problems, which is a key factor for process systems engineering (Von Stosch et al., 2014). The developed hybrid model can form an integrated element of future digital twins (Schneider et al., 2021).

Research Objectives

The overall objective of this study is to demonstrate that hybrid modelling can have an advantage in model accuracy when in use of complex process control of wastewater treatment applying high-resolution operational data. This is achieved by the following principal objectives:

Objective 1: A mechanistic model of a biofilm reactor for predicting nitrogen removal, power consumption, filtration and greenhouse gas (GHG) emission.

Objective 2: A machine learning (ML) model to predict the residual error for the main effluent quality variables providing a better accuracy of the hybrid model.

Objective 3: A hybrid model predictive controller (HMPC) that demonstrates the added value of hybrid models.

1.5 Originality

As will be demonstrated in the literature review (chapter 2), most of the biofilm reactor models simulating wastewater treatment by biofiltration have been calibrated with experimental data from a pilot or laboratory setup. In this study however, it was decided to use high-resolution operational data

from a very large size municipal wastewater treatment plant (in order to approach the behaviour of the real physical system as much as possible).

A key novel aspect of this study is the development of a hybrid model that combines a mechanistic biofilm reactor model and machine learning models and is trained on longer periods (3.5 months) of operational data from a very large wastewater treatment plant. The combined model results in improved model accuracy and allows strong validation.

This study culminates in a demonstration study in which the conventional feed-forward controller in use today at the plant under study is compared to a Model Predictive Controller of the methanol dosing to a biofilm reactor that takes advantage of the developed hybrid model. This, in the context of the hybrid model used here, has not been reported in literature before.

1.6 Structure of Thesis

This thesis is structured in accordance with the objectives described earlier: the three main chapters discuss the results of the calibration of the biofilm model (chapter 4), the outcome of the machine learning model training and testing in a hybrid modelling configuration (chapter 5) and the results of the case study with a hybrid model predictive controller (chapter 6).

The general structure of this manuscript is as follows:

1. A general introduction in chapter 1 to further discuss the context of this study, the problem statement and its objectives;
2. Chapter 2 presents a literature review on the challenges the wastewater treatment sector is facing today, and advancements accomplished that have resulted in modern tools for process control, with a state-of-the-art review of wastewater models applied to biofiltration;
3. Chapter 3 discusses the applied methods for data collection and reconciliation, for model development, including calibration and validation, for training and testing of DDM's and all the materials used;
4. The results of the successful calibration and validation of the mechanistic biofilter model are presented in chapter 4, and how it was applied to new operational data from a large size wastewater treatment plant;
5. The training and testing of machine learning models that can predict the mechanistic model's residual error is discussed in chapter 5;
6. Chapter 6 relates to the application of the developed hybrid model as part of a model predictive process controller of methanol dosing in a submerged biofilter denitrification process;
7. An overview of the general conclusions and perspectives for future studies is then presented in the last chapter 7.
8. An inclusive bibliography is listed at the end of the thesis. A more general description of wastewater treatment, as well as the results of the data cleaning method and sensitivity analysis that have been applied are provided in the annexes.

As chapters 4, 5 and 6 were written as independent research papers, some sections in these chapters might include recurring elements and show some similarities (e.g. methods and materials), although efforts were made to reduce textual duplication.

1.7 Publications and Presentations resulting from this Work

The specific chapters presenting the results of this research have been submitted to peer-reviewed scientific journals and are awaiting approval for publication. Chapter 4 has been submitted to *Techniques Sciences Méthodes* TSM (ASTEE) and is awaiting approval for publication. Chapter 5 is under review by the editors of 'Water Research' (IWA) whereas chapter 6 has been submitted to the journal of 'Environmental Science: Water Research & Technology' (Royal Society of Chemistry).

The development of the hybrid model in this study has so far been presented at several scientific conferences, in particular the 13th IWA Conference on Instrumentation, Control and Automation ICA (17-21 October 2022) (Serrao et al., 2022a), the 8th IWA Water Resource Recovery Modelling Seminar (WRRmod) (18-22 January 2023) (Serrao et al., 2023a), and the Water Environment Federation Innovations in Process Engineering Conference (June 6-9, 2023) (Serrao et al., 2023b), all co-hosted by the International Water Association (IWA). The study project was presented at the 25^{ièmes} Journées Information Eaux (11-13 October 2022) (Serrao et al., 2022b) organized by APTEN (*Association de Professionnels du Traitement des Eaux et des Nuisances*), an association of the French wastewater treatment sector. An accepted proposal for a presentation to the 11th IWA Symposium on Modelling and Integrated Assessment (WaterMatex) in September 2023 will provide a great forum to exchange gained knowledge on the application of hybrid models for control of complex systems.

The contribution of this research to successful international collaboration between academic and industrial partners in Europe and North America was presented to the Megacities Alliance for Water and Climate (MAWAC) - Euro-North American Region (ENAR) workshop, hosted by NYU under the auspices of UNESCO, March 20 and 21, 2023 (Serrao & White, 2023c).

More information about this research project, including video presentations by the author, can be found on the website of the **InnEAUvation** research program hosted by SIAAP (www.inneauvation.fr/MarcelloSerrao).

2 Literature Review

Municipal wastewater treatment is an essential process for the management and control of liquid waste, nutrients and pollutants with the objective to avoid environmental pollution and protect public health. Without it, the quality of natural environments, such as rivers, lakes and soils, would deteriorate rapidly and severely affect biological life and the health of the human population, especially in dense urbanized regions. The understanding of the threats of wastewater and the process technologies to treat the various components of wastewater have evolved over time. Nowadays, the importance of efficient and effective treatment has become a challenge in the face of the effects of climate change, water and energy scarcity, urbanization, the rise of a circular economy where waste is not more than a resource in need of reutilization and the presence of pollutants of emerging concern. The wastewater industry is transforming into energy-efficient resource recovery facilities where automation and digitalization form key elements of process control to achieve required levels of efficiency and compliance with stringent regulations.

This chapter synthesizes the main findings of the review of papers, articles and textbooks on the challenges urban wastewater treatment is facing today and how it has affected process design. The changing role of the wastewater sector in the face of recovery of resources and the ongoing digitalization is highlighted in the second part. The third and main part of this chapter provides introspection about the available knowledge on modern tools for process management and real-time process control, with a focus on process models as decision-support tools for the treatment by biofiltration.

Annex I presents a more elaborative overview of the historical development of urban wastewater management, an explanation of the different treatment phases and provides background information of biological treatment and the importance of biofilms.

2.1 Challenges the WWT sector is facing today

The wastewater sector faces various pressures and drivers that are changing the conventional role of WWTPs where the priorities were the collection and treatment of wastewater into a role, called a Water Resources Recovery Facility (WRRF), in which it will 'produce' resources to be reutilized in the context of a circular economy. The quest for energy-efficiency underlines processes that remove pollutants and nutrients with less energy. Evolving regulations push innovation even further, with new technologies capable to remove contaminants of emerging concern (e.g.: micro plastics, pathogens, pharmaceuticals). Global pressures sustained by climate change, urbanisation and rising living standards are giving rise to floods, water scarcity and freshwater pollution, and these are rapidly becoming major challenges for the wastewater industry.

Role of Regulator

The role of the regulator in environmental protection is an important channel for process efficiency and innovation. When considering the European context in particular, the evolving stricter norms for treated wastewater and discharge into the natural environment as described by the European Urban Waste Water Treatment Directive of 1991 (UWWTD 91/271/EEC) and the European Water Framework Directive (EU Directive 2000/60/EC) had an enormous impact on water treatment across Europe.

According to a recent study (Pistocchi et al., 2019) the objectives of the UWWTD are mostly achieved when looking at the EU scale for reducing raw municipal wastewater discharge and implementing at least primary wastewater treatment. Across the EU member states, there is a high level of compliance with 98% of wastewater collected and 92% satisfactorily treated (EU, 2022). The UWWTD proposes a timetable for all member states to meet minimum standards of effluent quality depending upon the population equivalent and the type of water into which the treated water is discharged. Organic matter, nitrogen and phosphorus removal is required for populated areas with more than 10,000 people equivalents (p.e.). In October 2022, the European Commission announced a stricter revision of the UWWTD to expand wastewater treatment to (small) conglomerations that currently fall outside the scope of the directive, as well as to include the treatment of micropollutants and the monitoring of health parameters (such as the presences of pathogens). The revised UWWTD includes storm water and rainwater management and requires the development of integrated urban wastewater management plans for all large cities. The new plan will set specific energy neutrality targets and the use of renewable forms of energy by 2050 for wastewater utilities (EU, 2022).

The 2030 Agenda for Sustainable Development of the United Nations established a specific Sustainable Development Goal (SDG) to ensure availability and sustainable management of water and sanitation for all. SDG 6 ('*Ensure availability and sustainable management of water and sanitation for all*') aims at clean water and sanitation and has targets for providing access to adequate and equitable sanitation (6.2), for improving water quality (6.3) and increasing water-use efficiency (6.4), ensuring sustainable withdrawal and supply of freshwater to address water scarcity (6.6). SDG 11 ('*Make cities and human settlements inclusive, safe, resilient, and sustainable'*) and SDG 12 ('E*nsure sustainable consumption and production patterns'*) have specific targets on water quality and availability (Ortigara et al., 2018).

The regulation of energy consumption and the promotion of the use of energy from renewable sources on a European scale are set forth in the EU Climate and Energy Directive (2009/28/E), that provides general guidelines for preservation and recovery of energy in the framework of a circular economy. In December 2015, the European Commission further adopted a Circular Economy Plan (EU Commission, 2015), that is an action plan for establishing a circular economy that aims at the development of a sustainable, low carbon, resource efficient and competitive economy (Niero and Hauschild, 2017). Although this roadmap does not in particular focus on wastewater management, the concepts of reducing, recycling, and recovering of waste have been instrumental to promote initiatives in the wastewater sector (Guerra-Rodríguez et al., 2020)

Water Quality

Traditionally, the objective of collective wastewater treatment was the protection of public health in urbanized regions. The treatment focused on the removal of suspended and floatable materials and the treatment of biodegradable organic matter from the wastewater before discharging it into surface water, such as a river. Then, as the protection of the natural environment became important, water quality standards and regulations in Europe and the US were expanded to prevent nutrients reaching the surface waters (Metcalf & Eddy, 2014). Treatment processes were expanded and upgraded to allow for the removal of nitrogen (N) and phosphorous (P) components. Effluent quality norms for more-sensitive ecological regions, such as US ecoregion-based ambient water quality standards for nutrients (USEPA 2015), give a good indication of future effluent requirements for wastewater treatment

operators nationwide. The USEPA (2015) norms for ecoregions are often between 0.01 and 0.1 mg/L for total phosphorus (TP), and 1 and 3 mg/L for total nitrogen (TN) (Daigger, 2017).

Recent studies also note the importance of micropollutants, antibiotics and other excreted pharmaceuticals that are increasingly being found in wastewater. These chemicals are among those forming a mix of chemicals, which cannot be easily removed by conventional treatment plants, as they require costly, more stringent and energy-intensive treatment techniques. Unregulated micropollutants at concentrations in the range of ng/L to µg/L can cause harm to the public health and aquatic ecosystems. Long-term exposure to micropollutants can result in high-risk health problems such as endocrine disruption and cancer (Ngo et al., 2020). Ajala et al. (2022) studied micropollutants of emerging concern in hospital wastewater and its remediation options and conclude that no single treatment technology can effectively detoxify the wastewater. Instead, a combination of methods such as photocatalytic adsorption or photofenton adsorption was found more effective (Ajala et al., 2022).

The removal of pathogenic organisms that may cause long-term health issues and environmental impact, are also receiving more attention nowadays, in particular related to the rise of reuse of treated wastewater. Disinfection by chlorination is one of the most common used methods to inactivate pathogenic microorganisms found in wastewater (Gonzalez-Rivas et al., 2019). However, recent studies reflect on the concern over undesirable chlorinated by-products, as well as the inability of chlorination in eliminating some epidemic microorganisms at low chlorine doses (Xu et al., 2002). Therefore, in recent years, electrochemical processes that allow disinfection have become more applied due to their very high efficiency and their easy applicability (e.g., Gonzalez-Rivas et al., 2019; Shirkoohi et al., 2022)).

Notwithstanding that in the early 1970s the presence of millimetre-sized plastic debris in the marine environment was already documented (e.g., Carpenter et al., 1972), microplastics, which are defined as plastic particles smaller than 5 mm (e.g., Talvitie et al., 2015), and microfibers in wastewater and runoff stormwater (e.g., Treilles et al., 2022) have only recently received more attention. Dris et al. (2015) analysed samples from the influent and effluent of wastewater treatment plants and found microplastics in high concentrations in the raw wastewater. However, the treated water had much lower amounts (83 to 95%) of microplastic contamination, while the removed sludge contained higher amounts of particles and fibers, in particular those of larger and longer sizes. This indicates that wastewater treatment is capable of removing large amounts of microplastics from the water by particle separation as it then accumulates into the sludge (Dris et al., 2015).

Resource Recovery

More recently, the emphasis has moved towards the recovery of resources and energy neutrality as part of a circular economy with wastewater treatment plants (WWTP) transitioning to water resources recovery facilities (WRRF) (Regmi et al., 2019). By establishing more circular resource flows, the water sector can contribute to sustainable development goals. As large-scale centralised WWTPs also represent centralised collection points for a variety of resources – namely water, energy, nutrients and other products – their redesign from treatment facilities into water resource facilities creates possibilities to contribute to a more circular economy. The recovery and reuse of resources can lead to a decrease in the consumption of raw materials, waste reduction, and improvement of energy efficiency. Potential opportunities are the reclamation and reuse of wastewater to increase water

resources, the recovery of nutrients or highly added-value products (e.g., metals and biomolecules), the valorisation of sewage sludge and the recovery of energy (Guerra-Rodríguez et al., 2020).

Reclamation and reuse of wastewater are considered tasks with high potential to implement resource recovery, although not yet exhaustively exploited on a global level (Fito and Van Hulle, 2021). Studies show that in the European Union the reuse of treated wastewater is one of the most commonly applied methods as 18 out of 27 countries reuse reclaimed wastewater at some level for agricultural, industrial, urban, environmental or recreational purposes (Levine & Asano, 2004).

Although the reuse of reclaimed water in agriculture for crop irrigation is a well-established practice, it lacks quality standards and uniformity of good practices across the globe. Furthermore, if not properly treated and monitored, the reuse of reclaimed water might have a negative impact on the quality of the soil (Fito and Van Hulle, 2021).

Energy Efficiency

Studies indicate that municipal wastewater treatment is very energy intensive and that it accounts for 1 – 5% of the national electricity consumption in European countries and in North America (Longo et al., 2016, Mamais et al., 2015). In more detail, Longo et al. (2016) report that data from Germany (Reinders et al., 2012) and Italy (Foladori et al., 2015) indicate that electricity demand for wastewater treatment accounts for about 1% of total consumption. For Spain, studies indicate that 2 – 3% of total electric energy consumption is used by wastewater management, while that percentage roughly doubles to 4–5% when drinking water and agricultural demand is included (Mateos and Rodríguez, 2012). In the United States, about 4% of the electricity demand is dedicated to the treatment and distribution of drinking water and the collection and treatment of wastewater (Goldstein and Smith, 2002).

In a conventional WWTP, between 25 to 40% of operating costs relates to energy consumption (e.g. Metcalf and Eddy, 2014; Elías-Maxil, 2014; Venkatesh & Brattebø, 2011). This value varies in the range of approximately 0.3–2.1 kW h/m^3 treated wastewater (Liu et al., 2004). Panepinto et al. (2016) studied the total electric energy demand of a large WWTP in Italy and found that total yearly consumption was about 66.78 GW h/y, 50% of which is due to the aeration during the biological oxidation processes. In Greece, a recent study determined similar results for aeration accounting for 40–75% of total energy requirements of WWTPs operations (Mamais et al., 2015).

The wastewater sector has a growing interest in the use of tools and methodologies to obtain adequate knowledge of the energy consumption profile, such as benchmarking and energy audits (Belloir et al., 2015). However, regarding energy audits of treatment plants, literature indicates that a uniform approach and a suitable methodology to perform audits that allows identifying the principal energy consumption indicators is still lacking (Foladori et al., 2015). Longo et al. (2016) also found that no single key performance indicator used to characterize energy performance would be applicable universally. This makes comparison among different plants and benchmarking difficult.

Longo et al. (2016) identified three main benchmarking approaches: (i) normalization, where the performance of a key performance indicator is normalized with respect to the reference indicator, (ii) statistical techniques that calculate the performance of a plant by analysing a subset, and (iii) programming techniques that use an optimization to define an optimal contour. The normalization

method provides easily understandable results, but it relies on a large sample of plants to provide a sound benchmark. Several partial indicators may also be needed to compare plants that have different treatment processes (Longo et al., 2016).

Azimi and Rocher (2017) describe the main findings of a longstanding energy mapping exercise and an energy audit performed for six wastewater treatment plants operated by SIAAP in the Paris region (with a combined 9 million P.E.). The annual energy consumption varied between 1 000 GWh and 1 150 GWh for all WTTPS combined. Since 2017, a slight decrease has been observed in the energy used per treated m^3 wastewater, from 1.36 KWh/m^3 (2017) to 1.16 KWh/m^3 (2019). Approximately half of the energy that is consumed was self-generated mainly in the form of methane biogas (46%), which is recovered during anaerobic sludge treatment, as well as green electricity produced as solar power (4%), the remainder is covered by purchased energy (Azimi and Rocher, 2017).

The energy audit revealed that the operation of reactive and compact technologies for water and sludge treatment (such as Biofiltration and Membrane Bioreactors) consumes significantly more energy when compared to the more conventional systems such as CAS. Wastewater treatment processes using physico-chemical lamellar settling, biofiltration, membrane bioreactors (MBRs), as well as thermal dryers for sludge treatment, all require frequent pumping for recirculation, the injection of oxygen, heated air or chemical reagents. Up-flow biofiltration systems and membrane bioreactors need a continuous high-water pressure supplied by pumps to force the flow of water through the supported media/bed and membranes (Azimi and Rocher, 2017).

Global Pressures

Urban wastewater treatment now also needs to address global challenges imposed by the effects of climate change, rapid population growth, freshwater pollution and depletion of water resources. Extreme weather events linked to climate change are among the factors that aggravate the situation by bringing heavier and more frequent rainfall in some areas, but water scarcity in others, affecting how well storm sewers and treatment plants operate. Excessive rainstorms can overwhelm and overload sewer systems and lead to combined sewer overflow (CSO) events, while a lack of rainfall can lead to problems in collection and treatment of wastewater. The rising sea level is threatening the low-lying zones of urban coastal centres and the operations of wastewater plants located near the coast (EEA, 2019).

Additionally, the challenges facing collective sanitation and centralized wastewater treatment in many parts of the globe, in particular in Asia and South America, are those related to a continuing rapidly increasing population, the ongoing industrialization and accelerating urbanization. It is not only the increasing number of persons that reside in an urban conglomeration, but also the increase of living standards that comes with the rise of a larger middle-class society in many lesser-developed regions of the world. This is leading to a change in consumption patterns that generates a higher wastewater per capita generation and brings with it an associated added stress on water resources (EEA, 2019).

From a climate change perspective, studies indicate that wastewater treatment worldwide is a significant contributor with approximately 2 – 3% of total global greenhouse gas (GHG) emissions (Maktabifard et al. 2019). In their study in Greece, Mamais et al. (2015) established that the annual GHG emissions varied significantly according to the treatment schemes employed and that the GHG emissions through wastewater treatment ranged between 61 and 161 kgCO2e/PE. The highest values

of CO2 emissions were obtained in extended aeration systems and the lowest in conventional activated sludge systems. The authors show that a lowering of the dissolved oxygen (DO) set points and sludge retention time can provide significant decrease GHG emissions (Mamais et al., 2015).

These challenges add to those routinely addressed by local authorities and water utilities, including financing construction, maintenance, operation and upgrades, as well as sourcing and retaining suitably skilled staff to deliver the necessary improvements in wastewater treatment. Moreover, on a more structural level, the wastewater industry is faced with an aging infrastructure dating from a first large development period in Europe and North America from 1960's – 1970's when many large centralized WWTPs were constructed. An earlier study in the US projected that the annual cost over the period for operations and maintenance (O&M), will average between $20.3 billion and $25.2 billion US dollars for wastewater systems (CBO 2002). Infrastructure asset management, with renewal in particular that aims to repair, rehabilitate or replace infrastructure to reinstate functionality, is currently an important subject of debate for wastewater utilities. This is not only because assets are aging, but also due to the need for reduction of combined sewer overflows, unsatisfactory hydraulic performance and expansion of the sewage network (Selvakumar and Tafuri, 2012).

2.2 Advancements in WRRF Design & Scope

The wastewater sector has always been evolving under the pressure of guidelines and regulations that have become more inclusive applying stricter norms. The industry has been redesigning treatment technologies and processes to allow the removal of more and different pollutants and nutrients to lower acceptable quality levels. Sustainability has become an important motto, with a focus on the overall energy balance of a facility, its greenhouse gas emissions, the total chemical usage and the carbon footprint of the 'production' of clean and safe water. All these elements have become of importance in the design and operation of modern sustainable wastewater treatment facilities.

2.2.1 Innovative Technologies

Wastewater treatment can be achieved by combining a variety of physical (e.g. screening, settling, filtration), chemical (e.g. coagulation, oxidation), thermal (e.g. drying, incineration) and biological (e.g. by suspended or attached biomass) processes. Globally, the most widely applied wastewater treatment technology is the conventional activated sludge (CAS) process, in which aerobic microorganisms metabolise the organic pollution present in the wastewater under constant oxygen supply. The microorganisms with their adsorbed organic material ('sludge'), are settling out at the bottom of a clarifier ('settling tank'), where they are collected and recirculated back into the aeration tank. This is an important process requirement to maintain the correct balance between the microorganisms and the wastewater (Recycle Activated Sludge - RAS). The Surplus Activated Sludge (SAS) has to be regularly removed (Henze et al., 2008). This process is well understood and can be fairly well controlled under designed conditions. However, it is considered unsustainable due to its low resource recovery potential and low-cost effectiveness, as well as its high-energy demand and large environmental footprint, such as by greenhouse gasses emissions (Kehrein et al., 2020).

The ability of a system to adapt to stress and changes by being transformed into another more desirable state is called the system's adaptive capacity (Folke, 2006) and is considered an important characteristic of a sustainable system (Holling and Gunderson, 2002). Within the domain of urban wastewater, a sustainable urban water system is one that can manage and adapt to changing circumstances (Daigger, 2012; Hering et al., 2013; Vairavamoorthy, 2009; Wilderer and Schreff, 2000).

This requires adaptive management (Pahl-Wostl et al., 2007) and technological systems that are adaptive and flexible (Daigger, 2017; Jeffrey et al., 1997; Spiller et al., 2015). In particular for wastewater treatment, systems need to achieve a high level of energy-efficiency to reduce the dependence on purchased energy and its associated ecological footprint. Also, systems should be flexible so that they can be adapted easily to changing objectives without incurring high capital costs. Another important aspect is the impact of the digital revolution and the compatibility of data and communication systems.

The development of new technologies has come with a shift in urban water management that needs to meet the challenges the wastewater sector is facing and allows to rethink the approach that was designed in the 19th and 20th centuries (Daigger et al., 2019). Since the innovation of anaerobic ammonium oxidation (anammox), its associated technologies are rapidly being implemented (O'Shaughnessy 2015). Membrane technology (Judd 2011), along with other advanced water treatment technologies – e.g., ultraviolet (UV) treatment, advanced oxidation, and biological activated carbon (BAC) – is facilitating increased water reclamation (Zodrow et al. 2017). Phosphorus recovery from used water is now a demonstrated technology, and work is on-going on bio- and electro- chemical technologies for further resource recovery. These trends are expected to continue, supplemented by further advances in biotechnology and materials science (nanotechnology) (Daigger, 2019).

Increased interest in the recovery of phosphorus also affects process choices. Chemically enhanced primary settlers represent one approach to increase the fraction of influent biodegradable organic matter diverted to anaerobic treatment. Here, substantial phosphorus removal also occurs, but methods for subsequent phosphorus recovery are limited because the removed phosphorus is chemically bound. High-rate biological treatment systems (either suspended growth or biofilm-based) and non-nitrifying biological phosphorus removal processes are other options for capturing carbon and diverting it to anaerobic treatment. Biological phosphorus removal processes require lower biological loadings with longer solids residence time (SRT) to allow the relatively slow-growing phosphorus accumulating organisms to be active in the system (Grady et al. 2011).

2.2.2 Biological Activated Filtration

Biological Activated Filtration (BAF), or biofiltration, is a technology that has recaptured a significant amount of attention recently. Although based upon a concept dating from the 1970's, it has been redesigned and improved and has from the beginning of the 21st century become a more applied treatment technology for process intensification for large cities and metropoles situated in Western-Europe and North America (Von Sperling, 2007).

Biofiltration is an intensive treatment technology using microorganisms in compact reactors. This technology combines both physical and biological purification processes in a reactor using immersed filter media onto which the bacteria sustain themselves and break down organic materials and consume nutrients. It employs a dual function of biological oxidation of BOD and/or ammonia soluble components after diffusion in the biofilm, and physical removal of particulate and colloidal solids by adsorption and filtration (Metcalf & Eddy, 2014).

Biological treatment of municipal wastewater is a method to remove pollutants and nutrients by creating conditions inside reactors that stimulate the growth of certain types of bacteria naturally present in the wastewater. It aims at degrading or adsorbing dissolved, colloidal, particulate and settleable matter into biological flocs or biofilms. Biological treatment is based on the natural role of

bacteria to close the elemental cycles of carbon, nitrogen and phosphorus (Henze et al., 2008). It follows a complex process that involves microbiological reactions to remove pollutants and nutrients from the water. For their growth, each type of bacteria feeds on a specific component, such as organic matter or nitrogen ('the substrate').

As mentioned, the technology most commonly applied globally for biological treatment is the conventional activated sludge process. Here, the bacteria are in suspension and feed themselves upon the substrates they encounter inside the water. When they become more numerous and develop into flocs, the biomass density increases and they can settle at the bottom of a sedimentation tank, where the biosolids are removed as sludge. In order to prevent the biological activity from decaying and keep the biomass 'activated', collected sludge is recirculated to the input of the treatment process. The addition of oxygen through aeration stimulates the growth. This process is well understood and can be fairly well controlled. However, conventional activated sludge processes require large bioreactors and settling tanks and are overall considered a slow process with Hydraulic Retention Times (HRT) of 10 to 24 hours (Henze et al., 2008).

Another approach is treatment by microbial growth that is occurring onto support media (such as polystyrene beads). Here, the activity occurs primarily inside a biofilm layer that develops around the support media. The biofilm is a thin layer composed of water, bacteria and Extracellular Polymeric Substances (EPS) that allow for the development of a large active biomass that is diverse (e.g. heterotrophic bacteria near the surface and autotrophic bacteria at the bottom of the biofilm) due to the ecological niches developed inside the biofilm. The main transport of nutrients from the wastewater ('bulk liquid') into the biofilm is through diffusion of soluble components. Due to the small diameter of the support media (3 – 5 mm), when filled-up in a tank or reactor, a large total surface area is created for the bacteria to develop upon; thereby making it an intensive process with an average HRT of 15 minutes in relatively small compact reactors (e.g. 10m x 17m x 4m).

Biological activated filtration is a wastewater treatment method applying fixed growth of biomass in compact reactors that contain a filter bed, upon which the bacteria are active inside the biofilm. Different configurations can be applied to promote environmental conditions for the microorganisms to convert nitrogen compounds, e.g. nitrification under aerobic conditions by autotrophic bacteria and denitrification under anoxic conditions by heterotrophs. Important process parameters are the recirculation rate of nitrified water to feed the upstream denitrification unit, the carbon content of the wastewater and the dissolved oxygen concentration in the case of nitrification.

Biofiltration is a complex and intensive municipal wastewater treatment process that requires a high degree of process control in near real-time to guarantee effluent water that meets the surface water quality objectives and to be able to recover resources with energy efficient technologies that limit the environmental impact.

The main advantages of BAF's when compared to the more commonly applied suspended growth processes (for example CAS) are the relatively small footprint of the installation that allows for it to be totally covered inside a structure (less malodours and visual pollution), the ability to treat diluted wastewaters and the absence of a requirement to settle the biomass. However, the process is more complex from an operational and maintenance point of view, has a high capital cost for installation and operation, and it is vulnerable to clogging with solids and accumulated biomass, requiring frequent backwashing to maintain high performance (Metcalf & Eddy, 2014). For these reasons, it is often

applied in a secondary treatment phase. Moreover, the energy requirements of BAF systems for aeration, water distribution and recirculation flows are of major concern and can be reduced by optimising internal recycling and adapted control strategies (for example: Rother & Cornel, 2004).

2.2.3 Energy Efficiency

The wastewater industry is continuously seeking new technologies to achieve a more efficient usage of energy in wastewater treatment and even energy self-sufficiency so that it can reduce the dependency on purchased energy and operate in a sustainable manner. Of major concern are the cost issues on a microeconomic level, the increasing risk of an energy-shortage due to restrictions on the production and import of energy, but also the associated carbon footprint of electricity consumption (Wett et al., 2007).

The largest energy consumer in a WRRF is the process of aerobic biological degradation. Methods to lower the required aeration energy are to decrease the amount of organic pollutants to be aerobically degraded by redirecting them to the anaerobic digester and/or to use shortcut nitrogen removal pathways instead of the full nitrification-denitrification process, with the latter using a significantly larger amount of oxygen and carbon source for nitrogen removal (Daigger 2014). Energy self-sufficiency can be achieved through production of biogas while simultaneously minimizing the energy consumption of treatment processes.

Biogas is produced in the anaerobic digestion process and can be transformed into thermal and electrical energy using a combined heat and power plant (CHP). Optimization of biogas production is an important measure for overall energy efficiency. As an example, Müller et al. (2006) showed that in Switzerland two thirds of all WWTPs that had undergone an energy analysis were able to reduce energy consumption by on average 38%, of which 67% was due to increased electricity production from biogas. The remainder was due to real energy savings accomplished in the biological stage and with improved energy management (Wett et al., 2007).

Nowak et al. (2011) reported that in Austria, two municipal wastewater treatment plants for nutrient removal and aerobic sludge digestion have become energy self-sufficient after a long optimisation process. The optimisation was mainly concerned with optimal aeration control and control of the aerobic section of the aeration tank to optimise denitrification and prevent degradation of particulate organic matter that should be degraded in the digester. Both plants were equipped with energy-efficient combined heat and power (CHP) units. The authors suggest however, that it is more sustainable to use the biogas as bio-methane/bio-fuel (Nowak et al., 2011).

As pointed out by Daigger (2019), energy neutral operations were more standard in the past when treatment was focused only on the removal of organic matter monitored by the BOD_5 and total suspended solids (TSS). Primary treatment was then often followed by either high-rate activated sludge or trickling filter biological treatment systems to minimize liquid stream energy requirements. Treatment was coupled with anaerobic digestion of the primary and secondary sludges, and the reuse of the biogas produced in combined heat and power (CHP) systems. This ensured that plants were able to produce sufficient power to meet their own plant-wide electricity needs (Daigger, 2019).

However, the demand to provide consistently higher quality effluent with lower BOD_5 and TSS concentrations, and the evolving regulations that impose removal of nitrogen and phosphorus from wastewater, forced many WWTPs to implement a lower rate activated sludge system and the resultant

increased aerobic stabilization of sludge. The net result was a reduced secondary sludge production, with less biogas for use in CHP systems, and an increase in energy demands for biological treatment, especially for nitrification. This made the overall process net energy negative (Daigger, 2019).

Consequently, the wastewater industry is searching for energy-efficient technologies. Anaerobic treatment technologies have a great potential to lower energy consumption (Tarallo et al. 2015). Several other studies propose a chemically enhanced primary treatment to maximize energy recovery and achieve energy-neutrality (e.g.: Jenicek et al., 2013; Dolejš et al., 2019), which can increase overall organic matter removal from 40% to up to 60% (Wan et al., 2016). Jenicek et al., (2013) describe how biogas production was optimized at a large wastewater plant by enhanced primary sludge separation, a thickened waste activated sludge, an implemented lysate centrifuge, increased operational temperature, and an improved digester mixing. With these optimizations, biogas production increased significantly. The authors conclude that, with the correct optimization of anaerobic digestion technology, wastewater treatment can either attain or come close to attaining energy self-sufficiency (Jenicek et al., 2013).

Descoins et al. (2012) search for energy efficiency by optimizing an activated sludge process coupled with anaerobic digestion in a large wastewater treatment plant and show the influence of primary settling efficiency on electrical autonomy. A limiting factor for the overall energy efficiency is the ammonium recycling from digestion to the activated sludge reactors, as well as the lack of carbon substrate availability for denitrification in the activated sludge reactors (Descoins et al., 2012).

Other measures that can improve the efficiency of operations significantly are control strategies for aeration systems and the mixers in activated sludge processes, and the installation of more efficient pumps (Sarpong and Gude, 2020).

Serralta et al (2002) use an ammonium-based supervisory control system that - through simulations – optimizes efficiency and reaches energy savings of up to 10% in blowers and 50% in recirculation pumps. Rieger et al. (2012) achieved up to 20% decrease in total energy requirements (30% in simulations) besides an improved effluent quality upon implementing full-scale dissolved oxygen (DO)-based and ammonia (NH3)-based aeration control strategies. An overview of aeration control systems applied to municipal WRRFs revealed as high as 27% aeration energy requirement reduction for a full-scale activated sludge process (Åmand et al. 2013), while optimization of aeration control strategies combined with electricity production and on-site utilization from AD biogas can reach a 30–50% reduction in energy costs (Wett et al. 2007). However, these values suggest that, with aeration control strategies and conventional CAS-based layout alone, a fully energy self-sufficient WRRF will not be achieved unless external resources are added, such as organic wastes for co-digestion (Shen et al. 2015), or by modifying plant layouts and incorporating other unit processes (Khiewwijit et al. 2015; Fernández-Arévalo et al. 2017).

Another way to reduce the high energy demand of the activated sludge process is to propose organics removal techniques that do not require aeration. As an improvement over chemically enhanced primary settling, Gikas (2017) proposes a process that is based on advanced microsieving and filtration processes with a rotating fabric screen (pore size: 100–300 μm) followed by a continuous backwash upflow media Filter. About 80–90% reduction in TSS and 60–70% reduction in BOD_5 has been achieved. Then the partially treated wastewater is fed to a combination of low height trickling filters, combined with encapsulated denitrification, for the removal of the remaining BOD_5 and nitrogen. The biosolids

produced by the microsieve and the filtration backwash concentrate are fed to an auger press and are dewatered to about 55% solids. The biosolids are then partially thermally dried (to about 80% solids) and conveyed to a gasifier, for the co-production of thermal (which is partly used for biosolids drying) and electrical energy, through syngas combustion in a co-generation engine. In this study, the energy requirements for complete wastewater treatment, per volume of inlet raw wastewater, are reported to be 85% below the electric energy needs of a conventional activated sludge process (Gikas, 2017).

Andrews et al. (2016) show that the emerging energy-efficient technologies that are most likely to be adopted in the near future are those focusing on mainstream shortcut nitrogen and pyrolysis/gasification, since these technologies are closest to being ready and the deployment timeline is estimated to be relatively short. There is a strong potential for net electric energy production through gasification/co-generation (Gikas, 2017). Mainstream shortcut N removal and mainstream anaerobic treatment are expected to have the biggest impact on energy use in the wastewater sector in the near term (Andrews et al., 2016).

Solon et al. (2019) demonstrate that there is potential for achieving future energy-positive water resource recovery facilities through novel integration of mature technologies. They have evaluated different simplified model layouts that include a high-rate activated sludge system, chemically enhanced primary treatment, partial nitritation-anammox, partial nitrification-denitrification over nitrite and anaerobic digestion, and then have compared the energy consumption and operational costs to a conventional activated sludge (CAS)-based process scheme. This study showed that the combination of novel and mature technologies has less aeration energy requirements, lower operational cost and higher biogas production than a CAS (Solon et al. (2019).

In another study, Khiewwijit et al. (2015) proposed a treatment layout that combines bioflocculation, deammonification, phosphorus recovery, AD and combined heat and power unit processes to maximize nutrient and energy recovery. This resulted in a simulated net energy yield, due to reduced aeration energy consumption and by an increased energy. Fernández-Arévalo et al. (2017) included thermal hydrolysis to improve sludge biodegradability and increase biogas production. They also modified the conventional plant layout by decoupling COD, nitrogen and phosphorus treatment to optimize resource recovery.

Overall, there is significant evidence of the importance in the overall design of treatment systems to make them energy neutral or even energy positive. However, currently, only a small percentage of WWTPs globally produce energy in large volumes for valuable (re-)use and few are energy self-sufficient (Sarpong and Gude, 2020).

2.2.4 Barriers to implementation of innovation

The implementation of innovative technologies is often a slow process. Literature research indicates several common barriers of structural and non-structural types.

The useful life of the physical structures, especially hydraulic transport infrastructures such as pipes and channels, and water storage tanks, is significantly longer than that of the specific technologies used. Geels (2006) and Neumann et al. (2015) argue that wastewater systems follow technological trajectories that are a result of decisions and choices made in the past for investment in technology.

Spiller et al. (2015) indicate in a similar manner this dependency on technology and the systemic effects of change in water and wastewater engineering. Engineered systems are determined in function of interconnectedness and coupling of system components. Structural changes in one part of the treatment system rarely occur in isolation, but rather affect other parts of the system (Ulrich, 1995). The importance of interconnectivity in engineered systems results in an interdependency that makes it a barrier for a transition towards sustainable urban water management and a change becomes difficult to implement (Marlow et al., 2013).

Alternative approaches to design have been exploited in order to reduce the dependency of the existing infrastructure, such as phasing of systems over time and installation of modules or blocks of system units (Spiller et al., 2015).

The role of the legislator in prescribing water regulations might also create a hindrance for innovation to flourish (e.g. Maere et al., 2016). There are generally two types of regulations that can be discerned; regulations that are technology based and lead to technology-based effluent limits (TBEL), and regulations that apply an environmental risk management perspective having set water quality-based effluent limits (WQBEL). There is an increasing tendency for developed countries to employ a mix of TBEL and WQBEL (e.g. the UWWTD and WFD in Europe). The UWWTD sets a minimum baseline of treatment and is rather easily understood and implemented. However, it is often inadequately equipped to meet many of today's challenges (Novak et al., 2015).

The shift to performance-based legislation that sets water quality effluent limits (i.e. WFD) allows for further improvements of the ecological quality. Moreover, prescriptive legislation that specifies the application of a certain technology could actually limit innovation. Since they prescribe specific technologies, instead of performances, and carry penalties when these regulations are not met, they leave out other potential design and operational flexibilities (Novak et al., 2015).

Certain legislative conditions can also promote the development of innovative solutions. Maere et la. (2015) discuss the importance of setting ambitious environmental objectives that challenge the current technological and organizational boundaries; of creating a flexible legal framework that allows for some freedom in achieving the environmental goals (i.e. performance-based rather than prescriptive); and of setting an appropriate financial framework that incentivizes innovation (Maere et al., 2016).

2.3 Important Aspects of Wastewater Treatment Plants of the Future

The role of WWTP is changing from solely environmental protection by producing effluent water that is conform the standards, into a facility that produces (raw) materials and energy in the context of a circular economy. The historical centralized approach will probably give way to a more decentralized form where optimized systems allow the local reuse of water (for agricultural purposes) and the recovery of resources. The ongoing digitalization, the use of big-data and development of digital twins will allow for further process optimization. Important aspects of these anticipated changes are now discussed.

2.3.1 Resource Recovery

The wastewater industry is currently facing dramatic changes, shifting away from energy-intensive wastewater treatment towards low-energy, sustainable technologies capable of achieving energy-positive operation and resource recovery. The latter will shift the focus of the wastewater industry to

the extraction of resources from the wastewater, as opposed to the conventional paradigm of treatment. The new and wider vision for a water resource recovery facility (WRRF) includes water sanitation, protection of water sources and the natural environment, energy reduction and production, and resource recovery (Regmi et al., 2019).

WRRFs of the future must be able to meet a variety of functional conditions. An important requirement remains the treatment of wastewater and the production of high-quality effluent water (water recovery) that meets quality standards for receiving water bodies (e.g. a river into which the treated water is discharged) and that meets the norms for reuse standards (e.g. as irrigation water in agriculture). Additionally, a WRRF is focused on the recovery of energy, nutrients and organic material, and may become an integral component of a bio-based economy. WRRFs are increasingly designed to maximize organic matter capture, rather than oxidizing it. This reduces the oxygen requirements for treatment and allows the organic material to be digested anaerobically for biogas production. The requirements are likely to vary over time, in nature as well as magnitude, and the facility must adapt to them (Daigger 2009, 2010, 2012a, 2012b).

2.3.2 Flexibility in Design

Another highly important aspect of the principal design considerations that is mentioned in literature is the flexibility to adjust to changing conditions and the ability to adapt to and incorporate evolving technologies. A modern wastewater treatment plant or recovery facility should not be conceived or designed with a fixed set of functional requirements in mind that are based on a fixed set of technologies. Rather, during the design phase of the layout and configuration of a treatment and recovery facility, future growth and the integration of uncertain requirements using evolving technologies should be accounted for. A 'building block' approach will allow adding additional treatment capacity and process upgrades as it becomes necessary over time (Daigger, 2019).

A treatment may be adequate early in the life of a WWTP – as designed-, but more stringent requirements, added over time, will often require the addition of technology systems to achieve evolving treatment objectives. A strategy enabling significant upgrades in effluent quality over time should be incorporated into the planning and layout of a WRRF. Technology changes can also be made to improve treatment capacity and/or as a result of changing economics. Functional adaptations include hydraulic expansion – e.g., adding associated treatment units – treatment process upgrades (e.g., to meet more stringent water quality requirements), new organic matter ('sludge' or 'biosolids') processing and resource recovery technologies/opportunities, and plant aesthetics – e.g., mitigation of external impacts like pollution caused by noise, odour, truck traffic and external lighting that can have deleterious effects on human health, wildlife, and environmental quality of the immediate surroundings (Daigger, 2019). In addition, social benefits derived from odour abatement link to nuisance reductions in the nearby population and improvements in occupational work environment for the operators within the WWTP (Estrada et al., 2011).

The external appearance of a WWTP and the mitigation of impacts of operations have become increasingly important to the design of new plants due to expansion of city developments in proximity of industrial zones. Also, the public becomes less tolerant of facilities which adversely impact adjacent properties and their market value. A number of practical architectural and moderate cost measures are mentioned, such as external screening using either engineered or natural materials. Alternative, new plants can be constructed (partially) covered in response to the increased attention to aesthetics.

Concerning the plant layout, Daigger (2019) points out the good practice to consider the full range of potential product water requirements over the full lifespan of the plant, and the ultimate hydraulic and organic loading capacity requirements. Space allocations based on existing technologies are often sufficient, as more efficient technologies, which will generally require less space, are likely to be available in the future should more stringent product water quality requirements be imposed. Future capacity requirements can be allowed for by allocating space for additional treatment units, while increasing the capacity of the plant's existing hydraulic structure by installing parallel pipes and/or channels (Daigger, 2019).

In contrast to the historical approach of centralized wastewater treatment, where the economy of scale was a major factor, decentralization is now recognized as an emerging solution and it has become a significant area of focus for many researchers (Li et al., 2023; Zhang et al., 2023). Decentralized WWTPs allow for a more sustainable and integrated urban wastewater system and are considered a complement to conventional centralized wastewater treatment processes. Decentralized treatment has the potential to improve the technical, environmental, and economic performance of existing centralized systems (Garrido-Baserba et al., 2022). Most of the studies focus on separate treatment of different wastewater fractions and promote the application of source separation where different effluents are collected separately (urine, faeces, and/or greywater) in order to increase resource recovery. Joveniaux et al. (2022) analysed source separation as an alternative to conventional wastewater treatment in several large cities in France and demonstrate that it is feasible, but still in an emerging phase. Extreme decentralization to building scale, where decentralized water systems are located in a building or a building block to allow treating wastewater and recovering resources for direct reuse, is a further topic of research. However, the application here is even slower mainly due to regulatory constraints and the risks associated with these systems (Garrido-Baserba et al., 2022).

2.3.3 Big data & Digitalization

Notwithstanding that the water industry is transforming and digitalization forms a key part of this transition (IWA & Xylem, 2019), most digitalization initiatives have focused on drinking water production and distribution applications. This delayed uptake in the wastewater sector can be explained by the complex biological reactions that are part of wastewater treatment processes that lead to specific requirements with data collection and model development (Torfs et al., 2022).

However, recently, interest in data-based technologies is growing in the wastewater sector, as indicated by surveys among wastewater utilities world-wide. Yang et al. (2021) performed an industry sector wide literature survey on relevant keywords ("digital twin") and concluded that there has been an incremental increase in the amount of literature published on digital twins after 2015. As big data techniques, Internet of Things, cloud computing, and artificial intelligence algorithms advance, the digital twin technology has entered a phase of rapid development (Yang et al., 2021).

In the wastewater sector, developments in monitoring technologies have allowed for inline sensors that monitor quality variables to be more precise, reliable and affordable. These sensors generate large quantities of data in high-frequent signals (e.g. every minute). Other remote sensing and communication technologies are also contributing to the proliferation of big data (Garrido-Baserba et al., 2020).

Therrien et al. (2020) explain (see Figure 2-1) that the development and application of smart process control tools have a strong need for process data, which has historically been difficult to collect. With

sensors more ubiquitous than ever, there is now an unprecedented amount of data being produced on a daily basis. However, due to the tough measurement conditions typically present in the wastewater environments, measurements are still subject to many faults that can reduce the trustworthiness of the data. Sensors are disturbed by bias, drift, precision degradation or total failure effects that cause the reliability of measurements to decrease (Yoo et al., 2008).

For sensors and novel data-driven tools to be considered in operation for process control, the data they generate must be available, accurate and verifiable in real or near-real time. This will result in interpretable data of high quality (Alferes and Vanrolleghem, 2014).

As indicated in Figure 2-1, the data processing steps required to create understanding and intelligence from raw measurements can collectively be pictured as a flow of data through a pipeline. When knowledge-based tools are transparent and robust enough to be trusted and are supported by data-based modelling approaches, intelligent decisions can be made with confidence (Therrien et al., 2020). Operational monitoring data is often collected and stored in raw format in a database. Prior to its utilization, it requires being preprocessed in order to remove outliers and erroneous data. This creates gaps in continuous time series data that need to be filled for certain applications. An automatic data quality assessment method recently described by Alferes and Vanrolleghem (2016) extracts information from individual water quality time series from on-line sensors and then, based on prediction models, detects outliers and smoothens the signal by removing faulty and unreliable data.

Aguado and Rosen (2008) apply multivariate statistical projection methods, like principal component analysis (PCA) and partial least squares (PLS) to make data more comprehensible by extracting relevant information. These methods make use of data directly collected from the process to build an empirical model which serves as reference of the desired process behaviour and against which new data can be compared. In addition, they provide graphical tools that are easy to apply and interpret by process operators. Applications of these approaches and modifications of them to wastewater treatment operation include continuous and batch processes (Teppola et al., 1999; Lennox and Rosen, 2002; Lee and Vanrolleghem, 2003; Yoo et al., 2004).

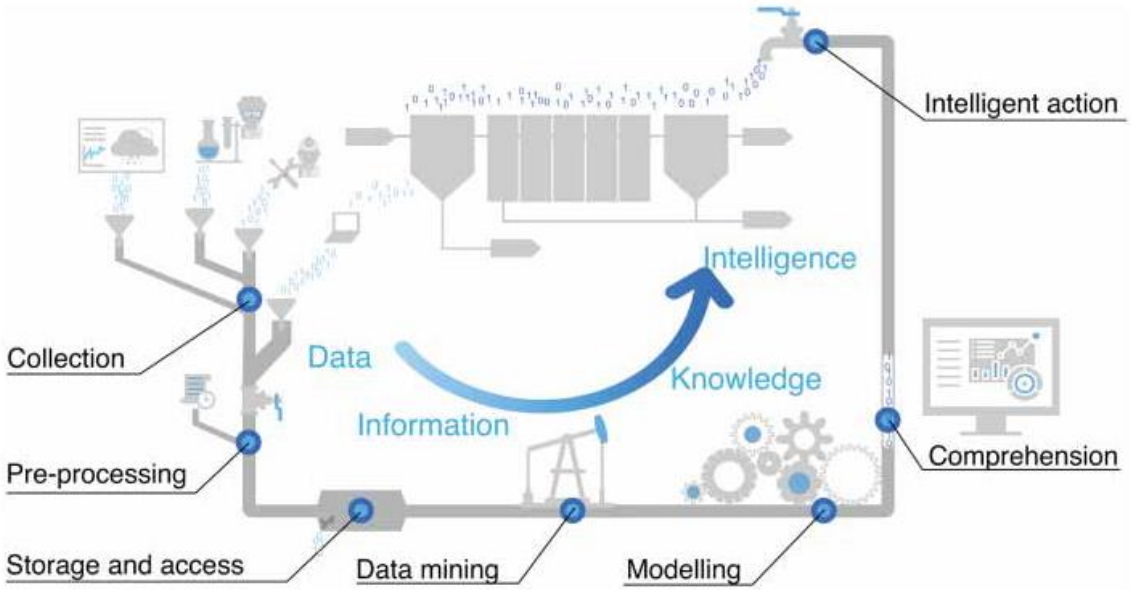


Figure 2-1: From data to intelligence. Schematics indicate for each step in the pipeline the most essential actions (Therrien et al., 2020).

Big Data

Modern WWTPs that are well equipped with inline sensors and other surveillance tools nowadays continuously monitor and collect data from unit processes, but the data are often underutilized. The large amount of data, the complexity of the datasets and the lack of data science background for operators and managers, makes it challenging to efficiently collect, manage, and analyse the data (Diebold, 2003; Kadiyala, 2018; Manyika et al., 2011; Regmi et al., 2018).

Most raw data are stored in their original format for potential future performance analyses with little consideration to their structure or the organization of the data repository (Plana et al., 2019; Newhart et al., 2019). As Therrien et al. (2020) point out, this creates a large pool of data or a “data lake” that allows for the storage of both structured and unstructured data, without concern for indexing or making sure that the metadata fits a specific schema (Nargesian et al. 2018). Also, data lakes provide no guarantee regarding the integrity of the data they hold, which might result in data swamps rather than lakes. Since there is no need to structure the data along with descriptive metadata upon writing, much of the context in which the data was collected may be lost once it is stored in a data lake. This increases the risks for the data to become ignored, misused or corrupted (Therrien et al. (2020).

As more data is being collected the storage and retrieval of data becomes more complicated, and specific large-scale data storage systems are required for the management of big data. Big data is characterized by the properties summarized by the four V's (Farley et al. 2018): (i) Volume of data being generated by all the sensors, actuators, alarms and type of laboratory measurements; (ii) Velocity of data creation and collection by in-line sensors and instruments; (iii) Veracity of the data being collected: the harsh environment WRRF sensors find themselves in mean that faults are omnipresent. Pre-processing is therefore critical for the data to be trustworthy; (iv) Variety of collected data: most of the sensor data collected in WRRFs are time-series data. However, a lot of information at WRRFs is nowadays being stored in other unstructured data formats such as photos, videos, spectral measurements, instrument data sheets and standard operating procedures (Therrien et al., 2020).

Actually, large amounts of data are being collected and stored in databases but remain unused since operators have difficulty in processing and analysing the data in a timely enough fashion to allow for better understanding or proper decision-making (Corominas et al., 2018).

Data usage can have multiple purposes, such as to calibrate and train mechanistic and data-driven models, to supply models with input data for simulation, to evaluate the state of the dynamic subsystems, to feed soft sensors, and/or to provide the necessary inputs for automated closed-loop control strategies (Torfs et al, 2022). The quantity and reliability of data that is being collected at WWTPs is on the rise (Olsson et al. 2014). New, more reliable sensor technologies combined with faster computations are making the storage of frequent and high-resolution water quantity and quality data more affordable. However, due to technological and monetary constraints, data collection and data management will remain somewhat limited, which means data access will ultimately dictate how to optimally exploit the available data (Torfs et al, 2022).

Other organisational activities that support operations (e.g. equipment and infrastructure maintenance, finance, resource planning, compliance, research, and development, etc.) are also generating significant amounts of digitised data. The information gathered from digital workflows provides the context for interpretation of observations made in the process systems. These metadata relate to process equipment, infrastructure, location, maintenance history, quality, purpose, measuring range, and typically do not need to be updated frequently (IWA 2021b). However, metadata still need to be checked and kept up-to-date on an active basis.

Data checks also needs to take into account the compatibility of data that is stored in a wide variety of formats (e.g. PDF files of P&IDs, screenshots of CAD models), which needs to be replaced with data in file formats native to the design software and that can integrate into a database (Brendelberger et al. 2019; Wiedau et al. 2019).

Fundamental to the mathematical modelling and data management is the IT software architecture and hardware infrastructure needed. Sophisticated data architectures and computational workflows, composed of structured and unstructured data import, storage, and processing, will be required (Torfs et al., 2022). Fuller et al. (2020) suggest that the cost of installing and running such infrastructure is likely to be the biggest challenge for any digitalization project. Furthermore, questions need to be addressed around data sharing, cybersecurity, and ethics (GWRC 2021).

Furthermore, the on-going digitalization is raising interest in the application of other data-based solutions. Integrating the power of big data analytics (including AI) with the management and control of urban wastewater systems represents a significant unexploited opportunity to achieve economic and ecological objectives of sustainable operations. The emergence of big data science and the application of machine learning will develop new ways to analyse, organize, and extract information from large data volumes and create opportunities for data-driven discovery (Garrido-Baserba et al., 2020).

2.3.4 Concluding Remarks

Wastewater treatment facilities of the future must not only adapt to evolving functional requirements but also to changing technologies, requiring a high level of flexibility and adaptability in the processes. This applies to hydraulic and treatment capacity, plant layout and design, and operation. A proven

engineering approach is that of the 'building block approach' to plant layout and design, and the choice of process tank configurations which can accommodate numerous technologies (Daigger, 2019).

The changing role of the wastewater industry that is leading it away from energy-intensive wastewater treatment processes towards low-energy, sustainable technologies capable of achieving energy positive operation and resource recovery, will also impact the process control tools and systems in place. Moving away from the conventional paradigm of treatment, the tools of the future need to be capable of monitoring, modelling and controlling processes that recover resources (e.g.: water, nutrients, organics, energy). The practice of design, operation and control of resource recovery technology will need modern tools that are able to perform with high precision in real-time to allow meeting stringent objectives related to water-product quality, process performance stability, and operating costs. As such tools lead to a better understanding of processes; this may also lead to new and innovative resource recovery solutions (Regmi et al., 2019).

2.4 Modern Tools for Process Control

For optimal plant control, advanced control strategies must be implemented (Dürrenmatt and Gujer, 2012). There is a need for numerical support tools to simulate and understand the effects of desired outcomes or proposed modifications in the processes. In fact, a small change in a local process setting can have impacts upon the performance of the other stages of the treatment process. The problem becomes even more complex when we add the economic effects and environmental impact of proposed changes to the performance criteria. Process modelling provides valuable insights into the behaviour of processes; in fact, it is a requirement for process optimization (Wentzel & Ekama, 1997). Modelling is widely used for water resource recovery facility (WRRF) design, upgrade and controller evaluation (Sweeney & Kabouris 2013). Such smart tools will contribute to decision-making and process control management.

In the last few decades, there have been experiences gained and lessons learned with the application of smart infrastructures in the sanitation sector (Ingildsen and Olsson, 2016). This is observed in particular in the fields of in-line monitoring of water flow and of qualitative process performance indicators used by automatic controllers (e.g. Sharma et al, 2011), as well as the development of computer aided mathematical models for simulation and scenario analysis (e.g. Benedetti et al., 2013). This has resulted in more stable and reliable process control, increased regulatory compliance and overall increase in our understanding of the hydraulic, chemical and biological processes taking place during the transport and treatment of municipal wastewater.

The development and application of smart process control tools allows optimization and efficiency, facilitates recovery of resources of high quality, and promotes the shift towards energy neutrality as part of a circular economy. Automation may also contribute to counter a shortage of a skilled workforce.

2.4.1 Process Control

Process control has the objective to optimize the efficiency of process performance in terms of removal rates of pollutants and nutrients, energy efficiency, the amount of reagent consumption. Process control allows operating a plant at optimum conditions in order to enhance the lifetime of a system and reduce unit product cost (Agarwal et al., 2016). Like many other industries, the wastewater treatment sector has been experiencing a sharp increase in the application of instrumentation, control and automation (ICA) since the 2000s (e.g. Ingildsen and Olsson, 2016; Garrido-Baserba et al., 2020).

The ICA technologies were already introduced in the 1960 - 70's in order to control operations and assure good quality of the discharged water in their natural environment.

Prior to the 1960s, control relied on manual adjustments and observations made by operators. Digital control was not yet an option due to the fact that the general knowledge was not yet well developed on the dynamics of treatment or the theory of process control for non-linear systems, like WWTP's (Olsson, 2012). Also, high investments were required for the use of computers and instrumentation, while the existing WWTPs at that time were not designed with sufficient flexibility to facilitate digital control (Newhart et al., 2019). Starting from the 1980s, more affordable computing power became available that led to the development of first principle models, although their complexity and lack of reliability made them poor advisory systems (Olsson et al., 1998). Moreover, the increase in plant complexity demanded a higher level of process control. Various control strategies for wastewater treatment plants were then proposed to better control dissolved oxygen (DO) levels and denitrification (e.g. Briggs, 1967; Busby and Andrews, 1975; Londong, 1992).

By the early 2000s, the digital revolution had reached the wastewater sector, and most WWTPs had integrated their own version of direct digital control into process monitoring in the form of programmable logic controllers (PLC) and supervisory control and data acquisition (SCADA) systems (Newhart et al., 2019).

Commonly, in literature two types of control strategies are differentiated: conventional (or classical) control and advanced control. In conventional strategies, such as PID, ratio cascade and feedback control, the control actions are directly determined by the measured signals of disturbances (e.g. temperature, pressure, flow rate, pH and quality concentrations). The signals are transmitted to the controller which then takes corrective action (e.g. either opens or closes a valve). It is the expectation that this action will return the variable to the desired set point within a short time span. In advanced control however, the controller allows for compensation of time delays and dead time in sensors and sampling apparatus and has override ability. Selection criteria for the type of controller are the stability, response time and the disturbances caused to the system (Olsson, 2008).

Nowadays, the most common process control practice is to maintain a set-point (i.e., target value) using inline sensor readings and feedback control. For example, dissolved oxygen DO concentrations can be controlled by adjusting air blower speed (i.e., aeration intensity) where online measurements in the system send a signal to the SCADA system that determines whether blower speed should change or stay constant relative to a measured value, like a DO concentration. Sufficient nitrification can be maintained by applying a constant aeration flow rate, by control of the DO level at a pre-selected set-point or by using a variable DO set-point controller based on ammonia concentration in the last aerated reactor of the plant (Ingildsen, 2002; Vrečko et al., 2003).

The denitrification process can be controlled by manipulating the external carbon flow rate or internal recirculation flow rate based on the nitrate concentration in the last anoxic reactor (Lindberg, 1997; Yuan et al., 2002). Chemical dosing to enhance contaminant precipitation and additional carbon for biological processes are other examples of feedback control based on inline measured nutrient concentrations. This method of process control is commonly achieved by a controller with a variable frequency drive to change operating conditions in a continuous, smooth, and automated manner (Newhart et al., 2019).

Different control algorithms have been proposed from simple ON/OFF and PI feedback control (Ayesa et al., 2006) over feedforward control to complex model predictive control (MPC) (Steffens and Lant, 1999). A simple feedback controller scheme shown in Figure 2-2, is where the output signal from the process model is transmitted to a comparator to be subtracted from the desired value of the signal (set point). The resultant signal from the comparator is the error signal and goes to the controller. Here, control algorithms or control modes perform mathematical operations on the error signal to calculate the required value of the signal to be transmitted to the final control element. The final control element then modifies one of the input variables to the process in order to bring the process output signal closer to the set point (Andrews, 1976).

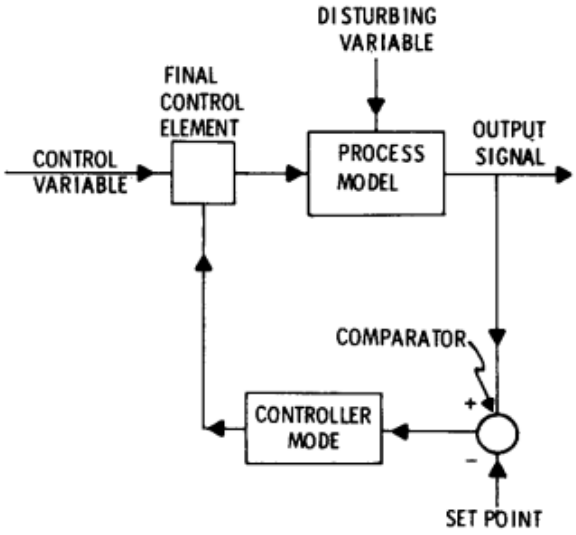


Figure 2-2: Block diagram of a simplified feedback control (Andrews, 1976).

A very commonly used control algorithm is that provided by the three-mode controller equation known as PID control, where the controller calculates the amount of control action as a function of four terms (equation 2.1). Term (1) is the bias and is provided to allow the amount of control action to be set at approximately half of the full value whenever the measured variable equals the set point. Term (2) provides proportional control in which the amount of control action is proportional to the error signal. The third term (3) gives control action proportional to the integral of the error signal and therefore reflects the length of time that the error has persisted. The fourth term (4) is proportional to the derivative of the error signal and thus reflects the rate of change of the error (Andrews, 1976).

$$C_A = K_B + K_P e + K_{IN} \int_0^t e\, dt + K_{DI} \frac{de}{dt} \qquad \textit{Equation 2.1}$$

where: C_A = amount of control action
K_B = bias control coefficient
K_P = proportional control coefficient
K_{IN} = integral control coefficient
K_{DI} = derivative control coefficient

The number of terms and the values selected for their coefficients is dependent upon the dynamic behaviour of the process. Determination of the values for the coefficients is known as controller tuning and is primarily a trial-and-error process although there are optimization programs available (Andrews, 1976).

Although optimal control performance cannot be expected from conventional PID or on-off controllers for the time-varying, nonlinear processes considered, their widespread use in industry and the resulting familiarity with their properties and concepts for design, have made that these regulators are the most widely applied in wastewater treatment processes (Vanrolleghem, 1995).

2.4.2 Model Predictive Control

Process control implies the use of models that represent or describe often in a simplified manner a wastewater treatment processes (Wentzel & Ekana, 1997). Models provide insight into system and plant performance (quantification of information, mass balances and data reconciliation), allow evaluating possible scenarios for upgrading or new designs in a quantitative manner, support management decisions, develop new control schemes and provide knowledge transfer (e.g. operator training) (Henze et al., 2008).

To optimize process performances, a tight control of manipulated variables is required. In complex wastewater treatment systems, conventional PID controllers cannot adequately perform due to non-linear and time-varying conditions. Here model predictive controllers (MPCs) appear to be better suited. In MPC, the process model takes into account the physical constraints of the actual system (Richalet et al., 1978).

As shown in Figure 2-3, in MPC, a process model is used to predict how the system would behave under a proposed sequence of control actions u(t). Using a numerical search algorithm an optimal sequence uOpt(t) is computed. To reduce the effects of model mismatch, only the first action of the sequence is implemented on the process after which the optimization exercise is reiterated (Vanrolleghem, 1995). This concept of a moving horizon computes a set of control moves over a larger prediction horizon, while only the initial control move is actually implemented and the remaining calculated moves are discarded. The process is then repeated at subsequent sampling times (O'Brien et al., 2011).

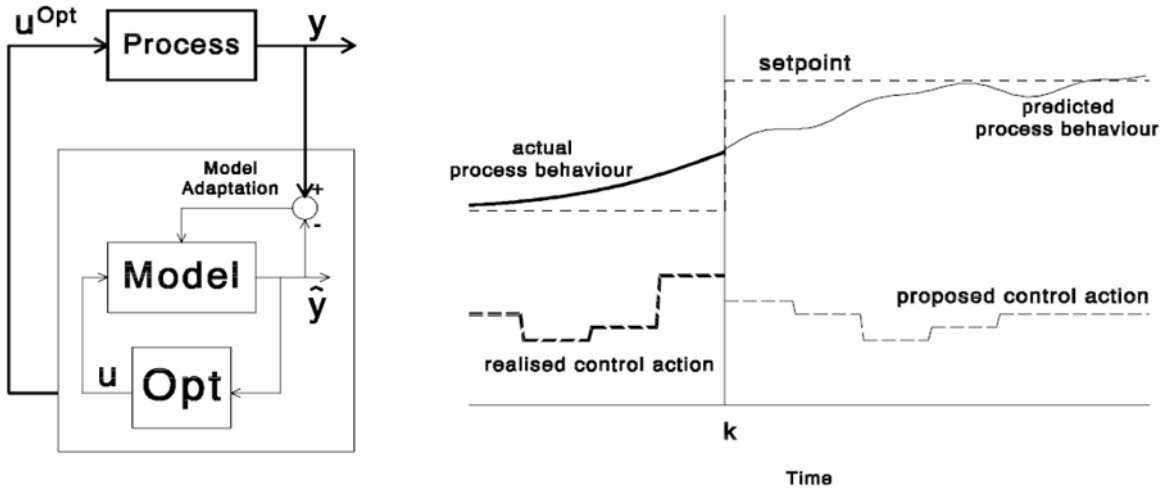


Figure 2-3: a) Principle of Model-based Predictive Control; b) Typical calculation performed by a Model-based Predictive Control algorithm (Vanrolleghem, 1995)

Stare et al. (2007) compared several commonly applied control strategies that include constant manipulated variables, various PI and feedforward control strategies, with model predictive control using an ideal process model. Results of the study show that with PI and feedforward controllers almost the same optimal operating costs can be achieved as with more advanced MPC algorithms

under various plant operating conditions. However, they conclude that MPC algorithms are advantageous in cases where the plant is highly loaded and if stringent effluent fines are imposed by legislation (Stare et al., 2007). Nevertheless, MPC has so far not been widely applied in WWT process control. Arguments are the complexity of the biological treatment processes and the mechanistic models that are required for simulation; as well as a lack of documentation of how to handle the nonlinearities encountered in the processes (Newhart et al., 2019).

Wastewater treatment has some specific characteristics when compared to other process industries, such as highly variable flow rates and pollutants loads that need to be removed to very low values under all circumstances. Unlike many industrial processes where an approximating linear model can be used to implement in a model predictive controller (MPC) (Borrelli et al. 2017; Rawlings et al. 2017), wastewater processing consists of several interacting physical, biological and chemical processes for both water and the settled sludge that exhibit highly non-linear and non-stationary (time-varying) behaviour, so that non-linear versions of MPC must be applied (Bernardelli et al., 2020).

Due to the large number of unmeasured states and the inherent non-stationarity related to wastewater treatment processes, model predictive controllers (MPCs) have been proposed as a better alternative for process optimization (te Braake et al., 1994). Model predictive control (MPC) has been developed considerably in the last decades both in industry and in the academic world. It is considered an appropriate way of dealing with a control problem in the time domain. Model predictive control integrates optimal control, control of processes with constraints, with dead-time, multivariable processes and uses future references when available. The finite control horizon allows the explicit handling of process and operational constraints (Camacho et al., 2010).

The control of linear processes with a linear MPC that applies linear constraints is a relatively well-established research field and the use of linear models are good approximations if the process is kept close to an operating point and the nonlinearities are not too severe (Camacho et al., 2010). However, in cases where systems are non-linear, such as wastewater treatment processes, a non-linear model is better suited.

More recently, several studies apply MPC control based on Artificial Intelligence (AI) models. A MPC with a nonlinear prediction model using either neural networks (Han et al., 2014; 2012) or fuzzy models (Yang et al., 2014) has shown to improve performance when compared to a MPC using a mechanistic model. Bernardelli et al. (2020) describe the design and testing of a MPC based on a neuro-fuzzy technique that is capable of estimating the main process variables and providing the right amount of aeration to achieve an efficient and economical operation. Their algorithm has been field tested on a large-scale municipal wastewater treatment plant of about 500,000 PE with encouraging results in terms of better effluent quality and energy savings (Bernardelli et al., 2020).

Another novel approach to MPC is to integrate Stochastic Differential Equations (SDEs) to predict plant performance, which provide a natural method to represent how mean and variance of a phenomenon evolves in continuous time (e.g. Stentoft et al., 2019). The proposed principle is used to reduce the operating costs of a wastewater treatment plant by enabling the flexibility to distribute the aeration load to periods with less expensive electricity prices. The performance of the proposed method is demonstrated for the operation of a single wastewater treatment plant and shows modest savings of 1.15% if only the day ahead market electricity price is considered. However, if the regulating and

special regulating electricity prices are included in the optimization the realized savings are in the magnitude of 7.23% and 27.32%, respectively (Brok et al., 2020).

2.4.3 Continuous Monitoring of Process Quality Indicators

Notwithstanding the progress made towards the implementation of smart tools, the importance of data quality checks, such as through autocalibration and built-in sensor checks, cleaning systems and reliable sample preparation units for data quality, has not diminished. Data preparation methods are still necessary as practical solutions to the frequently encountered problems with in-situ sensors in wastewater streams (e.g. Vanrolleghem and Lee, 2003).

Changes in the quality of influent and effluent flows into and out of a treatment system are important to detect early on as they serve as a strong indication of process performance. In-line monitoring allows early detecting and identifying any fault or abnormality that might negatively affect the process (Aguado and Rosen, 2008). The increasingly strict regulations have created a need for rapid and reliable monitoring techniques in real-time (Bourgeois et al., 2003).

Traditionally, in wastewater treatment plants there was often infrequent monitoring and/or automated systems are not as well developed as in other process industries. This is mainly because of the harsh environment – when compared to (drinking) water treatment - in which sensors have to be located (Lynggaard-Jensen, 1999) and the associated problems with instruments that foul, lose their measuring sensitivity and reproducibility, as well as have a need for frequent cleaning and recalibration (Bourgeois and Stuetz, 2000). Also, the high costs for investment and maintenance of a monitoring network with in- and off-line sensors often formed a bottleneck (Dürrenmatt and Gujer, 2012).

Continuous monitoring of quality variables is based on high-frequent values reported within short time intervals of a few minutes to a few hours by on-line sensors and instruments. These can actually collect data at frequencies up to a few seconds. Often the information content of this data is, however, relatively low since the dynamics of the process being measured are much slower than the sensor's sampling frequency (Olsson 2012). The reported values are therefore often the average or minimum values of a group of measurements taken over a longer period (e.g. every 15 minutes the minimum value measured from every minute gets reported back to a controller).

Continuous monitoring by inline sensors can support or even replace sampling either manually (grab) or by a composite sample that combines small volumes sampled over a day. The advantage of continuous monitoring is that process control can now be enforced more strictly to minimize cost (e.g. dosing of disinfectant). However, with the higher frequency and the harsh environment sensors find themselves in, faults are omnipresent. Pre-processing is therefore critical for the data to be trustworthy since the risk of failure is high (Therrien et al., 2019). It is noteworthy that laboratory analyses are considered to be of higher quality and are (still) the regulatory norm applied for monitoring quality variables.

There has been a technological development of sensors capable to be placed in-line in wastewater streams to monitor dissolved oxygen, suspended solids and organic matter, as well as several nutrients (such as nitrogen and phosphorus) over the last few decades. Today, these sensors can provide reliable data and have become more affordable (Ingildsen and Olsson, 2016). See for example the research by the LEESU laboratory in the framework of the Mocopée program that has resulted in the development of a 3D-fluorescence spectroscopy for rapid measurement of dissolved organic matter in wastewaters

entering the treatment plants (Goffin et al., 2018). Here, it is expected that fast and accurate characterization of the qualities and amounts of organic matter should achieve significant savings in terms of external carbon dosing and energy consumption for recirculation flows and aeration.

As mentioned, high-frequent measurement signals require precautions to be taken in order to produce good data quality, especially in areas with high effluent quality requirements (e.g. Papias et al., 2018). Methodologies for sensor management, installation and maintenance need to include data pre-processing techniques. Sensor calibration often involves an evaluation of performance in the laboratory and an adapted field protocol for calibration, data treatment and validation. It is crucial to consider the technical limitations of the measuring equipment and set appropriate maintenance and calibration methodologies in order to ensure an accurate interpretation of data (Papias et al., 2018). The huge amount of data recorded by inline sensors requires proper techniques to extract useful on-line information from the data (Rosen et al., 2003); it can then serve as input data for simulation models.

2.5 Process Modelling

Key to make improvements in the efficiency of WWTPs is our understanding of the physical, biological and chemical processes taking place inside a reactor. Modelling is a powerful tool for studying the behaviour of microorganisms and is advancing at a fast pace. Models based on mechanistic knowledge (such as empirical and phenomenological descriptions) are providing deeper insight in the biochemical and hydraulic processes and have found their way to support performance evaluation and decision-making for planning, design, control and optimization of systems (Rieger et al., 2012). Andrews (1976) already established that the dynamic behaviour of wastewater treatment should be considered during both plant design and operation so that performance of wastewater treatment plants would substantially improve. He argued that modern control systems for modification of dynamic behaviour should be incorporated into operations. Several powerful tools are mentioned for the study of dynamic behaviour and included among these are dynamic mathematical models and computer simulation (Andrews, 1976).

A conceptual model is an abstract representation of a physical system or process that requires assumptions or hypotheses to be made to simplify that reality. A mathematical model formulates these concepts so that the model can be applied to provide quantitative answers (Dochain & Vanrolleghem, 2001). Dynamical models describe the time evolution of processes. A further distinction is made between mechanistic models that use mathematical formulas based on the knowledge of physical laws (‘First Principles’), and analytical or empirical models that use empirical relationships based on the observed data to describe the behaviour of a process (‘Data-based models’).

In the wastewater sector, dynamic mechanistic wastewater treatment process models that describe the activated sludge process have since then become widely accepted and recognized as an important decision-making tool for design and process engineers, operators and managers. On the contrary, dynamic models that simulate the fixed growth processes inside a biofilm are still under development and do not yet have well established guidelines within the modelling community (Rittmann et al., 2018).

Consequently, there is a need for tools that aid in a better understanding of the (emerging) processes and support in decision making that guide utilities to strategically plan the implementation of

innovative technologies. Mathematical modelling of processes and process control tools can provide this support and permit WWTPs to transition from wastewater treatment towards resource recovery.

2.5.1 Mechanistic Process Modelling of Biofilm Systems

A mechanistic mathematical model is a systematic attempt to translate the conceptual understanding of a real-world system into mathematical terms. The main goals for biofilm modelling have been identified as increasing our understanding of the fundamental mechanisms of biofilm behaviour, as a tool to study the integration of different types of mechanisms occurring at different spatial and temporal scales; for experimental design and the evaluation of novel process designs; for the testing of operational strategies in order to improve process performance (Eberl & Morgenroth, 2006).

Eberl & Morgenroth (2006) describe the following steps in the development of such model: 1. identification of the important variables and processes acting in the system; 2. representation of the processes by mathematical expressions; 3. combination of all mathematical expressions into equations of mass or energy balances; 4. assignment of correct values to the model parameters; 5. solving the mass and energy balances by a technique that fits the complexity of the equations; 6. analysis of the model's output.

Dynamic models can be developed by applying material and energy balances to a reactor using fundamental transport, stoichiometric, thermochemical and kinetic relationships. Components in biofilm modelling are divided into particulate and dissolved components. This distinction is necessary since the processes related to particulates are much slower than the processes affecting the solutes. Furthermore, only dissolved components can undergo diffusion. Since that is the simplest approach to program and solve, most models apply a homogeneous mix of all components in the same proportions in all parts of the biofilm. A more complex approach involves the prediction of the spatial distribution of the components by solving the mass balances of all the particulates at different locations simultaneously throughout the biofilm.

The foundation of any mathematical biofilm model is the mass balance that states that for any time and location in a reactor, the net rate of accumulation (or total mass) for any component is the sum of all rates of transfer and transport added to the sum of all rates of transformation. A mathematical model translates these terms into expressions that need to be solved.

Mathematical models can be designed in a large variety; from 'simple' empirical correlations to complex and computationally intensive algorithms that describe three-dimensional biofilm morphology (Eberl & Morgenroth, 2006).

In contrast to a mechanistic model, an empirical model estimates the performance of a system based on finding patterns from a large quantity of data obtained under relevant operating conditions. However, most empirical models have a low performance with predictions made outside the range of the evaluated data.

Mechanistic models are developed based on the first principles of physics resulting in equations of mass, momentum and energy balances written for the system under consideration. Since these models are based on process knowledge, the mathematics are 'transparent' so that a change of model input can be traced by the codified relationships to an expected output. In the 1970's, the first-generation mechanistic models connected substrate fluxes into a biofilm layer to the utilization of substrate and

the transport of mass using one substrate that was considered as rate-limiting. These models were capable to simulate concentration profiles within the biofilm, although they contained the simplest possible biofilm geometry in the form of a homogeneous “slab” and a uniform biomass distribution. Examples are Harris and Hansford (1976), Harremoës (1976), LaMotta (1976), Rittmann and McCarty (1980, 1981).

From the 1980’s, second-generation models were able to describe a non-uniform biomass distribution inside the biofilm, that consisted of several substrates and included multiple types of bacteria. These models still maintained a simplified one-dimensional geometry, but spatial patterns for substrates and biomass type were included. Examples are Kissel et al. (1984), Wanner and Gujer (1984, 1986), Rittmann and Manem (1992).

Starting in the 1990’s, the next generation models provided highly detailed and complex 2D and 3D biofilm morphologies, which are often based on discoveries made by powerful analytical tools, such as laser scanning microscopy. These models provide mechanistic representations for the controlling factors of complex biofilm formation. A distinction of these models can be made based on the numerical solutions applied to the treatment of solutes and particulates. Concerning the soluble components, this could by the finite-difference method (Picioreanu et al., 1998a, 1998b, 2001, 2004; Noguera et al., 1999b; Pizarro et al., 2001; Laspidou and Rittmann, 2004) or by discrete methods (Pizarro et al., 2001; Noguera et al., 2004). For the computation of the distribution of the particulates, models could apply the cellular automata method (Picioreanu et al., 1998a, 1998b; Noguera et al., 1999, 2004; Pizarro et al., 2001; Laspidou and Rittmann, 2004); an individual based particles method (Kreft et al., 2001; Picioreanu et al., 2004) or the continuum method (Eberl et al., 2001). The continuum approach provides an average concentration of each component, while the two other methods describe the behaviour of individual cells. Overall, these third-generation models are computationally demanding and require a high level of expertise to operate (Eberl & Morgenroth, 2006).

Mechanistic models provide a representative description of a system or process derived from a fundamental understanding of its underlying mechanisms based on physical, biological or chemical laws or causal relations. This type of domain knowledge is then encapsulated in a set of mathematical functions, whose behaviour can be described by one or more mathematical equations (Schneider et al., 2022).

Mechanistic models can be classified according to the source of their knowledge; first-principles models are structured completely based on natural laws (the foundational understanding of a process or system); while phenomenological models describe empirical relationships and are structured according to domain knowledge, but where the observational data determine their parameters. Examples of the latter are Monod kinetics in ASM (Henze et al. 2000) or settling velocity functions in settler models (Takács et al. 1991).

The underlying dynamics are described using ordinary differential equations (ODEs), differential-algebraic equations (DAEs) or partial differential equations (PDEs) for the more complex water treatment systems, although algebraic equations can sometimes adequately be applied to simpler systems. In addition, there is a wide range of alternatives for equation-based models, including agent-based models, population balance models, and equation-free modelling (DeAngelis & Yurek 2015). However, these alternative types of modelling techniques have not found a wide acceptance within the wastewater sector (Schneider et al., 2022).

2.5.2 Mechanistic Modelling of Biofiltration

The modelling of biofiltration requires the combination of submodules that integrate the biological kinetics happening in the biofilm, the transport of solubles and particulates, the filtration of particulates and a backwashing model; each keeping track of the mass balances per component within each compartment. The development of mathematical models allows to perform dynamic (that is time dependent) simulations of the phenomena happening inside the biofilm. Unidimensional models describe the biofilm structure and the distribution of components in one direction, perpendicular to the surface of the biofilm. This keeps computational requirements relatively low (as compared to 2D or 3D models) and allows using the more commonly available commercial software products, such as WEST (DHI), GPS-X (Hydromantis), Simba (ifak), Sumo (Dynamita) and Biowin (EnviroSim).

For submerged biofilter systems, relatively few modelling studies have been reported in the literature. Most of the models include a transport model along the filter, a conversion model for the biological reactions in the bulk liquid and the biofilm, while some include a filtration or a head loss prediction model.

The foundation of the current process models was laid by the development of the Activated Sludge Model Number 1 (ASM1) first published in 1987 by the Task Group on Mathematical Modelling for Design and Operation of Activated Sludge Processes of the International Association on Water Pollution Research and Control (IAWPRC). It was the product of five years of research and resulted in the presentation of not only the ASM1 model, but also a guideline for wastewater characterization and development of programming codes, as well as a set of default values that have up till today not changed much (Henze et al., 2000). The ASM1 was designed with the specific characteristics of wastewater and the biological treatments in mind, so that the system of differential equations in the model is limited to the components being considered, and the process rates are selected as a function of the biological treatment conditions (Costa, 2022).

The ASM1 model was well received in the scientific community and served as the basis for further wastewater model development. In particular the so-called Guijer matrix notation that was introduced in the ASM1, which combines the soluble and particulate components affected by each of the simulated biological processes with their reaction rate terms, proved to be very practical to communicate about these complex models (Henze et al., 2000). The introduction of the ASM model was of great importance, providing researchers and practitioners with a standardised set of basis models (Gernaey et al., 2004).

The models became more complex over the years and included biological phosphorus removal (ASM2 in 1995), as well as denitrifying phosphorus-accumulating-organisms (PAOs) (ASM2d in 1999). To follow technological innovations in activate sludge processes, the ASM3 model (1999) was developed as a new platform to include the monitoring of internal storage components (Henze et al., 2000). In order to provide a complete description of the steps in the nitrogen cycle, a more detailed description of nitrification and denitrification with respect to ammonia oxidation inhibition, nitrite accumulation, and emissions of nitric oxide and nitrous oxide, was included in the ASMN model (Hiatt and Grady, 2008). The ASMN incorporates two nitrifying populations of ammonia-oxidizing bacteria and nitrite-oxidizing bacteria-using free ammonia and free nitrous acid, respectively, as their true substrates. The ASMN incorporates four-step denitrification (sequential reduction of nitrate to nitrogen gas via nitrite, nitric oxide, and nitrous oxide) using individual, reaction-specific parameters (Hiatt and Grady, 2008).

More recently, the scope of the ASM1 has been further enlarged by including new components or processes to the original model, such as the integration of the cellulose fraction (Reijken et al., 2018) and the formation of microorganism-derived dissolved organic nitrogen (Hu et al., 2020).

Thanks to the flexible configuration of the ASM1 that allows the matrix to be easily modified or expanded, the application of this model for biological reactions is not only possible for activated sludge processes with suspended bacterial growth, but also for other biological treatments with fixed growth. The ASM1 represents the basis for the formulation of new mathematical models for nutrient removal in WWTPs (Santos et al., 2020), in granular activated carbon filters (Acevedo et al., 2021), and for the application of activated sludge models to membrane bioreactors (Spérandio & Espinosa, 2008; Fenu et al., 2010; Orhon et al., 2021). An ASM1 model has been adapted to include kinetic reactions for the growth and decay of microalgae (Wagner et al., 2016).

The ASM model equations are founded on a simplification of complex processes. Moreover, there are still some relevant processes such as the formation of nitrous oxide (a potent green-house gas) and the conversions within the biological phosphorus removal process (due to the variability of phosphate accumulating organism metabolics (Gebremariam et al. 2011), which are not understood clearly enough to be put into a model for use in simulations. These knowledge gaps explain a part of the limitations of ASMs. Regmi et al. (2019) list several limitations, in particular the lack of ability to properly account for solids retention times (SRT), an inability to predict sludge settleability, the consideration of only one microbial group for a single process (e.g. nitrification) (Regmi et al., 2019).

Additionally, the half-saturation values used in ASMs do not truly represent intrinsic affinity constants but rather lump the effect of diffusion and spatial gradients of local environments (Arnaldos et al. 2015; Baeten et al. 2018). Substrate diffusion is largely dependent on floc properties which are a function of the local shear conditions, cohesion forces related to the exopolymeric substances characteristics, and turbulence intensity (Chu et al. 2003). This aspect of ASMs limits the model's ability to predict the outcomes of processes like simultaneous nitrification-denitrification as well as the conversion of micropollutants in biological wastewater treatment systems (Regmi et al., 2019).

In order to answer the many questions in the academic research world about the applicability of ASM models to simulation models of biofilm reactors (e.g.: Boltz and Daigger, 2010; Morgenroth and Wilderer, 2000), a Scientific and Technical Report (STR) was compiled with a focus on the “Mathematical Modelling of Biofilms” (Wanner et al., 2006). A key conclusion in the STR is that for many (not all) engineering applications 1-D biofilm models (as opposed to 2-D or 3-D) are sufficiently complex in representing local substrate and biomass gradients. Key differences in biofilm reactor modelling are related to describing mass transport limitations and a heterogeneous distribution of biomass components over the thickness of the biofilm. These differences are in particular (i) Diffusion of soluble and colloidal components, (ii) Attachment, detachment, and movement of particulate components within the biofilm, (iii) External mass transfer resistance (e.g., described as an external mass transfer boundary layer thickness). These biofilm-specific processes are in turn dependent on the specific type of biofilm reactor and on reactor operation (Boltz et al., 2012).

2.5.3 Development of Biofilter Models

Behrendt (1999) developed a model of a two-step nitrification biofilter that consists of three compartments; a gaseous phase, a bulk liquid and a biofilm surface layer, where all the mass balance equations are computed for. Ammonia conversion is described by nitritation and nitratation. However,

the model does not account for the evolution of biomass growth and decay or the transport of particulates (Samie, 2009).

Falkentoft et al. (2000) developed a model based on the ASM2d describing denitrification and phosphorus removal in a biofilter. It introduced the concept of the initial thickness of the biofilm that after each backwashing session is the starting point for calculations.

Viotti et al. (2002) built a model for a carbon and nitrogen removal in a BioStyr© biofilter that allows to simulate concentration profiles of COD and NH_4-N inside the biofilm and along the column height of the filterbed. It integrates an accumulation of biomass that can interfere with the filtration efficiency and the evolution of the head loss due to clogging of the filter.

Hidaka and Tsuno (2004) designed a model for carbon and nitrogen removal that simulates the concentration profiles of COD, NH_4-N and NO_3-N inside the biofilm and along the filterbed. The column height of one filter is represented by 5 completely mixed reactors placed in series. Each reactor contains three compartments (filterbed, biofilm and liquid) where transport, accumulation and release of particulates are simulated. However, the model does not account for diffusion of solubles and assumes complete substrate transfer into a uniformly distributed biofilm consisting of one layer. During each backwashing session, a constant part of the biomass is removed. The biological reactions are based on the ASM1 model (Henze et al., 2000b), while the authors have included heterotrophic endogenic denitrification, assuming that the heterotrophic biomass is capable of using nitrates as substrate. The model was tested with experimental data from a laboratory set-up. However, the model conditions (one layer for the biofilm; no soluble diffusion) restrict the simulation of different types of niches inside the biofilm and the transfer of solubles.

Vigne et al. (2010) applied a commercially available model (GPS-X) to a nitrification biofilter with operational data from a large WWTP. In the dynamic model, the biofilter was divided into six completely mixed reactors with an upwards flow direction, where each reactor contains a 5-layer biofilm, a liquid phase and a MTBL boundary layer with a fixed thickness. Here, solubles transfer by diffusion is included for the interface from the bulk liquid into the biofilm, as well as between the biofilm layers. This approach allows simulating a partial diffusion and the development of niches within the biofilm. The processes of filtration and backwashing are described, and a partial mixing of the support media occurs during washing of the filter bed reducing the vertical gradient of biofilm properties along the length of the biofilter. The authors apply an index to follow the head loss due to clogging of the filter. The biological reactions are based on the ASM1 model (Henze et al., 2000b). The simulation results show that the model is overall capable to predict the ammonia concentrations in the effluent under dynamic conditions. However, the model assumes a fixed dissolved oxygen content in the bulk liquid phase of each reactor, which is not realistic; and does not accurately simulate the headloss evolution.

Bernier et al. (2014b) have continued the development of the biofiltration model as designed by Vigne et al. (2010), but in a Matlab/Simulink (MathWorks) environment and named it SimBio. It uses a similar approach by simulating one biofilter that is divided into seven completely mixed reactors placed in series, thus adding one reactor as compared to the previously described Vigne et al. (2010) model. The number of biofilm layers was however reduced from five to two layers per reactor in order to speed-up calculations. The SimBio model accounts for variable dissolved oxygen concentrations in the bulk liquid as well as the effects of aeration flow rates and temperature on the oxygen transfer rate. It

computes headloss in function of the thickness of the biofilm layers. The ASM1 model applied for the biological reactions was modified to include nitrification and denitrification as two separate reactions by adding AOB and NOB type bacteria. For heterotrophic denitrification under anoxic conditions, the ASMN (Hiatt and Grady, 2008) was applied. In addition, the assimilation of orthophosphates by bacteria was added to the SimBio model. The model calculations however do not allow estimating operational costs and the emission of greenhouse gasses.

The Bernier et al. (2014) model was calibrated and validated with operational data from biofilters from various large WWTPs in the Paris region that include carbon removal, nitrification and post-denitrification. The results show that the model is largely capable of predicting effluent quality concentrations and filter headloss for the different biofilter configurations applying averaged conditions as the model input. Since then, the Seine aval WWTP has undergone several upgrades to the secondary treatment stage by adding pre- and post-denitrification phases. The model is therefore no longer in agreement with reality and would require updates and recalibration of the parameters.

Fiat et al. (2019) provide a model extension to the SimBio model to include the main nitrous oxide (N_2O) biological pathways during nitrification and denitrification. Modifications of the model structure were made to include the gas phase as a model compartment containing four common gasses in the nitrogen cycle (O_2, N_2O, NO, N_2). The mass balance on the gaseous compounds was expanded to correctly describe the N_2O gas-liquid partition and allow for calculations of N_2O emissions, based on the equations of N_2O in the ASMN (Hiatt & Grady, 2008) and Pocquet et al. (2016). The model was calibrated with full-scale data from a large WWTP with nitrifying biofilters. The results indicate that the model is capable of predicting nitrous oxide emissions without diminishing the performance of the predictions of other nitrogen variables, such as NH_4 and NO_2. The model increased the understanding of microbial behaviour and was able to demonstrate that the heterotrophic denitrifiers are capable of consuming a part of the N_2O produced during nitrification; thereby lowering this greenhouse gas emission.

Finally, Zhu (2020) developed an enhanced biofiltration model based on the model of Bernier et al. (2014) that is applicable for denitrification and nitrification biofilters. Compared with the original model, a variable thickness of the MTBL boundary layer (instead of a constant thickness) and five biofilm layers were again used (instead of 2 biofilm layers) to describe the mass transport process and the biofilm heterogeneity more precisely. Concerning the particulate components, the new model does not take into account detachment by shear forces, as the effect of such events was determined to be negligible in the original SimBio model. However, the model does take into consideration a mixing of the support media during washing, and extended this phenomenon also during normal operation, albeit at a lower rate. To maintain the spatial distribution of the components, the exchange of particulates is only possible in the model between the reactors directly underneath and directly above it and always within the same biofilm layer. Estimations of energy consumption (aeration and pumping) and biological production of N_2O as well as a simplified gas-liquid transfer process were integrated in the model in order to evaluate treatment costs and GHG emissions respectively. The aeration control scheme used in the plant was included in the model to better predict the necessary air flow injected during treatment. The model was calibrated and validated with intensive (15-minute interval) operational data from a large WWTP collected over several months.

Notwithstanding that mechanistic biofilter models provide insight of the behaviour of biological, physical and chemical processes, such wastewater treatment systems are highly nonlinear and time-depending resulting in complex relationships between model inputs, parameters and outputs variables. Other sources of error for model predictions include inaccurate or incorrect calibration, non-ideal process behaviour and lump-sum parameter assumptions (Olsson and Newell, 1999).

In addition, the cost of model development is high, in particular when the process system is large and has a complex non-linear behaviour. A good understanding is required to calibrate the model parameters needed to fine-tune the model and make accurate predictions. Overall, it thus remains difficult to apply mechanistic biofilter models to simulate complex treatment systems with a high precision, which is a requirement for process control and optimization.

2.5.4 Mechanistic Model Calibration

Without calibration and only default parameter values, an ASM model may produce results that are only accurate within an order of magnitude related to the variability of the measured data (Gujer, 2011). Calibration of the model parameters is therefore crucial. Even though calibration protocols have been developed, model calibration is still regarded a hindrance to modelling (Mannina et al., 2011). What follows is an overview of the literature on model calibration, in particular for knowledge-based biofilter models.

Model calibration aims to adjust the values of certain parameters so that the simulation results are as close as possible to the observations. This is followed by model validation, where different data sets are used to verify model performance (Rieger et al., 2012).

Rieger et al. (2012) proposed unifying guidelines for the calibration and validation of mechanistic ASM models by describing the key points that should be adhered to. It contains the internationally available knowledge and experiences of the members of the Good Modelling Practice (GMP) Task group of the International Water Association, and thereby sets the standards for a widely acceptable protocol. Inconsistent approaches and insufficient documentation had made quality assessment and comparison of modelling results difficult.

Prior to that, other guidelines for ASM model calibration had been published that served the modelling community. The most widely circulated protocols in the beginning of the 2000's were: (i) STOWA (Hulsbeek et al., 2002; Roeleveld & van Loosdrecht, 2002), (ii) BIOMATH (Petersen et al., 2002; Vanrolleghem et al., 2003), (iii) WERF (Melcer et al., 2003) and (iv) HSG (Langergraber et al., 2004). A SWOT analysis of these can be found in Sin et al. (2005), while an overview is also given in Rieger et al. (2012).

Nevertheless, most of the model applications that had been developed were mostly based on ad hoc approaches and expert knowledge. Moreover, each modelling study was using a different calibration approach: e.g. different influent wastewater characterization methods, different kinetic parameter estimation methods, different selection of parameters to be calibrated, different priorities within the calibration steps, etc. The lack of a standard approach made it difficult to compare different calibrations of ASMs with each other and to perform internal quality checks for each calibration study (Sin et al., 2005). These authors performed a critical analysis of the existing protocols and provided a list of the strengths, weaknesses, opportunities and threats. The study strongly suggested to test these different protocols on a case study and to develop a unified protocol combining all the strengths and

opportunities of each protocol. An interesting recommendation made was to further research the potential of automating some steps of the calibration procedure by use of mathematical algorithms (Sin et al., 2005).

The unified good modelling practice (GMP) protocol was built around five main steps. It starts with project definition, where its definition, objectives and requirements are defined. The next step is data collection and reconciliation. This is followed by the model set-up, selection of submodels and model checks. The fourth step is the model calibration and validation, and the last step is the simulation and results interpretation (Rieger et al., 2012).

A more detailed and systematic procedure for the calibration of activated sludge models was presented by Mannina et al. (2011) that was based on a comprehensive global sensitivity analysis and a novel step-wise Monte Carlo-based calibration of subsets of influential parameters. The calibration procedure starts in phase 1 with a prior screening of the most influential model parameters and by simplifying the problem of finding the optimal parameter set by splitting the estimation task into steps. In the second phase, the calibration is undertaken for sub-groups of output variables instead of solving a complex multi-objective function. However, even with this step-wise approach parameter identifiability issues may occur, but this is dealt with by using the general likelihood uncertainty estimation (GLUE) method, that so far has rarely been used in the field of wastewater modelling (Mannina et al., 2011). A real case-study demonstrated the effectiveness of the proposed calibration procedure.

Different calibration protocols that have been described over the past decades were mainly focused on activated sludge models where the biomass is present in suspension. As mentioned earlier, these protocols cannot be directly applied to models that simulate a fixed biomass growth (Boltz and Daigger, 2010; Rittmann et al., 2018). Moreover, the GMP protocol developed by the IWA task group (Rieger et al., 2012) does not go into detailed description of calibration procedures and remains applicable only as general guidelines.

Few biofiltration model researchers describe in detail how they have calibrated their models (Bernier, 2013). Conversely, Vigne (2007) does propose a protocol for the modelling and calibration of a nitrifying biofilter that starts with a local sensitivity analysis of the model parameters and their impact on selected effluent variables. The analysis is performed by changing one parameter at a time. A difference is made between parameters that are structural and measurable and therefore can be fixed at a value and those that are non-measurable or only indirect measurable. The calibration is then described step by step, starting with the parameters that are influential on the ammonia (NH_4-N), followed by nitrate (NO_3-N), TSS and head loss. Although well described, this method is applicable only for this specific biofilter configuration using a dataset and was not easily applicable elsewhere (Bernier, 2014).

For the modelling of a biofilter (Bernier et al., 2014a & 2014b) described their calibration procedure in more detail (Bernier, 2014). The effluent variables to which the model parameters were calibrated were selected on the basis of the filter configuration (e.g. denitrifying and nitrifying filters). The calibration started with a global sensitivity analysis to identify the influential parameters on the selected effluent variables, which was performed by Monte-Carlo simulations and a multiple linear regression based on the method described by Saltelli et al. (2007). The Monte-Carlo simulations were based on model parameters randomly selected within a range between minimum and maximum

values using Latin Hypercube Sampling (LHS). This sampling method divides the space in equal sizes and selects a value within each subspace. The LHS allows covering effectively and statistically the entire sampling space with just a small number of samples. The parameter range was based on expert opinion and default values found in literature. The sampled parameter values were then used for simulation. As mentioned, a multiple linear regression was applied to the results of the modelled variable (Y) and the parameter values (X) and the standard regression coefficients obtained allowed identifying the influential parameters. The performance indicators used for evaluation were the mean error (ME) and the mean absolute error (MAE) (Bernier, 2014).

For his modelling work, Zhu (2020) provides a detailed description of his calibration method based on the two-step approach developed by Mannina et al. (2011). It started with the selection of practically identifiable parameters using a global sensitivity analysis. Out of a total of 142 parameters of the biofiltration model, 49 parameters such as biofilter dimension and physical constant parameters were set at fixed values and did not participate in the analysis. The remaining 93 parameters were sampled using a LHS of 473 experiments. The sampling range for each parameter was 20% of the default parameter value for relatively well-known parameters, such as biological kinetic parameters used in the ASM1 model, while for parameters with less available information, the sampling range was extended to 50% (cfr. Brun et al., 2002). The sampled parameter values were then screened by a Monte-Carlo simulation in steady state, based on the method further described by Sin et al. (2011). According to the standardized regression coefficients (SRC) obtained from the global sensitivity analysis (see Mannina et al., 2011), the most influential parameters for different effluent variables were identified and divided into the following independent calibration groups: (1) filtration and carbon removal (effluent TSS and COD), (2) ammonium removal and oxygen consumption (effluent NH_4, NO_3 and DO), (3) nitrite production and removal (effluent NO_2), (4) orthophosphate consumption (effluent PO_4) and (5) N_2O production (N_2O emission factor). The first two groups hold about ten influential parameters and the other groups five parameters (Zhu, 2020).

For the second step in the calibration method, Zhu (2020) performed Monte-Carlo simulations for the first four calibration groups that contain effluent quality variables. Each sampled parameter set was used for a 30-day steady-state simulation to initialize the model followed by a dynamic simulation over another 30-day period during which nutrient loading decreased due to summer holidays. To identify the best parameter values, differences between simulations and measurements were quantified by calculating the mean error (ME), mean absolute error (MAE) and root mean square error (RMSE). When one group of parameters was calibrated, the parameter values of this group were fixed and the next group was tackled (see Mannina et al., 2011). After the Monte-Carlo simulation, the identified parameter values of the four groups were validated with longer term data and parameters adjusted manually if necessary (Zhu, 2020).

Since N_2O emission was not measured, the model parameters for the fifth group were calibrated approximately and adjusted according to the observed overall emission factor values reported in literature where measurement campaigns were realized (Bollon et al., 2016).

2.6 Evaluation of Model Performance

To allow for an objective evaluation of wastewater simulation model performance, efficiency criteria have been compiled and compared by Hauduc et al. (2015). A total of 30 efficiency criteria used in environmental modelling were classified into the following six groups according to different modelling

objectives: (1) single event statistics, (2) absolute criteria from residuals, (3) derivative error criteria, (4) relative error criteria, (5) total relative error criteria, and (6) comparison of residuals with reference values and with other models. In a first step, criteria with equivalent functional form were identified, leading to 18 groups. From each group a representative criterion was sub-selected for quantitative evaluation in an illustrative wastewater treatment plant modelling case study considering four target variables and two operating conditions (Hauduc et al., 2015).

The standard practice in activated sludge modelling (Rieger et al., 2012) is to compare the modelled effluent variables to the measured data and define statistical error indicators. The most commonly applied are the mean error (ME, equation 2.2), which measures the bias of the model (an optimal value close to 0 indicates a low bias); the mean absolute error (MAE, equation 2.3), which indicates the accuracy of the model without taking into account the direction (e.g. positive and negative) largest absolute error, and the root mean square error (RMSE) (equation 2.4), which measures the average difference between values predicted (P) by a model and the observed (O) values. The RMSE is another measure for model accuracy, where a value of zero would indicate a perfect fit to the data (Hauduc et al., 2015).

Alternative criteria calculate the relative criteria by dividing the indicators by the mean of the observations (ŷ). The relative error (or percentage error) is defined as the ratio of each of the absolute errors calculated earlier to the mean of the observations. This allows determining the magnitude of the absolute error in terms of the observed average value of the measurement. Relative criteria are dimensionless and allow for comparison across different state variables; it makes a clear distinction possible of the overall accuracy of the model. The closer the relative criteria values are to 0, the better the accuracy of the model (Hauduc et al., 2015). Finally, the Janus coefficient (equation 4), with an optimal value of 1, makes it possible to compare a model's performance across different datasets, such as one used for the validation with those data used for the calibration (Rieger et al., 2012).

$$ME = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i) \qquad Equation\ 2.2$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \qquad Equation\ 2.3$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \qquad Equation\ 2.4$$

$$Janus\ coefficent = \frac{RMSE_{validation}}{RMSE_{calibration}} \qquad Equation\ 2.5$$

2.6.1 Evaluation of Controller Performance

Monitoring and control strategies can be efficiently evaluated and compared by realistic model simulation as long as a standardised procedure is followed (Aguada and Rosen, 2008). The COST/IWA simulation benchmark (Copp, 2002) for control strategy development and evaluation has been widely used by the wastewater research community as a standardised simulation protocol. The Benchmark

Simulation Model no. 1 (BSM1) includes the model, control system, benchmarking procedure and evaluation criteria. The model parameters and influent files characterising the influent wastewater are based on default values and are all provided for.

The user can test any control strategy by setting the specified control handles and sensors. The evaluation is done according to some pre-defined criteria, based on a one-week period. Despite the usefulness of the BSM1, its limited evaluation period does not allow for long-term evaluation (Aguada and Rosen, 2008). Rosen et al. (2004) considered an extension to the BSM1 for long-term evaluation and comparison of control and monitoring strategies; named the long-term BSM1 (BSM1_LT). It also includes temperature variation and kinetic parameter dependencies, realistic models for sensors and actuators, as well as equipment and process faults. Jeppsson et al. (2006) developed an extension to allow control strategy development and performance evaluation at a plant-wide level, that is including the sludge treatment line (Aguada and Rosen, 2008).

2.7 The future of Mechanistic Modelling

As described above, process models have been developed and validated in the wastewater industry for designing, upgrading, and optimizing wastewater treatment plants, in particular for conventional activated sludge processes. However, due to the changing role and functionalities of WWTPs that include water quality, protection of water sources and the environment, energy reduction and production, and resource recovery, the scope and structure of the process models of the next generation could include new approaches such as thermodynamic, hydraulic, or economic models (Regmi et al., 2019). Some of the new approaches in the wastewater modelling field that have been proposed, describe physicochemical models (Batstone et al. 2012), energy and economic cost models (Rahman et al. 2016), greenhouse gas models (Mannina et al. 2016) and methods about how to integrate all these aspects in a plant-wide context (Solon et al. 2017).

Moreover, as Regmi et al. (2019) stated, hydrodynamics and mass transport are important for the optimum design and operation of WRRFs, or novel technologies dedicated to resource recovery. Computational fluid dynamics (CFD) models can not only simulate hydraulic phenomena, but they have a great potential for modelling physicochemical processes where gaseous, solid, and aqueous phases interact and can be a very valuable tool for the optimization of the aeration process or resource recovery (Samstag et al. 2016).

However, the increased complexity of new modelling approaches can become an obstacle for understanding and implementation of the model's output. Future model development will likely put emphasis on resource recovery (water, nutrients, organics, and energy) rather than wastewater treatment. The practice of design, operation and control of resource recovery technology will need models that consider stringent objectives related to water-product quality, process performance stability, and operating costs. Unit process models for resource recovery will likely be integrated within broader frameworks (e.g. automated dynamic process control, sustainability, etc.) and at various scales (e.g. sewershed, watershed) to target combined social, economic, and environmental goals. Effective cost and price models will need to be developed for the different parts of the WRRF value chain in order to provide input to economic assessments (Regmi et al., 2019).

2.7.1 Digital Twins

A Digital twin is best described as a simulation model of a physical counterpart (e.g., a process or system) that has an automated data connection, preferably bi-directional, that allows for dynamical

updates to the model so that it remains an accurate description of that reality and retains its relevance (Torfs et al., 2022). The use of a digital twin has the potential to reduce costs, improve resource recovery, increase water quality and increase customer engagement (IWA and Xylem Inc. 2019). Interest in Digital Twins has greatly increased in the recent years across both academia and industry, accompanied by a growth in the number of related publications, processes, concepts, and envisaged benefits (Jones et al., 2020).

The origin of using the DT terminology goes back to the early 2000s. However, it was first used in publications during the early 2010s as a new tool for the management of a product's lifecycle in the manufacturing industry (Glaessgen & Stargel 2012; Grieves 2014). Since then, the concept of DTs has spread throughout many different sectors including the water industry (Torfs et al., 2022). Industry-wide, this growth is largely driven by advances in related technologies and initiatives such as Internet-of-Things, big data, Industry 4.0, real-time sensors and sensor networks, data management, data processing, and a drive towards a data-driven and digital manufacturing future (Jones et al, 2020). For the urban water industry in particular, there's a rising interest in the development of DTs thanks to the advances in instrumentation and the increasing availability of online data and computing capacity for utilities (e.g. via cloud computing) as well as the business drivers around both improved capital and operational efficiency targets (Valverde-Pérez et al., 2021).

A recent White paper by the IWA on DTs (Valverde-Pérez et al., 2021) show a large potential for economic savings in the water sector with the application of DTs (e.g. online energy optimisation), more effective protection of the environment (e.g. model predictive control for effective nutrient removal) and increased benefits for society (e.g. improved storm water management to minimise risk of flooding in urban areas). Two direct added values of DTs are the visualisation of data (such as performance indicators) by digital twins (Garrido-Baserba et al., 2020) and the ability of a digital twin to use model-based control and data-driven optimisation that can achieve more reliable control with higher efficiencies than conventional process control (Valverde-Pérez et al., 2021). This can result in reductions of energy and reagent consumptions, and thus lower capital expenditure (Stentoft, 2020).

2.7.2 Definition of Digital Twins

A variety of definitions employed across industry and academia remain (Jones et al., 2020) and a uniform definition for a DT cannot consistently be found in publications. Moreover, the distinction between a conventional First-Principles process model and a DT is not always clear (Torfs et al., 2022). Grieves (2014) describes the Digital Twin as consisting of three components, a physical product, a virtual representation of that product, and the bi-directional data connections that feed data from the physical to the virtual representation, and information and processes from the virtual representation to the physical. Kolditz et al. (2019) define digital twins as virtual systems that 'contain all important characteristics and features of the real system, depending on the specific purpose for an application'. Digital twins can be regarded as combinations of models and real-time data that provide a digital representation of a specific part of the water system's behaviour (Valverde-Pérez et al., 2021). Torfs et al. (2022) give an overview of the components of a DT and its connections to the real entity and relevant stakeholders (*Figure 2-4*). Visible in the Figure is the real physical system ('Real WRRF Entity'), the virtual representation of it ('WRRF Digital Twin') as well as the various data connections that feed both components. The data connection from the system being monitored to the DT should be automated, while the inverse connection from the DT to the physical system (to the SCADA system)

could involve a human decision maker. External information sources, such as weather forecasts and water level predictions in the watershed or sewage network, can also provide data into the DT.

Torfs et al. (2022) list the following features of a digital twin: (i) a DT should be connected to a real entity (e.g. equipment, process or product); (ii) a DT should have an automated, live data connection to the real entity that is preferably in two directions. Whereas the data connection from the real entity to the DT should be automated, the virtual-to-real entity connection may be either automated (direct input from the DT into SCADA) or manually performed; (iii) DTs should evolve and follow changes that happen to the real physical system (Torfs et al., 2022).

A crucial aspect it seems of a DT is thus the data connection between the physical system and its virtual representation, which is something that sets it apart from conventional process models (Jones et al, 2020) and gives a DT a much higher level of usefulness as compared to a conventional simulation model executed independently (Torfs et al., 2022).

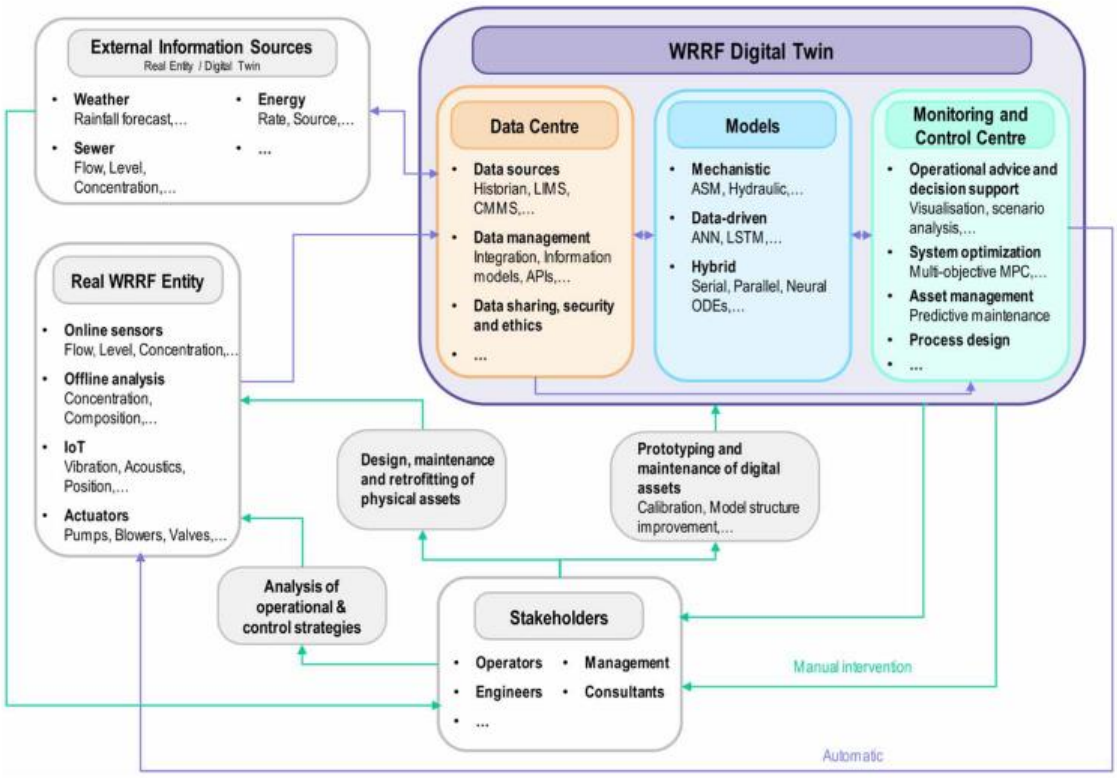


Figure 2-4: Schematic representation of a Digital Twin and its interaction with the physical system, control systems and stakeholders (Torfs et al., 2022).

Therefore, a digital twin is reliant on every part of the data pipeline mentioned before (see *Figure 2-1*). It receives extensive, fault-free and continuous data from the physical system; it represents process knowledge through a constantly updated model; it provides insight into the plant via interfaces, visualization and analytics, and it supports decision-making by operators and taking of actions via automatic control (Therrien et al., 2020).

Regarding the nature of real-time data, Torfs et al (2022) argue that the time horizon for model updates and simulation should be defined by the objectives of the digital twin (e.g. instrument failure detection, real-time control) and the relevant system dynamics (aeration vs. sludge age). The automated data-

feed and dynamic model updating defines a DT, while the time horizon over which this is performed, can remain flexible (Torfs et al., 2022).

2.7.3 Digital Twin applications in WRRFs

DT applications include (but are not limited to): (i) data-driven decision support for the selection of different operational strategies and operator training (Johnson et al. 2021); (ii) online (multi-objective) system optimisation (e.g. model-predictive control) for energy or resource savings or compliance management (e.g. to minimise carbon footprint) (Stentoft et al. 2020, 2021); (iii) failure analysis (Jain et al. 2020); (iv) asset management and predictive maintenance (Bartos & Kerkez 2021); (v) investment planning (Ruohomäki et al. 2018); and (vi) decision support for policy making (Poch et al. 2020). Although the potential for application of DTs is far-reaching, only a limited number of successful implementations are documented in the literature for WRRF systems, and they mainly focus on operational decision support and online system optimisation through advanced control. In the list above, for example, only Johnson et al. (2021) and Stentoft et al. (2020, 2021) describe applications in the WRRF sector, whereas other examples are taken from related domains such as water network management, smart cities, and the manufacturing industry.

Torfs et al. (2022) provide examples of application of digital twins for operational decision making and online optimisation, such as the DT developed for the Singapore PUB Changi Water Reclamation Plant (Johnson et al. 2021). It has been implemented as a hybrid model that uses both mechanistic and data-driven models and includes some level of automatic calibration through the use of both the SCADA and laboratory data connections. Also presented is a digital twin for model predictive control that allows for online optimisation. This DT consists of a catchment model (including both the sewer line and dry weather inlet flow forecast) as well as an effluent quality model, both being empirical models calibrated on an hourly basis. It is further described by Stentoft et al. (2020), who present practical examples of a digital twin developed for design of future WRRFs. Alex et al. (2020) illustrate how a detailed process model of a plant including its control system and actuators can be used for virtual commissioning of the automation system. This process model developed during the planning and design phase of the automation concept can continue its lifecycle as an operational DT once the automation system is connected to the real entity (Torfs et al., 2022). A very recent development is the design of a Digital Twin by Sparks et al. (2022) that serves to predict nitrifier populations and kinetics, where the authors deploy Machine Learning algorithms and model-based controllers for process optimization.

With the introduction of digital twins in the wastewater sector that include process simulators with real-time data feed, the demand for efficient data treatment and hybrid process models will only rise in the predictable future (Therrien et al., 2020).

2.7.4 Concluding Remarks

Increased data availability in combination with improved computational capacity will continue to shape the structure of future modelling frameworks. The newly developing synergy between first principles and data-driven models has the potential to create very powerful tools for further innovation, development and decision support. However, balancing the efforts for model development and complexity, data collection, data quality assurance, and integration of different frameworks will be challenging and will require diverse technical skills (Regmi et al., 2019).

2.8 Data-Driven Process Models

With the increasing number of big-data collected and stored, data-driven methods are gaining importance to extract useful information from the observations (e.g. Lee et al., 2006). Machine learning or data-driven models (DDM) calculate output variables exclusively from the input data by exploiting the ability of finding complex patterns within the data. They model the system behaviour solely by mapping the input to the output data rather than from process knowledge (Shirkoohi et al., 2022). The number and nature of the model parameters are flexible and not fixed in advance based on any knowledge of the system. To contrast these with mechanistic models, DDMs are also called non-parametric models, since the model parameters hold no physical representation of the system (von Stosch et al., 2014). This paragraph describes in a brief overview the usage of DDMs. For more background information on the use of DDM's and smart water technologies by SIAAP in the Greater Paris Region, see Annex II.

Over the past decade, DDM models have attracted a lot of attention in the academic world and found applications in the domains of image processing and natural language processing. Another recent example are the artificial intelligence chatbots that are powered by large generative language models that have been trained and fine-tuned with big-data using both supervised and reinforcement learning techniques. The advancements made with data-driven models have also found a highlighted interest in the process industry (e.g. Qin and Chiang, 2019; Shang and You, 2019, Bikmukhametov and Jäschke, 2020) and in the water sector (e.g. Al Aani et al., 2019) .In the wastewater sector there is also a growing interest in the application of data science in modelling of the operation and control of WWTPs (eg: Fan et al., 2018; Antwi et al., 2019; Bahramian et al., 2022). Data-driven models use statistical and data-based methods with machine learning models to find unknown relationships in the data and are very fast in terms of computation time (seconds to minutes). These models describe the system solely based on information extracted from the data and have strong interpolation capabilities (Newhart et al., 2019). This makes a DDM easier and faster to develop in comparison to a mechanistic model, although it requires larger volumes of data for training, testing and validation and is considered less–transparent than a first-principle model (von Stosch et al., 2014).

In the wastewater sector, data-driven analysis is being used for fault detection, variable prediction and advanced process control (Newhart et al., 2019). Examples of data-driven methods and models that are frequently applied in process engineering are Gaussian processes, polynomials, multivariate adaptive regression splines, partial least squares, principal component analysis, partial least squares regression, support vector machines, decision tree-based methods and artificial neural networks (Schneider et al., 2022). Multivariate statistical methods have been used for fault detection in process monitoring (Lee et al., 2005a, Lee et al., 2006, Rosen et al., 2003, Yoo et al., 2004) and software-sensor design (Jansson et al., 2002, Masson et al., 1999), while artificial neural networks have been applied for prediction (Belanche et al., 1999, Dellana and West, 2009, Kim et al., 2009, Raduly et al., 2007) and monitoring and control (Baeza et al., 2002, Luccarini et al., 2010).

Machine Learning models include a wide range of techniques that include artificial neural networks (ANNs), genetic algorithms (GA), support vector machines (SVM), random forests (RF), boosted regression trees (BRT), simulated annealing (SA), Monte Carlo simulation (MCS), particle swarm optimization (PSO), immune algorithm (IA), ant colony algorithm (ACA), imperialist competitive algorithm (ICA) and decision trees (DT). ANNs are frequently adopted to predict the pollutant removal processes because of their capabilities of self-learning and self-adapting, while genetic algorithm (GA)

and particle swarm optimization (PSO) efficiently search for the global optima (Fan et al., 2018). Shirkoohi et al. (2021) indicate that artificial neural networks (ANN) are an effective method for data-pattern matching. ANNs are popular tools in machine learning thanks to their ability to learn rather complicated functions, which are non-linear statistical data modelling and data mining tools (Hu et al., 2010).

A recent bibliometric review on the use of AI technology in wastewater treatment research (Zhao et al., 2020) found that the number of published articles applying AI to wastewater treatment research was 19 times greater in 2019 than in 1995, and papers had 36 more citations on average. Zhao et al. (2019) showed that several articles received more than 150 citations discussing the applications of ANNs and neural fuzzy models for simulating and predicting the performance of biological treatment (e.g.: Côté et al., 1995; Gernaey et al., 2004; Hamed et al., 2004).

Still according to the bibliometric analysis of Zhao et al., (2019), the main topics research has focused on are the application of Machine Learning models for the prediction of pollutant removal, the economic and energy efficiency of AI technology in its application to wastewater treatment, the contribution of AI technology to the management of wastewater treatment, and AI research on resource extraction and reuse. The study shows that ANN and fuzzy-logic models are the most widely applied methods and that, in most of the studies reviewed, were trained and tested with experimental data (Zhao et al., 2020).

Several other examples of DDM applications in water research are well documented by Newhart et al. (2019). Ebrahimi et al. (2017) used multiple regression models to predict concentrations of water quality variables (phosphorus, nitrogen, TSS) in a full-scale wastewater treatment plant. The model was capable to explain between 71 and 87% of the total variation in the dependent variable. However, the authors did not demonstrate that the model had sufficient accuracy for stand-alone fault detection (Newhart et al., 2019). Another example of variable prediction is the study by Vasanth Kumar et al. (2008). The authors trained a neural network model with experimental data under various conditions to predict equilibrium concentrations after the uptake of dye by powdered activated carbon (PAC). The results were nearly perfect (R^2 = 0.96). However, hundreds of data points were needed to calibrate the model prior to the predictions; there was not significant variation among the input variables; only a single contaminant was used; and separate testing data were not used to verify results (Newhart et al., 2019).

Tinelli and Juran (2019) presented an automated statistical model and artificial intelligence supported algorithms that allow early detection of anomalies in data measured at the physical-chemical enhanced primary clarifier of a WWTP plant. The developed algorithms exploit the concept of expert pattern recognition of deviations from normal conditions and assess the probability of an event or an anomaly, filtering false alarms. The study evaluated the removal rates of organic matter and suspended solids and the coagulant reagents used in a chemically enhanced primary settler and provided a risk-based classification of potential anomalies applying Supported Vector Machines (SVM) and Artificial Neural Networks (ANN). The SVM showed a classification error of 2% and of 4% using the ANN. The authors presented a method for anomaly detection and risk-based classification, as well as the ability to isolate anomalies according to a risk severity scale (Tinelli and Juran, 2019).

2.8.1 Training and Testing of Data-Driven Models

Similar to mechanistic models, Machine Learning models within the context of Data-Driven Models (DDMs) require training and validation of the model as well as testing the model's performance on a new dataset.

The general procedure of training a Machine Learning model starts with data collection and data preprocessing to remove faults and outliers, as well as fill data gaps. Normalization and standardization are commonly applied data management techniques that avoid numerical overflows. The second main step is constructing and training the model, which includes a selection of the model's architecture, training algorithm, activation function and model hyperparameters. During the third step, the model is tested and optimized based on performance criteria. The final step is the selection of the parameters that provide the best-fit for the trained and tested model (Shirkoohi, 2022).

DDMs can be categorized based on their training (or learning) algorithms; unsupervised, supervised or reinforcement learning. For methods based on supervised learning, including classification and regression algorithms, the mapping function is learnt from input and output data and their associations (Razavi 2021). With unsupervised learning, such as clustering and association, a machine learning model learns from unlabelled input data and discovers common features to identify data groups with similar input patterns (Abrahart et al. 2008; Borzooei et al. 2020). With this learning approach, there is no labelled response data or any predetermined (in-)correct answer. With reinforcement learning, the model learns from the data using trial and error or through a reward-punishment process (Sutton & Barto 2018). It requires less data, since the model is only trained with correct or incorrect responses (Shirkoohi, 2022).

The performance of the data-driven models depends heavily on the quality and quantity of the input data (since no domain knowledge is included), and relies on the learning methods applied. Therefore, the interpretability of data-driven models is a critical bottleneck (Schneider et al., 2022). Also, due to the non-stationary characteristics and significant temporal variability of municipal wastewater, frequent retraining of data-driven models is recommended (Torfs et al., 2022).

2.9 Hybrid Process Models

Hybridization is the process that combines a mechanistic model, which incorporates relevant process knowledge, usually chemical and physical phenomena, such as mass and energy balances, with a data-based model that allows increasing the predictive power by including information about lesser-known sub-processes. Hybrid modelling is able to combine the advantages of the two approaches by linking mechanistic models, which are based on our knowledge of the processes, with data-based models (Von Stosch et al., 2014).

The first research to combine these two approaches started in the early 1990's, where the hybrid models consisted of a combination of either ANN's or radial basis functions with mathematical expressions describing mass and energy balances by respectively data-driven and mechanistic components. The original intention of hybrid models was to attribute information absent from mechanistic models using available (measurable) data.

Recent bibliometric analyses (e.g. Zhao et al., 2020) and reviews (e.g. Corominas et al., 2018) indicate the progress of academic studies applying an artificial intelligence-based data-driven approach to improve the reliability of wastewater systems, increase water quality and lower energy consumption.

Recently, a comprehensive text mining analysis by Schneider et al. (2022) revealed that 534 publications studied the use of hybrid models in wastewater treatment during the period 1990 - 2022, of which more than 50 papers in 2021 (Schneider et al., 2022).

As well explained by Schneider et al. (2022), hybrid modelling is emerging as a useful research tool that allows combining the benefits of knowledge-based models with data-driven models. As indicated in Figure 2-5, in a hybrid model, the requirements to understand all the physical processes happening within a system are lower, such that the data-based part will be able to compensate for unknown or missed dynamics. Since both modelling techniques interact to some degree in a hybrid model, the transparency of the model calculations and the interpretability of the model predictions remain of a higher level than with data-driven models only, which are important criteria of the acceptance and credibility of the model's performance. The main advantage is a higher benefit/cost ratio for solving complex problems, which is a key factor for process systems engineering (Schneider et al., 2022).

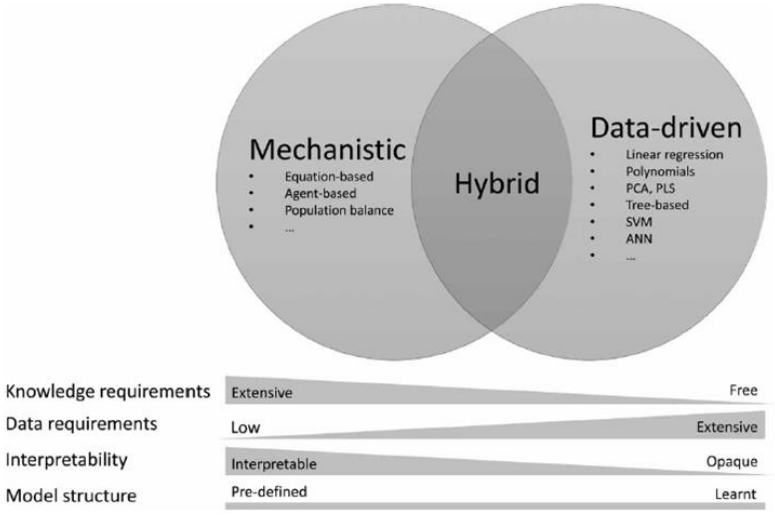


Figure 2-5: Venn diagram of hybrid modelling, schematically showing the requirements of expert knowledge and data, the interpretability, and the model structure. (Schneider et al., 2022)

Research on hybrid models for wastewater applications in particular have focused on generation of influent data, on process modelling, and on process monitoring and control (Schneider et al., 2022).

An influent generator (IG) is a model, aiming to provide the flowrate and pollutant concentration dynamics at the inlet of a WRRF for a range of modelling applications. Hybrid models have been developed that use machine learning methods trained on historical data (e.g.: Ahnert et al., 2016; Zhu et al., 2018a). Li and Vanrolleghem (2022) propose a data-driven IG model only using routine data and weather information, and without need for any additional data collection. The model was constructed by an artificial neural network (ANN) and completed with a multivariate regression to generate time series for certain pollutants. The model is able to generate flowrate and quality data (TSS, COD, and nutrients) at different time scales and resolutions (daily or hourly), depending on various user objectives. The model performance was analysed by a series of statistical criteria. It was shown that the model can generate a very reliable dataset for different model applications (Li and Vanrolleghem, 2022).

Hybrid models have been developed for enhancing the performance of a mechanistic model (e.g.: Anderson et al., 2000; Lee et al., 2005b) and for model parameter identification (Villez et al., 2019).

Hybrid models for process monitoring and process control have been described for time-series prediction (Fernández et al. 2009; Zhu et al. 2018a) and fault/anomaly detection (Wade et al. 2005; Haimi et al. 2016; Cheng et al. 2019). In these roles, the hybrid models provide support to other automated systems as soft-sensors making predictions of water quality variables based on other measured variables for influent quality (eg: Wang et al., 2019; Zhu & Anderson, 2019) and effluent quality (Shi & Xu, 2018; Niu et al., 2020).

2.9.1 Hybrid Model Architectures

Different architectures are applicable for combining mechanistic and data-driven components within a hybrid model (Figure 2-6). Serial and parallel structures are the two most commonly developed architectures where the two components collaborate to increase performance (Lee et al., 2005b).

In the serial structure (top of Figure), the output of one model is used as input to the second model with the objective to improve the second model's performance. The serial hybrid model is specifically adaptable to those cases where a specific sub process is not defined well. Commonly, the output of the data-driven model is used as input of the mechanistic model. The data-driven component represents the dynamics that are not explicitly modelled by the mechanistic component, such as ill-defined or poorly understood reaction kinetics. An early example is a work done by Psichogios & Ungar (1992), where a serial hybrid model showed higher accuracy than either a mechanistic or data-driven model alone. With the inverse approach, the output of a mechanistic model is used by a data-driven model, with the objective to support the features of a data-driven model with domain knowledge, directly feeding the data-driven model with relevant data transformations. A relevant example of this is the representation of inter-connected processes at different spatial or temporal scales, where a multiscale approach utilises both mechanistic representation for dynamic simulation and data-driven methods for higher-scale inference of process behaviour (Li et al. 2019; Hannaford et al. 2021). Tsen et al. (1996) show that this architecture outperforms other architectures and the mechanistic model itself. Nevertheless, this approach has not yet found much application in the wastewater sector in particular due to its lower degree of interpretability (Schneider et al., 2022).

The parallel architecture (shown in the middle part of Figure 2-6) is in particular useful when a mechanistic model is limited in describing the dynamics due to incomplete domain knowledge or oversimplification of the processes. In this approach, the outputs of both models are combined to produce a modified and improved output. In a first variant, a cooperative parallel hybrid model, the data-driven model is trained to learn the mismatch between the mechanistic model and historical data, defined as the residual error. Then both results are merged, usually by addition. A recently described alternative in this approach is to learn the dynamics' residuals instead of the state's residuals (Quaghebeur et al. 2022). In a second variant, a competitive parallel hybrid model, the mechanistic and data-driven models are trained to make the same prediction. Each component solves the same problem independently to improve the hybrid model performance (Hu et al. 2011; Abba et al. 2020). These predictions are then weighted and combined in a final output (Dors et al. 1995; Peres et al. 2001; Galvanauskas et al. 2004; Ghosh et al. 2019).

The cooperative parallel hybrid model structure is in particular relevant in the context of improving mechanistic process models of complex treatment systems, such as biofiltration. These process models are based on process knowledge, but due to the high degree of complexity of the model, their development and parameter estimation as part of the calibration process is much more arduous, time-

and data-consuming; such that it can become a constraint limiting the application (Vanrolleghem et al., 2005). In addition, these first-principles biofilm models require high computing power, which makes them less suitable for real-time control applications with a short time action horizon (Schneider et al., 2022). The data-driven component in a hybrid model uses statistical and databased methods with machine learning models to find unknown relationships in the data. The DDMs are very fast in computational time and have strong interpolation capabilities (Newhart et al., 2019).

A different type of hybrid model shown in the lower part of Figure 2-6, is surrogate modelling (Razavi et al. 2012), also known as metamodels (Jin et al. 2001), substitute models (Romijn et al. 2008), emulators (Conti & Hagan 2010; Mahmoodian et al. 2018), or response surface models (Wang & Georgakis 2019). These models are trained on the output of a large (and heavy) model to create a computationally less demanding (reduced) model for rapid or large-scale simulations. Once developed and calibrated or trained, the surrogate can then autonomously be applied for real-time model predictive control, uncertainty analysis, or optimisation problems. Although surrogate models could be of any model type (mechanistic, statistical, deterministic, empirical, etc.) that computes faster than the original model, often data-driven models are applied for their speed.

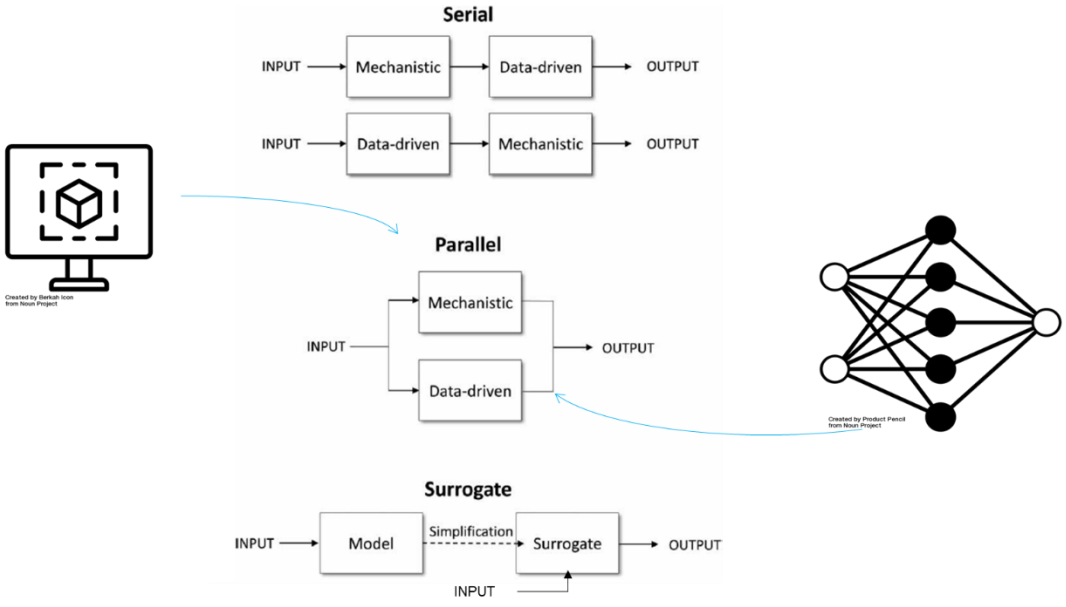


Figure 2-6: Schematics of the three model architectures used in Hybrid Modelling, based on Schneider et al. (2022). The Surrogate model architecture was slightly modified to include a data input flow during 'normal' (non-training) operation.

2.9.2 Hybrid Model Predictive Control

The research field of hybrid model predictive control (HMPC) is a non-fully developed domain with many open challenges. The concept of a Hybrid Model Predictive Controller (HMPC) in the context of this research focuses on the 'model' component of a MPC, that is to say that the model is a hybrid system consisting of a mechanistic model (MM) supported by a data-driven model (DDM). The MM here simulates process behaviour and makes predictions of variables based on 'First-Principles', the process knowledge acquired. The DDM is there to support the MM and correct the mechanistically modelled output variables in order to reduce the prediction error and thus increase the overall model

performance inside the MPC structure. The control actions will therefore be based on modelled variables that have a higher level of precision. This is expected to result in a more efficient process control. As stated by Newhart et al. (2019), a MPC that uses a mechanistic model has a tendency to have a poor accuracy in fault detection, but combined with statistical process control may be a superior problem-solving approach. A review of the literature found that there are not many academic studies on the topic of HMPC as defined here and that the term 'hybrid' can refer to many different concepts. In most cases, references to hybrid systems in the context of HMPC, discuss dynamic systems with both continuous-state and discrete-state and event variables. The physical system has time-driven and event-driven dynamics, the controller affects both time-driven and event-driven components, and it may deal with continuous and/or discrete signals (Camacho et al.,2010). It principally refers to the implementation of MPC in hybrid processes.

Giorgetti et al. (2007) attempt to find the optimal control law of a constrained system by utilizing model predictive control (MPC) built around hybrid dynamical systems, an application they named 'hybrid model predictive control', that aims to control multiple systems by simulations using mixed-integer quadratic programming. Oberdieck and Pistikopoulos (2015) present an approach for an explicit hybrid model predictive controller, in particular using multiparametric programming. The authors define a hybrid system as a system that features continuous as well as discrete dynamics, and the online optimization step corresponds to a mixed-integer quadratic programming.

Bemporad et al. (2000) describe the development of optimal controllers for hybrid systems. They elaborate on a procedure for synthesizing piecewise linear optimal controllers for hybrid systems and investigate conditions for closed-loop stability. The authors define mixed logical dynamical (MLD) systems as hybrid systems since these processes evolve according to continuous dynamics, discrete dynamics, and logic rules (Bemporad et al., 2000).

HMPC concepts have been applied as well in the domain of water management. Van Ekeren et al. (2011) propose a model predictive control (MPC) approach that is aimed at automatically coordinating the different actions related to the opening and closing of water gates for flood protection over a large watershed. In order to reduce the computational effort required to solve the hybrid MPC problem, they authors use time-instant optimization (TIO-MPC).

In the context of wastewater management, Ocampo-Martinez et al. (2007) and Ocampo-Martinez and Puig (2009) describe a Hybrid Model Predictive Control of Sewer Networks, where the hybrid systems considered are mixed logical dynamical (MLD) systems as mentioned above. Their motivation is that by considering hybrid dynamical systems in discrete time, a number of mathematical problems are avoided that allow for tractable analysis and optimal/predictive control. In Ocampo-Martinez and Puig (2009) a HMPC with fault-tolerant capabilities is formulated within the hybrid systems framework that allows to adapt the system model on-line by taking into account the fault information provided by a fault diagnosis and isolation module. In this way, the controller can cope with the considered faults (Ocampo-Martinez and Puig, 2009).

Bartman et al. (2009) describe a MPC control system, where a dynamic nonlinear lumped-parameter model is derived using first-principles, and its parameters are computed from experimental data to minimize the error between model predictions and observations. This model then is used as the basis for the design of a nonlinear control system. The control system is implemented on an experimental reverse-osmosis membrane water desalination system, and its set-point tracking and disturbance

rejection capabilities are evaluated. In general, the system's performance was more satisfactory with the dynamic model than with a conventional controller.

Since in wastewater treatment, the process has a nonlinear behaviour, the model in the MPC must be able to account for this. However, the use of nonlinear models is computationally intensive to solve, and accounting for too many nonlinearities can substantially slow a controller's response.

Alternative methods of MPC discussed in Newhart et al. (2019) are the application of a piecewise linear MPC where multiple linear models provide an approximation of the nonlinearity (e.g. Ocampo-Martinez, 2010), and to use adaptive MPCs, where linear model parameters are adapted to fit changing conditions in the operations (e.g. Zhang and Zhang, 2006). Such adaptive MPC controllers have shown a better performance when compared to conventional PID controllers when faced with non-linear processes. However, Hermansson and Syafiie (2015) found that strong dynamic and nonlinearities are still better controlled using a neural network (Newhart et al., 2019).

In conclusion, it can be stated that hybrid models enable more comprehensive WWTP automation with real-time control as they apply fast-computing data-driven models. As the robustness of techniques used at each step grow, it will become possible to integrate them into larger systems such as digital twins that will enable the production of actionable information for process control and optimization in (near) real-time (Therrien et al., 2019). Research into the application of hybrid models is only expected to increase as the complexity of mechanistic models will continue to rise to reflect the more complex and innovative treatment processes being developed. In addition, the trends in digitalisation of the water sector will lead to a wider acceptance of data-driven approaches (Corominas et al, 2018). The number of hybrid models being applied in real-world cases is expected to increase in the future (Sin and Al., 2021).

2.10 Concluding Remarks

Mathematical modelling is an important scientific tool used for a better understanding of biological wastewater treatment processes. Models allow performing system analysis and predicting performances under a wide range of scenarios. The use of process models has become a widely accepted practice in the wastewater sector for the design, the analysis and monitoring of a system's performance. Unlike the ASM mechanistic models (Henze et al., 2000) that describe suspended biomass growth processes found in activated sludge processes and that have a solid base within the scientific community, mechanistic models simulating the biofiltration process, which is based on a fixed biomass growth inside a biofilm, are still in full development. The biokinetics equations considered in the ASM cannot be directly applied in a biofiltration model due to the importance of the transfer processes by diffusion, the biological kinetics that take place in the biofilm and the effect of regular backwashing. Additionally, there are no widely accepted protocols to aid the community in the development of biofilm models (Rittman et al., 2018).

Despite the successful applications of mechanistic wastewater process models that are based on process knowledge, there are important limitations; these models are built on assumptions and simplifications of the complex microbial processes taking place, such that several underlying processes remain incomplete in the model; they require laborious model parameter calibration and model validation, that need a lot of data and time; and do not incorporate novel treatment techniques (Schneider et al., 2022). Furthermore, the greater complexity of mechanistic models makes their development and calibration much more time- and data-consuming; such that it can become a

constraint limiting the application (Vanrolleghem et al., 2005). These biofilm models require high computing power, which makes them less suitable for real-time control applications with a short time action horizon (Schneider et al., 2022).

Quite differently, data-driven models use statistical and data-based methods with machine learning models to find unknown relationships in the data and are very fast in terms of computation. These models describe the system solely based on information extracted from the data and have strong interpolation capabilities (Newhart et al., 2019). Nevertheless, due to the non-stationary characteristics and significant temporal variability of municipal wastewater, frequent retraining of this type of models is recommended (Torfs et al., 2022).

Hybrid modelling is able to combine the advantages of the two approaches by linking mechanistic models, which are based on the knowledge of the processes, with statistical data-based models (Von Stosch et al., 2014). The main advantage is a higher benefit/cost ratio for solving complex problems, which is a key factor for process systems engineering (Schneider et al., 2021).

Recent advancements in data-driven process control and performance analysis could provide the wastewater treatment sector with an opportunity to reduce costs and improve operations; and to reconcile two partially conflicting objectives of pollutants removal and energy conservation. Big-data integration at the operational level of WWTPs will have the most substantial impact on process control. However, the utilisation of big-data in operations is not yet commonly applied, in part due to the lack of data science knowledge as well as the voluminous size and complexity of datasets (Newhart et al., 2019).

All types of models are sensitive to the quality of the data provided. This highlights the importance of the quality of data represented by the "data-pipeline", a process that describes the necessary steps to turn raw data into useful information in order to make intelligent decisions. The data must be of high quality in order to derive useful information from it (Therrien et al., 2020). High frequency data (e.g. 15-minute intervals) from in-line sensors in harsh wastewater treatment conditions in particular, are exposed to larger variations and show signs of drift, bias and errors. A data preprocessing procedure based on the detection of outliers, smoothing and faulty data will reduce the impact of measurement errors (Alferes & Vanrolleghem, 2016).

Increased data availability in combination with improved computational capacity will continue to shape the structure of future modelling frameworks. The newly-developing synergy between first principles and data-driven models has the potential to create very powerful tools for further innovation, development and decision support. However, balancing the efforts for model development and complexity, data collection, data quality assurance, and integration of different frameworks will be challenging and will require diverse technical skills (Regmi et al., 2019).

The research field of hybrid model predictive control (HMPC) is a non-fully developed domain with many open challenges. A literature review in the context of applying a hybrid model for the ‘M’ in ‘MPC‘, did not reveal actual references for application in water management.

3 Methodology & Materials

This chapter discusses the methodologies and materials used for this research. It follows the main project development phases as described by the International Water Association (see Figure 3-10) that have been applied in this study. The chapter starts with a description of the data collected at the study-site and the methodology used for data cleaning. This is followed by an in-depth description of the mechanistic model that this study started from, the calibration and validation procedures as well as the evaluation criteria that were applied for its future development. The methods used for the selection, training and testing of the data-driven model are then discussed, to conclude with a presentation of the developed structure of the hybrid model and the hybrid model predictive controller.

3.1 Operational Data

3.1.1 Study Site

The data for this project was collected from the biofiltration wastewater treatment lane of the 6 million PE, 1.25 million m^3/day (dry weather) wastewater treatment plant Seine-aval (SIAAP Paris, France) located in Yvelines near the city of Achères, along the Seine River downstream of Paris (Figure 3-1). The Seine-aval (SAV) wastewater treatment plant (WWTP) was selected as the case study for this research since it is has a great potential for optimization, being one of the largest plants in Europe with a rated flow design of 1.7 million m^3/day. The dry weather treatment capacity corresponds to an average daily flow of 14.5 m^3/s, while the maximum flow allowable is 2 300 000 m^3/d (26.6 m^3/s) (Zhu et al., 2018b).

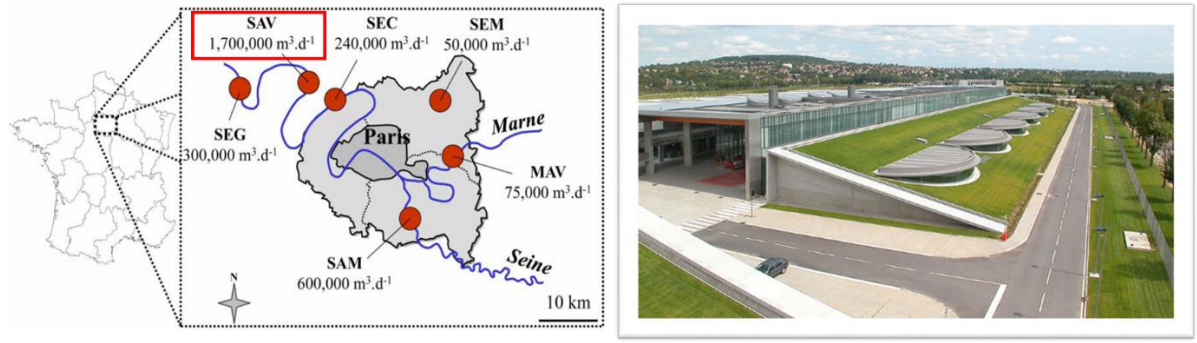


Figure 3-1: Map with locations of SIAAP's WWTPs (left) and indication of the Seine-aval. Picture of the Seine-aval WWTP.

The wastewater treatment chain consists of a pre-treatment stage (screening, grit and oil removal), followed by a primary settling phase enhanced by dosing of coagulants and flocculants. The secondary biological treatment is performed either by biological filtration in a configuration of pre-denitrification (preDN), nitrification (NIT) and post-denitrification (postDN), or by membrane bioreactors (MBR). Figure 3-2 illustrates the Seine-aval WWTP under typical normal dry weather conditions in the current configuration.

Primary treatment is currently performed by 20 conventional settling tanks and 9 enhanced clarifiers ('Actiflo'), where the wastewater is flocculated with micro-sand and a polymer, facilitating the formation of flocs that improve the settling process. This provides better control over the primary effluent quality in order to assure a low concentration of suspended solids in the effluent of the

primary stage. Moreover, the faster settling process allows reducing the tank volumes required significantly, and thus reduces the footprint of the plant.

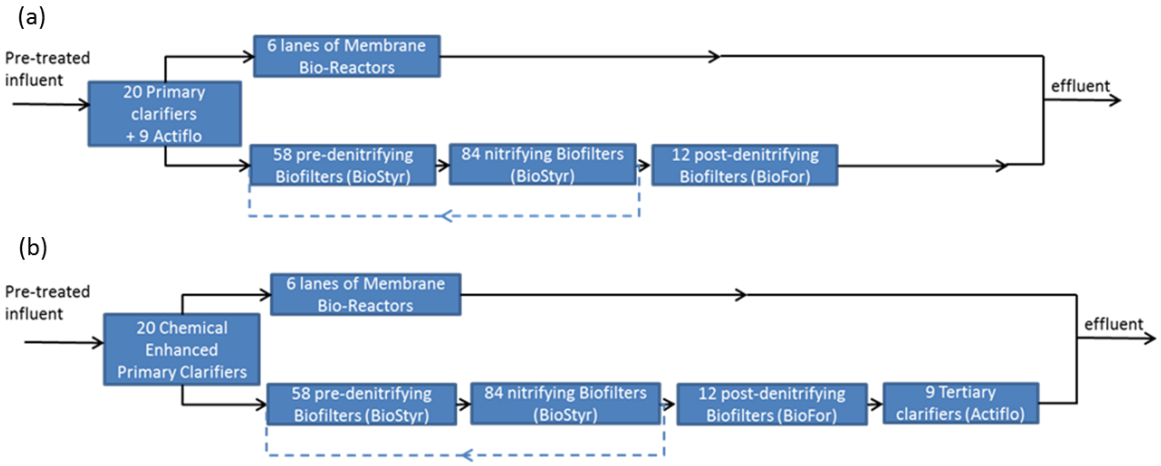


Figure 3-2 : (a) Current configuration of Seine-aval WWTP; (b) configuration after the renovation works have finished.

For secondary treatment, the plant is equipped with two parallel and independent biological processes; the main track is the biofiltration track that treats about 80% of the primary effluent under dry weather conditions. The objective of the biofilters is the removal of nitrogen and organic matter, as well as some phosphorus. This is performed over three stages; a predenitrification (preDN) stage where the bacteria remove nitrates and nitrites using organic matter; a nitrification (NIT) stage where the biomass removes ammonia and creates nitrites and nitrates; and a postdenitrification (postDN) stage to remove the remaining nitrate and nitrite. The treated water is then finally discharged in the Seine river. During intense rainfall events, the flow path is adapted allowing some wastewater to by-pass certain treatment processes if the maximum capacity is surpassed.

Plant Performance

The average daily flow rates and treatment performance at the Seine-aval plant for the period 2019 – 2021 are shown in Table 3-1; with on the top side the concentrations of quality variables in the overall effluent and on the bottom side the regulatory norms. The plant performance is evaluated on daily average concentrations for TSS, BOD_5, COD, TKN, NH_4 and TP, as well as removal rates that need to fall between 70% (for phosphates) and 90% (for TSS). In addition, there are yearly concentration averages for TN (10 mgN/L) and TP (1mgP/L) and removal rates to be adhered to. As prescribed by the authorities, the treatment norms are to be respected under all conditions for flow rates up to the daily reference flow rate (2 300 000 m^3/day) and 45 m^3/sec for a short duration of intense flow rates such as during storms.

Table 3-1: Seine-aval WWTP treatment performance on average for the last three years for the quality variables monitored and their compliance norms.

		Seine-aval WWTP Effluent concentration						
Year	Flowrate	TSS	BOD_5	COD	TKN	NH_4	TN	TP
	m^3/d * 1000	mg/L	mgO_2/L	mgO_2/L	mg/L	mgN/L	mg/L	mg/L
2019	1388	15.3	13.8	54.3	7.1	-	17.5	1.58
2020	1297	18.6	13.9	59.7	6.3	-	18.9	1.15

2021	1438	21.8	12.6	62.7	7.79	-	19.8	1.07
		Seine-aval WWTP Effluent compliance norms						
max conc. (day)		30	20	90	8	5	-	2
min Removal Rate (day)		90%	80%	75%	80%	81%	-	70%
max conc. (year)		-	-	-	-	-	10	1
min Removal Rate (year)		-	-	-	-	-	70%	80%

The Seine-aval WRRF is also a plant where water and energy are recovered and reused for industrial processes. About 50 000 m^3/day of treated effluent from the membrane bioreactor lane is used as an industrial water source for the cleaning and cooling of processes. In addition, the plant generates large volumes of methane gas thanks to the treatment of the produced sludge. This biogas is directly reapplied in the wastewater and sludge treatment.

Biological Activated Filtration

Under normal conditions during dry weather, the biofiltration units treat approx. 80 % of influent. They are equipped with completely submerged upward flow biofilters that contain media that supports fixed-growth conditions for the microorganisms growing on the pollution present in the wastewater. Two types of biofilters are being applied in the WWTP, namely biofilters that are equipped with low-density media (polystyrene beads), type Biostyr©, and biofilters that have argile media in the filter bed, type Biofor© (see Figure 3-3).

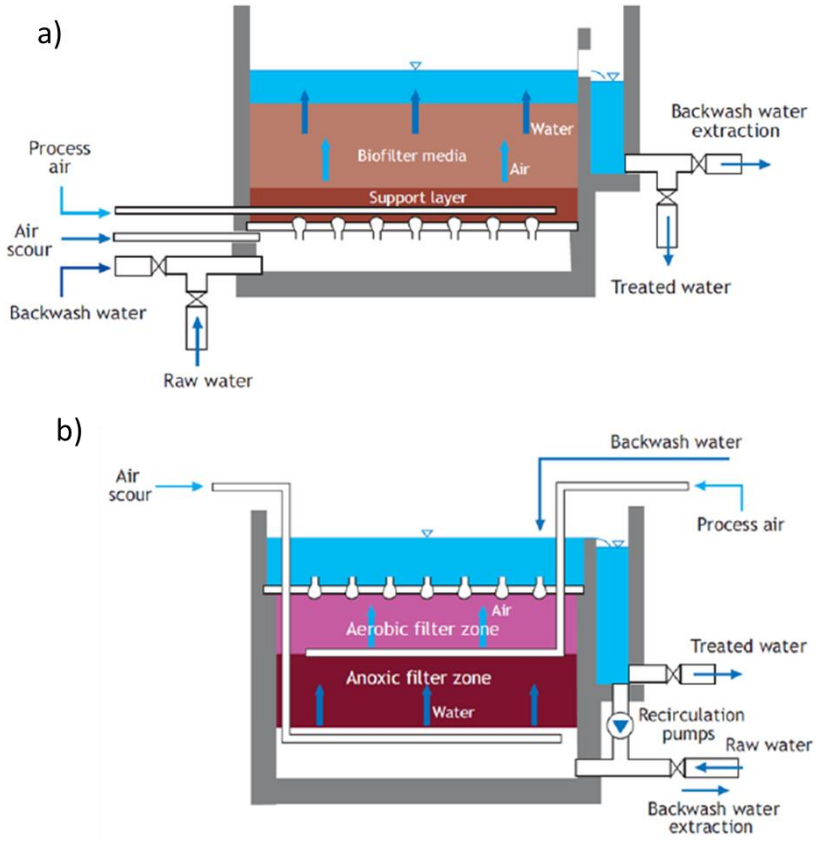


Figure 3-3: Configurations of biofilter type (a) Biofor© and (b) Biostyr© (Morgenroth, 2008).

Regulating the environmental conditions within a biofilter allows promoting the growth of the desired types of bacteria.

Predenitrification (preDN) is concerned with the removal of carbon and transformation of nitrites and nitrates into nitrogen gas under anoxic conditions (low concentrations of dissolved oxygen, high concentrations of dissolved NOx). The preDN stage consists of 58 Biostyr® (Veolia, France) biological filters (173 m^2 * 3.5 m height/biofilter) that contain polystyrene beads, 4.5 mm in diameter, and are divided over three batteries (preDN A, B, C). Denitrification of nitrates (NO_3-N) and nitrites (NO_2-N) happens by heterotrophic (HET) bacteria using the organic matter (OM) present in the settled water. Dosing of external carbon sources (methanol or ethanol) and phosphoric acid is required when the influent organics and phosphorus are insufficient for the biomass to perform. Denitrification treats the effluent from the primary clarifiers containing the necessary organic matter, and the nitrates present in the part of the effluent from the nitrification stage that is recycled (30 – 80%). This recirculation of nitrified water from the NIT outlet to the preDN inlet guarantees the inflow of nitrates and nitrites, which are present in the primary effluent only at very low concentrations. The backwashing water from its biofilters as well as treated water from sludge treatment and dewatering processes is also injected in this preDN stage.

Nitrification (NIT) transforms ammonium (NH_4-N) into nitrates and nitrites under aerobic conditions (sufficient concentration of dissolved oxygen) that necessitates the forced continuous injection of air. This stage consists of 84 Biostyr® (Veolia, France) biological filters (173 m^2 * 3.5 m height/biofilter; with polystyrene beads, 4.0 mm in diameter, and organized in six batteries: NIT A1, A2, B1, B2, C1, C2). It treats the effluent from the preDN stage, as well as the return backwashing water from its own filters. In case of high flow rates, the NIT filters will treat the part of the outflow from the primary clarifiers that has bypassed the pre-denitrification stage. In this way more biological treatment capacity is available under such hydraulic overloads.

In the nitrification stage of the biofiltration, the rate of aeration for each treatment cell is controlled by feedforward and feedback control rules. The feedforward control estimates the air requirements for the treatment cell according to the outlet ammonium set point and the ammonium concentrations measured at the entrance for the cell. The feedback control enhances or reduces the air flow rate as proposed by the feedforward control rules. It does this based on the difference between the measured effluent ammonium concentrations and the ammonium set point.

Postdenitrification (postDN) is the final biofiltration stage and is concerned with the transformation of nitrites and nitrates into nitrogen gas under anoxic conditions. It consists of 12 Biofor® (Degremont, France) biological filters (147 m^2 * 3m height/biofilter; clay granular material of 4.5 mm in diameter; 1 battery) that denitrify the NIT effluent. Here, the addition of an exogenous carbon source in the form of methanol (CH_3OH) and the dosing of phosphoric acid (H_3PO_4) is (almost) always necessary to stimulate biological activity due to the low levels of organic matter and phosphorus still remaining in the water. After treatment, the effluent of the postDN stage is mixed with the effluent from the membrane bioreactors. This causes a further dilution of any remaining pollutants and nutrients, before the treated water is discharged.

The flow path between the different biofiltration stages can be varied according to the influent's flow rate and pollutant concentrations. This makes a total maximum of 150 active filters, given that there

is always one filter in each treatment phase that is momentarily off-line for cleaning or maintenance purposes.

The biological filters are backwashed daily on a fixed schedule by reversing the flow using treated water stored on site ('backwashing') and accompanied by the injection of large air bubbles. During the washing (see Figure 3-4) that lasts for 30 minutes per day per filter, the trapped particles and part of the biomass growth are removed in order to restore the filtration capacity. Additional short backwashing sessions of 10 to 15 minutes can be applied to a filter when the pressure loss over the filter reaches a determined value. The collected backwash water contains high concentrations of biosolids (a mix of particles and released microorganisms) that requires on-site treatment to remove the sludge before it is sent back to the entrance of the biofiltration lane.

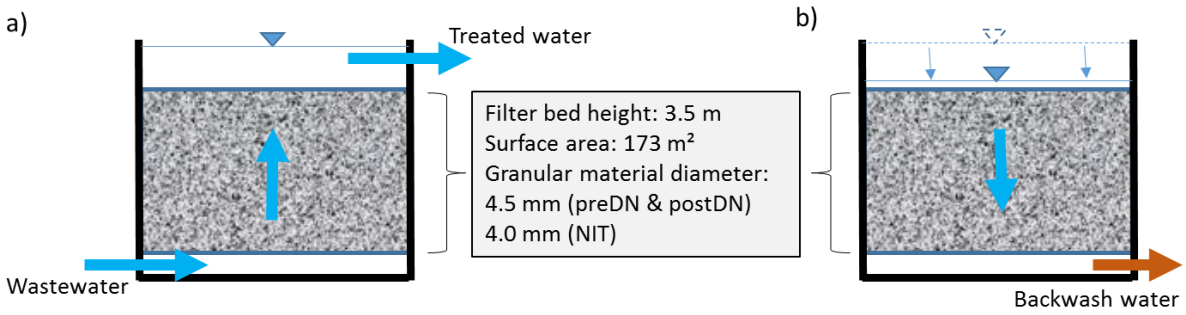


Figure 3-4: Principle of submerged biofiltration during (a) normal operation in upward flow direction, and (b) during filter cleaning where the flow is downward using stored treated water. The backwash water is subsequently treated on site.

Modernization Works

The wastewater and sludge treatment processes at the Seine-aval WWTP have been the subject of an intensive overhaul to achieve a performance that continues to meet the stringent regulations and lowers the environmental impact. The renovation works for the wastewater treatment plant started in 2009 and are planned to be finished by 2024. It includes the construction of new biofiltration units and membrane bioreactors. These processes are more reactive and require a smaller footprint when compared to conventional biological treatment, thus clearing large parts of the terrain previously occupied by the plant operations.

A principal component of the modernization, the upgrade of the biological treatment processes – a project named BioSAV – had as objective the improvement of the performance of biological treatment by integrating pre-denitrifying biofilters and membrane bioreactors. The new pre-denitrifying biofilters take advantage of the organic matter present in the wastewater and thereby reduce the need of external carbon dosing. These filter units were installed in 2017 and became fully operational by mid-2018. Furthermore, the membrane bioreactors require lower dosing of chemical reagents, if supplied with properly treated primary effluent. Overall, these new systems are expected to limit the consumption of reagents, lower the environmental impact and have an economic advantage.

Currently, the renovation of primary treatment is underway. The works include the commissioning of 20 new physical-chemical lamellar clarifiers (Densadeg®, Degremont Technologies) that will replace the activated sludge treatment systems (Figure 3-2b). These physico-chemical lamellar settling tanks can remove larger amounts of suspended solids (TSS) and phosphorus thanks to the increased surface area provided by the lamellar inclined tubes, as well as by dosing iron chloride and anionic polymer as

coagulant and flocculent reagents. The construction works on these chemically enhanced settlers are planned to be finished by 2024 and become operational soon afterwards. The 9 existing enhanced clarifiers will be changing role as part of a tertiary treatment phase, becoming responsible for final treatment of the effluent of the secondary biological treatment, removing any remaining suspended solids and nutrients.

3.1.2 Data Collection

For this research, data was collected from December 1st, 2019, till September 31st, 2020. The start of this data range coincides with the installation of in-line probes for COD and nitrate measurements at the entry and exit of the predenitrification stage in the Seine-aval WWTP. These sensors became fully operational on November 26th, 2019. The mechanistic biofilter model is calibrated on the first three and a half months available (01.12.2019 – 15.03.2020) which covers 106 days, while for the validation the data from 01.06.2020 – 30.09.2020 (122 days) is used. The period in between, from March 16th till end of April does not contain coherent and consistent data due to the Covid-19 lockdown.

Figure 3-5 indicates the locations at the Seine-aval WWTP biofiltration lane of the sampling points for the laboratory analyses as well as the location of the inline analysers and sensors for water quality and flow rates. As can be seen, most sampling points are located at the main entrance channels to the preDN, NIT and postDN stages, thus before the flows are distributed over the various batteries. Only for the preDN stage, there is actually one battery (preDN battery C) that is monitored both at its entry and exit for nitrogen and COD.

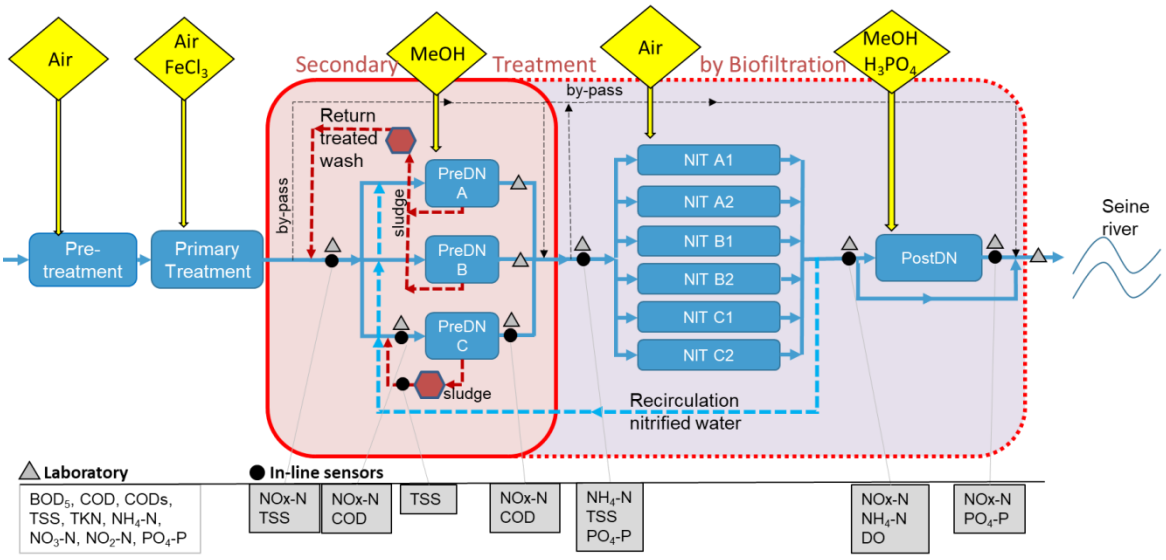


Figure 3-5: Schematic diagram of the biofiltration wastewater treatment lane at the Seine-aval WWTP with indication of processes (blue boxes) and main dosing actions (yellow diamond shapes). Outlined in the red dotted frame are the secondary treatment predenitrification biofilter units (preDN), the nitrification battery units (NIT) and postdenitrification units (postDN). The focus of this study is on the preDN stage. Also indicated are the sludge treatment units for preDN (brown hexagons) and the location of in-line sensors (black dots) and sample locations for laboratory analyses (grey triangles).

Considering nitrogen, the inline analysers are located at the entry of the preDN lane before the recirculation of nitrified water is mixed with it. These sensors actually monitor the combined compounds of nitrate and nitrate (NO_2-N + NO_3-N = mg NO_x-N/l) in the water.

Also, not all streams are actively monitored, but information can be obtained from analyses performed at other locations. An example is the recirculation flow of the effluent of the NIT stage that is not actively monitored. To retrieve the concentration of NOx (mg NO_x-N/l) in the nitrified recirculation lane, which is an important water quality variable for the process control of the preDN, it can be assumed that the quality of the influent water at the entry to the postDN is identical to the effluent quality of the NIT stage.

Database

Water quality and quantity measurement data from inline sensors, analysers as well as daily laboratory analyses and other observational data (e.g., maintenance reports) from the Seine-aval WWTP were accessed in the 'BaSta Virtuel' database that allows visualizing and exporting data in tables and spreadsheets (Figure 3-6). This secure database that is managed by SIAAP also contains for each sampling point essential meta-data, such as information concerning the sampling frequency, measurement units, responsible laboratory staff and dates of installation.

Besides water quality and quantity measurements, process control data is registered in the database. These are data related to the number of filters in operation at each 15 minutes time step, the different set points for nutrient removal (e.g. for nitrate concentration in the effluent of the preDN and postDN, the ammonia set point in the effluent of the NIT stage), as well as information about the compressors used for aeration and the distribution of the water flow (number of hours; energy usage; etc.), the duration of backwashing cycles and the quantity of reagents dosed.

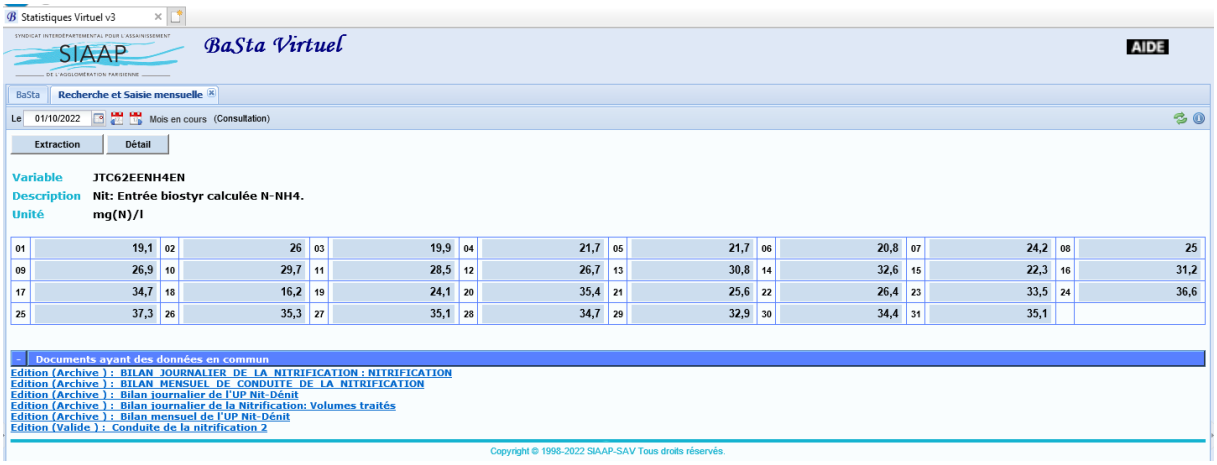


Figure 3-6: Print-screen of the BaSta Virtuel database that collects the observations made at the wastewater and sludge treatment lanes at the Seine-aval WWTP.

3.1.3 Data Pre-processing

Data pre-processing of the raw measurement data is necessary to remove bias, together with faulty data and outliers, as well as to perform gap filling. This allows obtaining a complete set of reliable and consistent data. Clean data is an important factor in modelling, to make sure it is accurate, complete, consistent and trustworthy. It avoids errors from affecting the model's performance. Due to the tough measurements conditions typically present in wastewater treatment processes, sensors in WWTPs are prone to malfunctions. Sensors are known to be affected by bias, drift, precision degradation or total failure effects that cause the reliability of measurements to decrease. The reconstruction of sensor data can yield a better monitoring efficiency (Yoo et al., 2008).

A preliminary step therein is the visualization of data and performing plausibility checks based on (mass) balances of components, i.e. the mass at the entrance of the treatment equals the combined total of components leaving the treatment by effluent water and sludge.

Alferes and Vanrolleghem (2016) proposed a method for automatic data quality assessment extracting information from individual water quality time series from on-line sensors. Data mining techniques based on forecasting models are used to detect and remove unreliable data from the raw data sets. A posteriori analysis is applied to remove noise and detect abnormal situations and potential sensor faults. The authors have successfully tested their method on water quality time series collected from different water and wastewater systems (Alferes and Vanrolleghem, 2016).

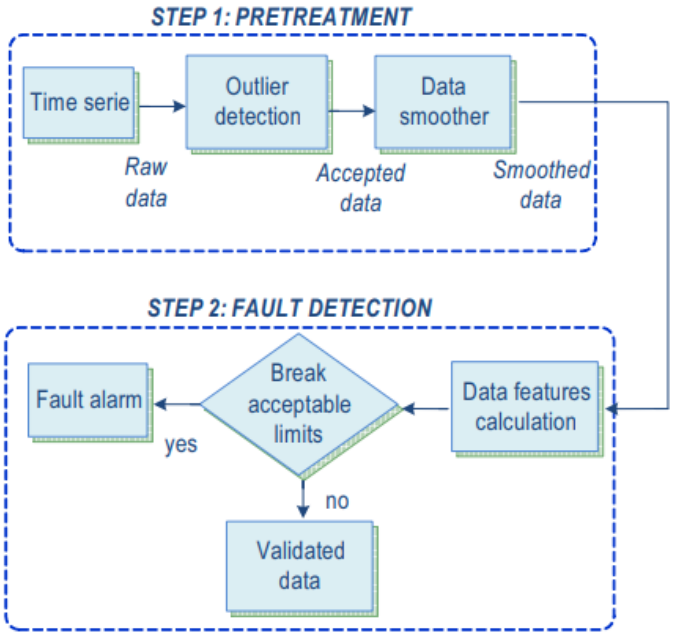


Figure 3-7: Univariate time-series analysis as proposed by Alferes & Vanrolleghem (2016) that comprises a data pretreatment step to identify outliers and a fault detection step that cleans data (Alferes & Vanrolleghem, 2016).

The methodology applied for data cleaning in this study is the method described by Alferes and Vanrolleghem (2016) illustrated in Figure 3-7, which is based on statistical and forecasting methods. It includes two main steps: (1) identification of unreliable data points, and (2) fault detection. In the first step, forecasting techniques are applied to predict future expected time series data by using the historical information contained in on-line water quality time series. Unreliable data are identified as outliers by comparing a new measured value with its forecasted value and the associated dynamic prediction acceptability interval. These outliers are subsequently removed from the time series. In the second step that aims at fault detection, several statistical data features and their acceptability limits are calculated.

3.2 Mechanistic Biofiltration Modelling Strategy

3.2.1 Biofilter Model

The mechanistic model applied here is a 1-dimensional dynamic biofilm reactor model describing a 'fixed-culture' submerged upward flow biological filter. The model was previously developed (Zhu et al., 2019; Zhu, 2020) based on an earlier model (Bernier et al., 2014) that uses ASM1 biokinetics (Henze et al., 2000). It consists of the combination of sub-models for the hydraulic representation of the filter

media, the transfer and transport of soluble and particulate substrates to/from and within the biofilm, the kinetics of bacterial growth and decay, and backwashing to avoid clogging of the filter. The model simulates the biological and physicochemical conversion of carbon, nitrogen and phosphorus. In total, 22 biological conversions are accounted for by describing the evolution of substrates by a mass balance.

The performance of biofilters is directly related to the filtration of suspended particles and the biological activity of purifying biomasses (Henze et al., 2008). Figure 3-8 shows the approximation of the vertical mixing of a filter into seven successive completely mixed reactors. The volume of each mixed reactor is divided between the biofilm compartment (in orange), a bulk liquid phase (blue), a gaseous phase (grey) and the inert media (white). The Mass Transfer Boundary Layer (yellow) simulates the transfer resistance of soluble components from the liquid phase to the biofilm by diffusion, which is a representation of the stagnant liquid film between the biofilm and the liquid phase (Boltz et al., 2011). To better simulate the impact of this boundary layer, the model considers a variable thickness of the boundary layer as a function of the media diameter and the bulk liquid's Sherwood number based on the equation of Ohasi et al. (1981). This allows simulating with more precision the impact of different hydraulic conditions near the surface of the biofilm that directly affect the diffusion resistance (Zhu, 2020).

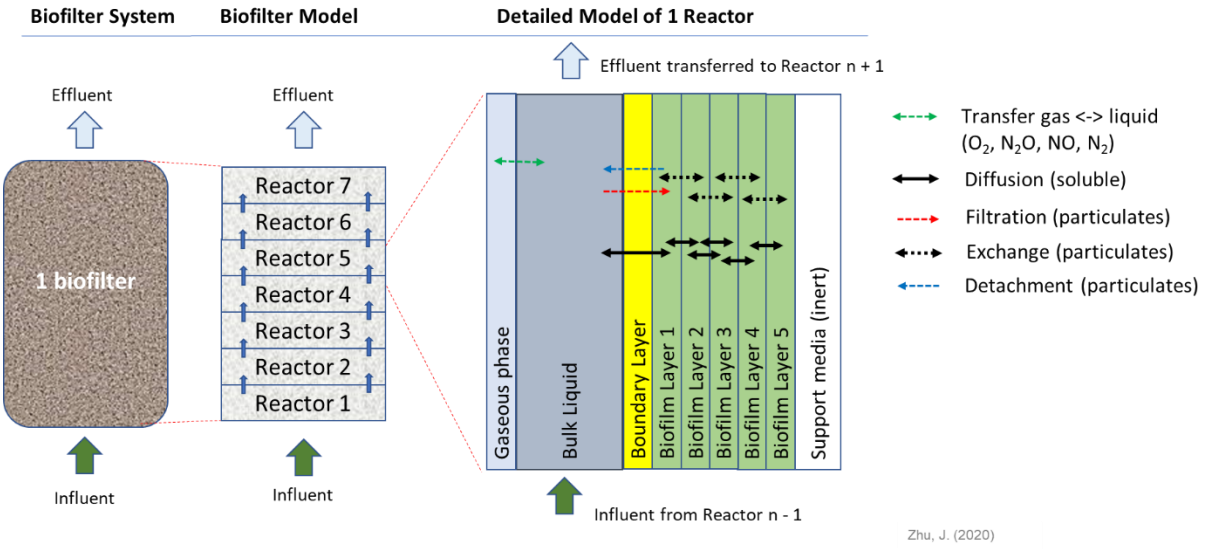


Figure 3-8: General concept of the biofiltration model developed by Zhu (2020) that simulates the behaviour of 1 biofilter, divided over seven vertical reactors describing the vertical spatial variations in the biofilter. Each reactor has compartments for gas, bulk liquid, the Mass Transfer Boundary Layer and the biofilm; the latter divided into 5 layers to describe the spatial variation of components in the biofilm. The horizontal arrows indicate the main physical phenomena modelled.

As illustrated in Figure 3-8, the biofilm compartment (symbolized by orange-coloured vertical rectangles in the figure) is split into five homogeneous layers, a number applied previously by Vigne et al. (2010) that was found to not take too much computational time and still accurately simulate the concentration gradients of soluble substrate inside the biofilm (Zhu, 2020). The model assumes a dynamic biofilm volume with a variable thickness of each biofilm layer, based on the mass of particulates present in each layer, the surface area of liquid to media and a constant biofilm density.

The physical processes considered in the model (the arrows in Figure 3-8) in each reactor are the transfer to and transport within the biofilm of soluble components by diffusion, the transfer and

transport of particulate components by filtration, attachment and detachment and exchange between the layers of the biofilm; as well as a flux of four gaseous components (O_2, N_2, NO, N_2O). These gazes are considered as soluble components in the biofilm layers and transported by diffusion (Zhu, 2020). In addition, the model redistributes the particulate components between the successive mixed reactors during treatment and especially during backwashing by representing continuous media mixing at different intensities for normal operation and backwashing. The model allows estimating energy consumption related to aeration and pumping, as well as greenhouse gas emissions (Zhu et al., 2018). For a more detailed description of the mechanistic model, see Zhu (2020).

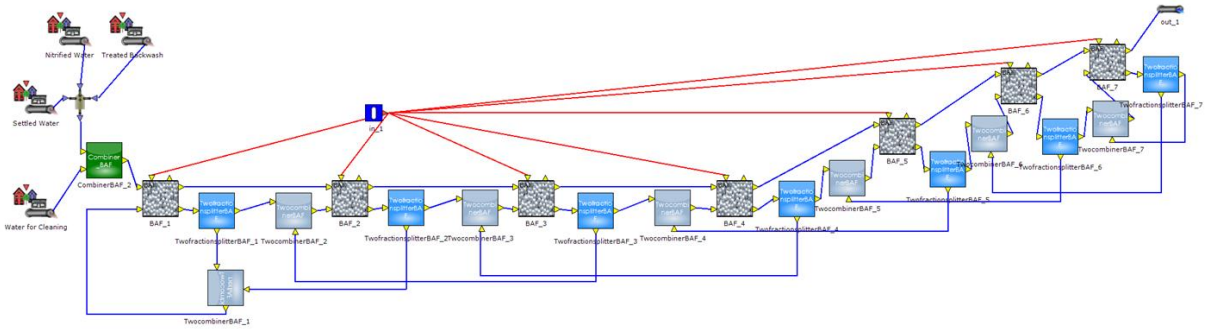


Figure 3-9: WEST layout of the biofilter model consisting of a biofilter divided into 7 reactors (BAF_1…BAF_7) with continuous back-mixing (blue recirculation lines) of part of the filter bed between the reactors. Flow direction is from left to right. On top left are shown the different types of influent water quality taken into account by the model (Zhu, 2020).

Figure 3-9 illustrates the configuration of the biofilter model with one filter divided over seven bioreactors (indicated by the boxes containing the polystyrene beads), each one represents 1/7 of the total column height (3.5 m / 7 = 0.5 m). On the left-hand side of the figure are shown the four different influent flows that are considered by the model. These are the effluent of the primary settlers (settled water), the recirculation of the effluent of the NIT phase (nitrified water), the treated backwash water delivered during filter cleaning (treated backwash) and the water used for cleaning the filters (water for cleaning). The flow direction in the model is thus from left (BAF_1) to the right side (BAF_7), with the final effluent monitored by the out_1 icon.

As mentioned, the model allows for the possibility of a continuous back-mixing of part of the biofilm-support beads. However, the mixing occurs only with the filter bed of the reactor directly above and below it. This is visible in the layout of the figure by the blue arrows to the side and underneath the bioreactors. The red lines above indicate control rules and variables that affect process performance (e.g. temperature profiles), which are defined in an input text file accessible through the smaller size blue coloured box.

Although the biofilter model simulates the behaviour of one biofilter only, it can be applied to simulate the behaviour of a group of filters (battery) and a group of batteries (a treatment stage). In order to do so, Zhu (2020) equipped the model with an integrated converter at the entrance of the first reactor that distributes the incoming flow over the (time-varying) number of active biofilters. This number has to be supplied by the user through an input file. In addition, the model calculates a ‘best’ number of active filters based on the flow rate and the pollutant load.

In the biofilter model the controller that activates the necessary number of filters by comparing two calculated numbers and selecting the largest of these. The controller first calculates the minimum number of biofilters to be activated based on the influent flow rate at the preDN stage and the

maximum filtration rate per biofilter predefined by the model user. Then, the second minimum number of biofilters to be activated is calculated according to the maximum NOx-N load per biofilter predefined by the user and the NOx-N load observed at the influent of the stage. Finally, the number of active biofilters is set to the greater of these two minimum numbers. These comparisons take place at each time step in the model (Zhu, 2020).

Finally, the biofilter model developed by Zhu (2020) was used to create a virtual representation of the whole biofiltration lane at the Seine-aval WWTP by placing three filters in series and adding the process controllers for treatment in the predenitrification, nitrification and postdenitrification stages. This integrated model contains submodules to control the dosing of methanol and phosphoric acid in preDN and postDN, the aeration in the NIT, as well as all possible by-passes.

3.2.2 Influent Files for Biofilter Model

The model requires several influent files to be prepared in advance to characterise the quality of the influent to the treatment stage. However, the model requires concentrations of nutrients and pollutants to be in the format of the model's 'state variables', which are the variables on which the ASM family models are based (see chapter 2 for more information). The data to be supplied can either be the directly measured concentrations of organic matter, nitrogen and phosphorus compounds, which, once they are submitted to the model, can be fractionated by the model based on the scheme and parameter values provided by the user. Alternatively, the user can calculate the state variable values him-/herself and supply these as model input. Either way, the user has to decide on a fractionation scheme and the values of the fractionation parameters used.

The selected fractionation scheme and the parameters are those that were developed and applied previously by Zhu (2020). The scheme remains valid as long as none of the state variables are modified. The parameters have been checked with literature and found to be in the range of generally accepted values.

For this study, high-frequent measurements taken by inline sensors and analysers are used when available. However, as indicated in Figure 3-5, sensors are not situated in all streams. Laboratory analyses are much wider available and function as valuable back-up since these are considered of higher quality. The composite samples for laboratory analysis are based on automated 24-hour composite samples, starting from 07:00 AM each day. Whenever measured water quality data availability is limited to daily average composite samples from the laboratory, these daily values can be linearly interpolated, where the laboratory data is 'positioned' on the mid-day point, that is at 7:00 PM. Whenever possible, cleaned sensor data are preferable.

Due to the complexity of the operations and the availability of the measurements, the different flows of wastewater that are mixed at the entrance of the preDN are to be inputted individually. This way, in a future extension of this research project with the development of a digital twin, the high-frequency sensor measurements can as much as possible be used to feed the model.

The required influent files for the biofilter model thus include the following:

(1) the settled water. Only laboratory analyses are available for daily average concentrations of BOD_5, total COD, soluble COD, TSS, TKN, NH_4-N, NO_3-N, NO_2-N and PO_4-P. The flowrate is measured per 15-minutes. The most important variables controlling denitrification rates measured at this point are the soluble content of organic matter dissolved in the water (solCOD) and available for the

microorganisms, as well as the TSS concentration, from which the amount of biomass present in the influent stream is calculated. A sensor measurement of the soluble COD concentration in the influent is available at the entrance of preDN Battery C. However, at this point, the water quality is already affected by the mixing with treated backwash water that contains high concentrations of organic matter, and forms no longer a good representation of the soluble COD content of the settled water.

(2) the recirculation of nitrified water. High-frequent sensor data is available for concentrations of NO_x-N (NO_3-N + NO_2-N) and NH_4-N. Daily average laboratory concentrations are available of BOD_5, total COD, soluble COD, TSS, TKN, NH_4-N, NO_3-N, NO_2-N and PO_4-P. The 15 minutes NOx data is multiplied with the daily laboratory ratio NO_3-N to NO_2-N in order to retrieve high frequent nitrate and nitrite influent data.

(3) the retour of treated backwash water. No measurements are available for this flow, except for a sensor monitoring the TSS concentration at the outlet of the settler treating the backwash water, as well as a flow rate of the backwash. The TSS can be used to estimate particulate COD and the biomass present. For the other components, the water quality is assumed to be similar to that of the effluent of the preDN stage, since that water quality is used for the filter cleaning.

(4) the washing water. The water quality of the water used for washing is assumed to be that of the effluent of the preDN stage. However, since the model calculations during the washing period are not considered to be part of the final effluent water quality (since in the physical system the backwashed effluent is treated separately), an average water quality for the influent of the washing water is sufficient.

For more information on the applied influent characterization, see paragraphs 4.3 (Preparation of the Model Input') and 4.4 ('Parameter estimation').

3.2.3 Methodology for Calibration & Validation of Mechanistic Model

A model needs to be calibrated and validated for correct and reliable application (Mannina et al., 2011). The method applied for the calibration and validation of the mechanistic biofilter model was based on the Good Modelling Practice guidelines by the IWA (Rieger et al., 2012) as extended by Mannina et al. (2011). This method has been applied in the research and development of the biofilter model by Bernier (2014) and Zhu (2020).

The Good Modelling Practice combines the key aspects of several modelling guidelines that were prevalent at the time with the aim to develop a unified protocol for the modelling community. The proposed protocol is illustrated in Figure 3-10 and comprises the following steps: (1) Project Definition, (2) Data Collection and Reconciliation, (3) Plant Model Set-up, (4) Calibration and Validation, and (5) Simulation and Result Interpretation (Rieger et al., 2012).

In the project definition step, the objectives of the modelling project are defined, stakeholders and their responsibilities are identified, and budget and schedule constraints are agreed upon. In this study, the project definition (step 1) was made preliminary to the start of the research and the project objectives are discussed in chapter 1.

The second step concerns data collection and reconciliation of data sets for simulation projects. A stepwise procedure to analyse collected data is provided including dedicated methods based on statistical analysis, expert opinion and calculating mass balances.

In this study, the mechanistic biofilter model was calibrated with high-frequent (15 minutes) operational data originating from the Seine-aval WWTP collected from December 2019 to March 2020. The method described by Alferes & Vanrolleghem (2016) was applied for data cleaning.

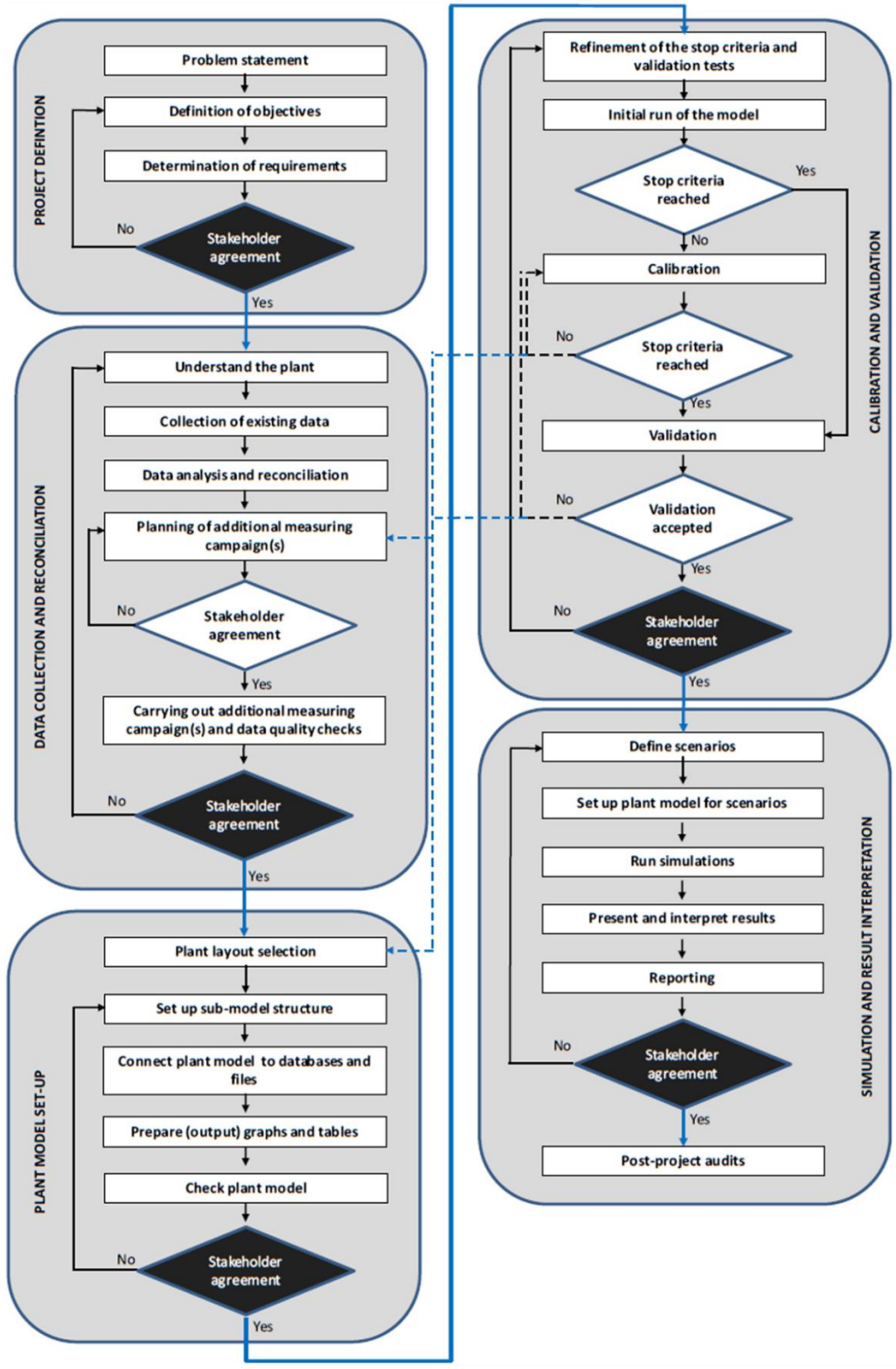


Figure 3-10: Good Modelling Practice of the IWA indicating the five steps of model development (Rieger et al., 2012).

In the third step, the plant model set-up, the model layout, including sub-models and the simulation outputs (e.g. graphs) are to be decided upon. In this research, the mechanistic modelled output will be sent to text files that can be used as input by the data-driven model and hybrid model.

Step four involves the actual model calibration and evaluation of the simulated results to be within an acceptable error to the measured data. Validation tests are performed in order to ensure the use of the model with the level of confidence required to meet the modelling objectives.

The evaluation criteria used to evaluate the model performance in this study are the mean error (ME) (equation 2.2) between observed and simulated data, the mean absolute error (MAE) (equation 2.3) and the root mean square error (RMSE) (equation 2.4), as well as their relative values. The relative errors are defined as the quotient of the error by the mean of the observations ($\hat{y}$) and make a clear distinction possible of the accuracy of the model. The closer the MAE/$\hat{y}$ and RMSE/$\hat{y}$ values are to 0, the better the accuracy of the model, but values of up to 15% - 20% are considered acceptable. The Janus coefficient (equation 2.5), which has an optimal value of 1 and is statistically considered acceptable up to a value of 2, makes it possible to state whether the model is valid based on the ratio between the RMSE for validation compared to the RMSE of the calibration (Hauduc et al., 2015).

In the final phase (step five), the calibrated and validated model is used to run the simulations that have been asked for in the project definition as the objectives of the modelling work. This step includes defining scenarios; setting up the plant model and input files for these scenarios; running simulations; presenting and interpreting results (Rieger et al., 2012).

Although comprehensive in its approach, the GMP Unified Protocol is considered as a general guideline more than as a detailed roadmap for model calibration. Mannina et al (2011) presents a detailed procedure for the calibration of an activated sludge model based on a comprehensive sensitivity analysis and a step-wise Monte Carlo-based calibration of the subset of influential parameters. The main advantage of this method and its key point is that the calibration is undertaken for sub-groups of variables in a step-wise procedure instead of solving a complex multi-objective function.

The methodology of Mannina et al. (2011) consists of multiple steps divided over two major phases (shown in Figure 3-11). In the first phase, the various parameter subsets are selected by a procedure that is based on the methodology proposed by Weijers & Vanrolleghem (1997). The first phase starts by selecting the output variables of interest (e.g. NO_3-N) and setting the parameters at their default or best-known values found in relevant literature with an uncertainty range and distribution. Monte Carlo simulations are then performed where each time a single parameter is varied at a time (OAT - *One At a Time*) within the specified uncertainty boundaries in order to evaluate the influence of the parameter on the model output. This allows calculating sensitivity coefficients, which are then scaled and sorted. A model parameter is considered influential if its average scaled sensitivity coefficient is larger than 0.1. This procedure is basically established to ensure that at least one parameter will be selected for each output available (Weijers & Vanrolleghem 1997). The last step in the first phase is the evaluation of the sensitivity coefficients which will permit to select the set of most influential parameters for each representative output.

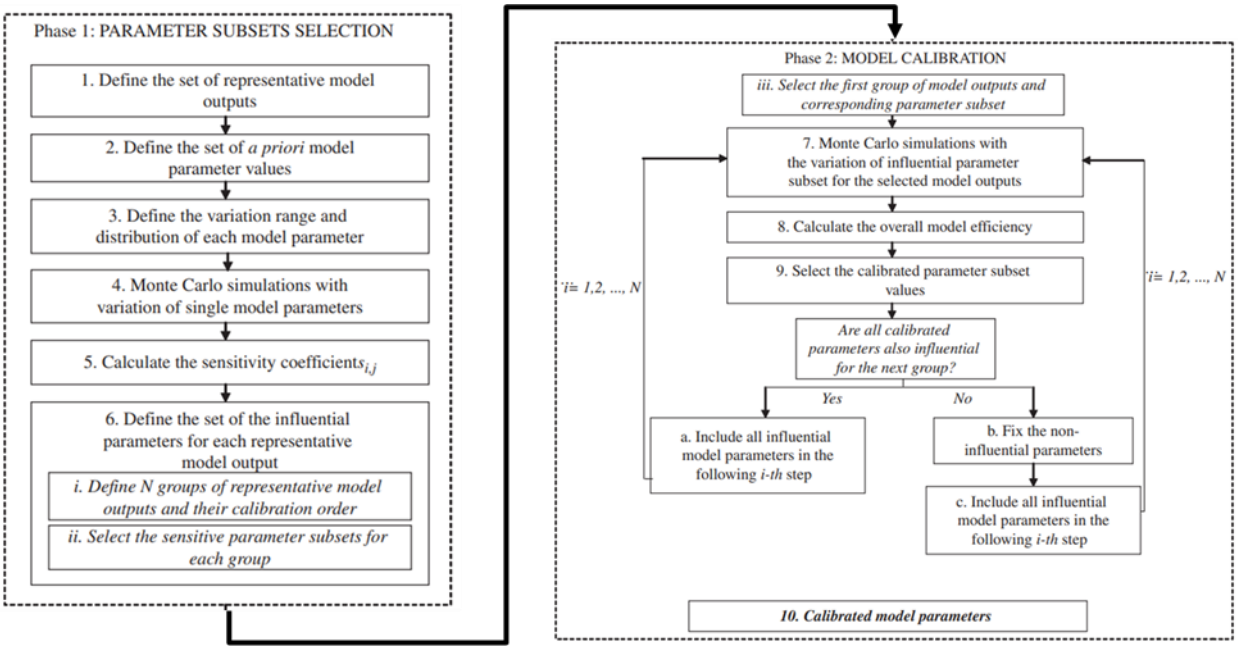


Figure 3-11: Two-step calibration procedure proposed by Mannina et al. (2011) with parameter subset selection based on a sensitivity analysis in phase 1, followed by a group-wise model calibration based on selected effluent variables in phase 2 (Mannina et al., 2011).

In the second step, the actual model calibration is performed by means of a group-wise Monte Carlo technique using only the subsets of model parameters selected in phase 1. The procedure begins with the selection of the first group of output variables of interest and the corresponding model parameter subset. The model parameter subset is calibrated according to the model outputs and an objective function by carrying out a number of Monte Carlo simulations to evaluate the quality of model fit for different parameter values. For each Monte Carlo run, the model fit with respect to each representative model output is calculated. After the Monte Carlo simulation has been performed, the overall model fit is determined, and the best parameter values are selected.

Parameters that are influential to more than one group of output variables are included in the calibration step of that other group. The parameters that are influential only for the first group are fixed at their calibrated values for the next calibration step with a new set of parameters influencing the second group of output variables. This second phase is then repeated for each of the groups of output variables with corresponding parameter subset.

3.3 Data-driven Model Development

Data-driven models (DDM) using machine learning can be built via a supervised training, where the model is trained on known input data and known responses to that data to generate reasonable predictions for the response to new data. Supervised learning uses classification and regression techniques to develop predictive models.

Another approach is unsupervised training that uses only the input data to find hidden patterns or intrinsic structures in the data. Unsupervised learning is mostly used for exploratory data analysis to find hidden patterns or groupings in data (such as clustering).

Since the objective of this research is to build a model that makes predictions of continuous responses based on input data that comes in a continuous format, the supervised machine learning approach using regression techniques to develop predictive models was selected for the training of the DDM.

3.3.1 Methodology for Training & Testing of Data-Driven Model

A commonly applied methodology for machine learning model development consists of the following main steps: data preparation/cleaning, feature extraction, feature selection, algorithm selection, model training, and model evaluation (Cohen-Shapira and Rokach, 2022).

In the data preparation step, the collected operational datasets cover the period 01.12.2019 to 15.03.2020, corresponding with those used for the calibration/validation set of the mechanistic model. The data were split 80% – 20% for training and testing purposes in chronological order to respect the time-dependent character of the observations. The training dataset contains 8 000 data points for each of the response variables NO_3-N, TSS, total COD, with a 15-minute interval covering the period 1.12.2019 to 28.02.2020. The testing of the ML models was performed with the remainder of the data (1 300 points) covering the period 1.3.2020 to 15.3.2020, which are the last two weeks of the dataset used for calibration of the mechanistic model.

The datasets used for model input required gap filling of missing data points, for which the moving median method was applied. Data input for the ML models is normalized using the z-score, with a mean $\hat{y}$ and a standard deviation σ, and scaled in the range of 0 to 1 to provide uniform data across all features that still have the original shape properties.

Selection of the right machine learning model is partly based on trial and error and requires trading off some strong points (e.g. computational speed, accuracy, and model complexity) against another. Highly flexible models tend to overfit data by modelling minor variations that could be noise in the data. Simpler models are easier to interpret, but might have lower prediction accuracy. The best approach is to try out a variety of machine learning models and choose the best according to an evaluation criterion.

The evaluation of the performance indicators for the DDM model calibration took specifically into account the root mean square error (RMSE) (see equation 3), since this indicator provides a quantitative measure of how well the model fits to the data average. The mean error (ME) (equation 1) and mean absolute error (MAE) (equation 2) were supplementary indicators used to get a sense for the model's bias and of the error's size on average. Cross-validation was applied to divide the training data into multiple (k) folds (subsets) and the holdout method is repeated k number of times. In the holdout, each subset was used once for validation. The average error across all partitions is then reported as the average validation error. This method increases training performance such that the score of the model has a lower dependency on the selection of the training data.

Optimization of the selected ML model was done by changing hyperparameters, such as the number of neurons, the weights and the biases. Feature importance analysis was applied using the F-Test method that enables the selection based on which features are influential on the response. The F-test examines the importance of each feature individually and allows defining the likelihood that an observed improvement of a fit to data is worth the use of the extra feature. It tests the hypothesis that the response values are drawn from datasets with the same mean against the alternative hypothesis that the population dataset means are not identical. Larger scores (including infinites) indicate greater

importance. Effectively, it measures to which degree features add any new explanatory power. The 'cost' is indicated by the loss in the degrees of freedom due to the added number of features (Helsel & Hirsch, 1992).

The selection, training and testing of the best machine learning model was performed in the toolkits environment for Regression Learning provided in the MATLAB R2022a (MathWorks) software.

3.4 Hybrid Model Development

The hybrid model developed in this study involves an interaction between a mechanistic model, which serves as the fundamental basis for the predictions, and a data-driven model that has learned to apply a correction to the predictions made by the former. The mechanistic model ensures the preservation of knowledge of the treatment process and offers good extrapolation capabilities. The data-driven model provides enhanced predictive capabilities through implicit learning of unknown relationships between the modelled mechanistic model input and measurements of the output variables. These capabilities can overcome the dynamics missed by the mechanistic model (Schneider et al., 2022).

The hybrid model developed here (see Figure 3-12) is defined as a cooperative parallel hybrid structure (Schneider et al., 2022) where the DDM predicts the difference between the output of the mechanistic model and the observations (the residual error Δ). Thus, the mechanistic model serves as the fundamental basis for the effluent quality predictions, and the data-driven model using machine learning techniques is asked to add a correction. This requires that the mechanistic model first predicts the values of process output variables (blue line in the figure), such as the effluent water quality variables and biomass concentrations. These are then communicated to the DDM model (black box) that as well receives the same influent data and important information about process control parameters (black line). The DDM uses this information to estimate the residual errors of the mechanistic model simulations Δ (green arrow), which are defined as the most probable expected model errors based on the training differences between previously observed (measured data) and simulated data. Finally, for the output of the hybrid model the two results are added together to obtain corrected values of the effluent variable (light blue highlighted arrow).

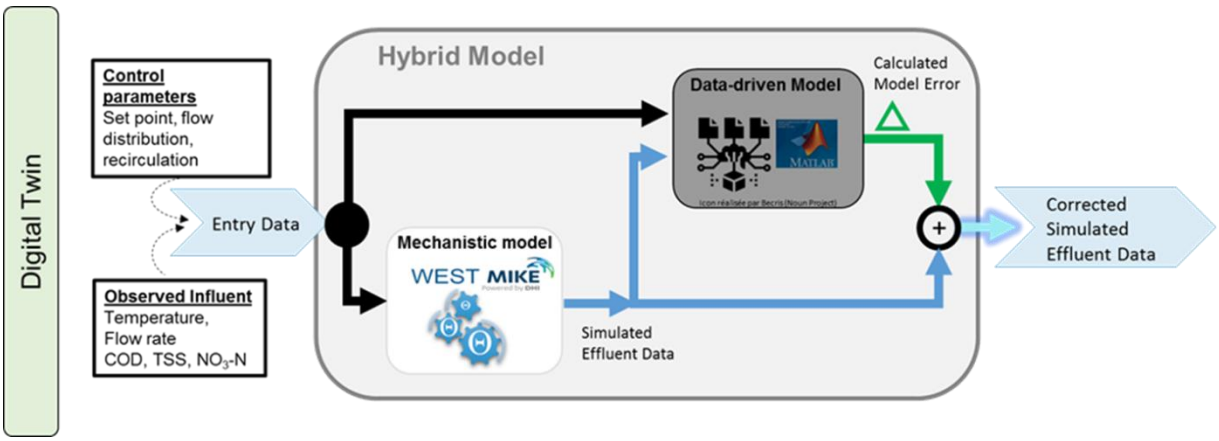


Figure 3-12: Structure of hybrid model (larger grey box) combining a data-driven model (black box) and a dynamic mechanistic biofilm reactor model (white box) that is based on the knowledge of the bio-physico-chemical processes.

The training objective is that the DDM learns to generate the proper model errors. During the training phase of the DDM, illustrated in Figure 3-13, observations of influent and effluent quality are transmitted to the DDM, as well as the simulated effluent variable by the mechanistic model. The

model error that is defined as the difference between observed and simulated effluent variables is calculated and applied as the response variable for the DDM model training with a Bayesian optimization algorithm. Here, the model error serves as labelled response in a supervised machine learning training environment.

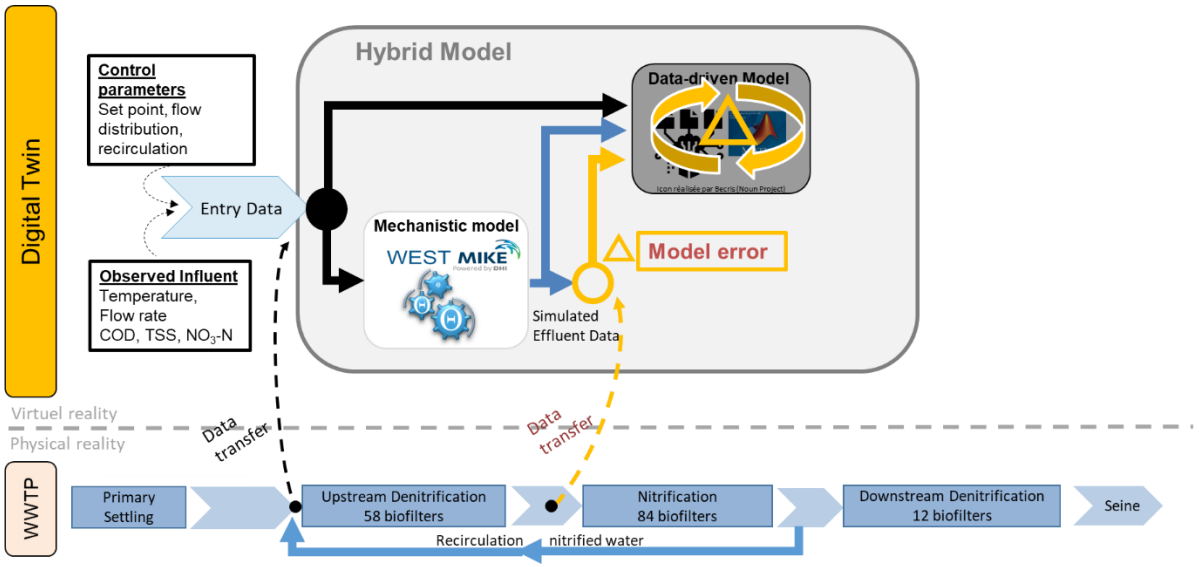


Figure 3-13: Training configuration of the hybrid model, which receives observations of influent and effluent data from the physical WWTP process, calculates the mechanistic model's error with respect to the observations and uses this information to train a DDM to predict the residual model error.

In this approach, a machine-learning model is trained for each of the output variables, where for each variable, a corresponding subset of process input variables or features is used.

3.5 Hybrid Model Predictive Control

The trained hybrid model is used in the hybrid model predictive control (HMPC) of methanol dosing in a predenitrification biofilter.

3.5.1 Structure of HMPC

As illustrated in Figure 3-14, the HMPC first searches for an optimal methanol dosing rate sequence by running multiple scenarios with the hybrid biofilter model over a certain prediction horizon. The sequence of dosing rates is varied over a so-called control horizon at a fixed number of times, e.g. 4 times). The MPC model (blue box in the figure) receives the measured influent data and relevant information about the process operations (e.g. nitrate set point) together with the methanol dosing rate sequence. In the future this information may easily be extended to include cost prices of methanol and environmental costs (e.g. GHG emissions) associated with the methanol dosing to make the optimization objective more comprehensive and practical.

The dosing of methanol has a direct negative effect on the nitrate concentration in the effluent. Higher doses of methanol will increase the metabolic activity of the heterotrophic organisms, which will consume more nitrates and thus lower the nitrate concentration in the effluent. The optimizer searches for the smallest error between observed nitrate and nitrate concentration set point.

The first of the sequence of optimal dosing rates is then identified and transmitted to "Reality", here represented by a second mechanistic biofilter model ("Reality" Process Model in red box). That model

represents the physical process. Feedback about the evolving effluent water quality in the physical system is assured by linking the modelled effluent from the "Reality" Process Model back to the HMPC, where it is used as a representation of the 'observed' effluent values. When a model-mismatch is detected, this will be corrected for by adjusting the NO_3-N set point of the MPC model. This model mismatch correction through set point adjustment is implemented as follows: when the output of the Reality Process Model is compared to the predicted output by the MPC Model, the averaged difference between the two models of the last 15 minutes period (symbolized by Ɛ in Figure 6-5) is communicated back to MPC model, where it is used to modify the set point applied during the next iteration. In this regard, the MPC model will strive to reach the new set point and thereby forcing the output of the Reality Process Model to move closer to the desired effluent concentration.

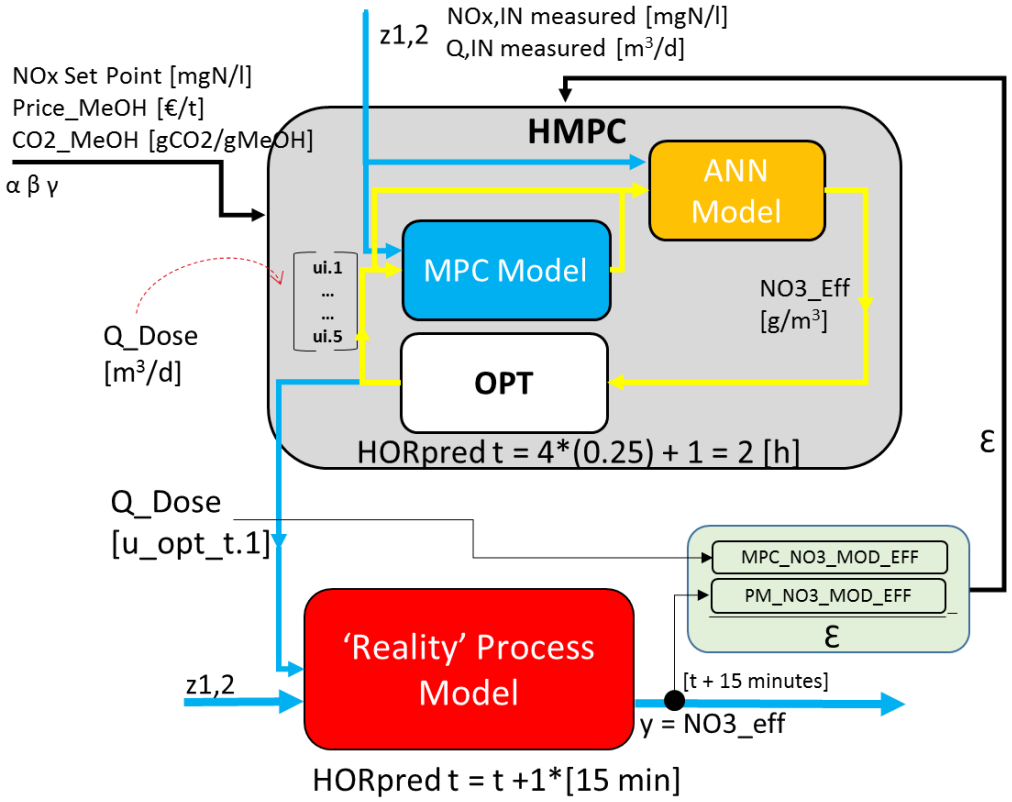


Figure 3-14: Flowchart of the developed Hybrid Model Predictive Controller (grey box) that consists of a First-Principles process model (blue box) and a DDM (orange) in an optimization loop. Only the optimal dose of the first-time step (u_opt_t.1) is applied to the process, represented by the "Reality" Process Model (red box). The HMPC receives feedback of the modelled effluent quality from the Process Model.

After the first iteration, the HMPC moves forward one-time step in a receding prediction horizon approach. The time step size has been chosen as 15 minutes intervals since the hydraulic retention time of the water in a biofilter is about 20 to 30 minutes. Any reagent dose injected at t = 0 will pass through the biofilter within 20 minutes, and its effects on the biological activity is assumed to be of the same order. For the optimization of the control actions, the HMPC model simulates over a prediction horizon of 120 minutes and allows variation of the methanol dosing over the first hour at 15-minute intervals. The last value of the methanol dosing is kept constant till 120 minutes. So, the optimization has five degrees of freedom, i.e. five dosing rates imposed over the 120 minutes.

The Process Model predicts only one-time step into the future (that is 15 minutes ahead) before a new optimization run is initiated, so that every 15 minutes information is sent back to the MPC model about the effluent water quality.

The HMPC is written in the Python language using Spyder in order to allow automatic execution of the DDM and the biofilter models in WEST (DHI). The development of the DDM was performed in MATLAB R2022a (MathWorks).

3.5.2 Scenarios Workflow in Python

For an automated execution of the different scenarios, the (hybrid) model predictive controller has been written in python using modular scripts. This allows executing the data-driven model and the mechanistic models in an automated mode and facilitates data exchange between the models.

The Python scripts that have been developed are responsible for data preparation, analyses and execution of xml files that run the WEST mechanistic model and the trained ANN model. The performance of each optimizer scenario is then evaluated in terms of the RMSE, and the scenario with the lowest error is selected as the best. The proposed action of the first-time step that belongs to that best-fitting scenario is then finally implemented in the Reality Process Model. In the case of the HMPC, the mechanistic model's effluent is first corrected by the DDM that estimates the model's error for each scenario. The evaluation that follows is then identical to the path just described. The details for the different scenarios and the workflow for the MPC and HMPC controller are described in more detail in chapter 6.

3.6 Concluding Remarks

The development of the Hybrid Model and its application in the Hybrid Model Predictive Controller will be further discussed in chapter 6. One of the aspects that deserves more research for future applications is the link with the real physical system. A connection to the real system, collecting data and providing methanol dosing rates can replace the knowledge-based "Reality" Process Model used now in the demonstration study.

4 Mechanistic Modelling of Wastewater Treatment for Nitrogen Removal through Biofiltration

This chapter discusses the performance of the biofiltration model developed by Zhu (2020) for the prediction of the concentrations for the effluent quality variables of the predenitrification (preDN), nitrification (NIT) and postdenitrification (postDN) stages in the biofiltration lane. As a reminder, the mechanistic model simulates the physical and biological phenomena taking place in a submerged biological activated filter with a 'fixed' culture of micro-organisms that grow attached to inert polystyrene beads (4.5 mm diameter). The bacteria are metabolically active within a biofilm, a thin layer surrounding the beads.

The development of the biofilter model (Zhu, 2020) had as objective to improve existing models for the biofiltration treatment lane at the Seine-aval WWTP or similar WWTPs by providing detailed phenomenological descriptions of the process behaviour of each biofiltration treatment process and simulating long term trends of effluent quality and estimate economic costs and environmental impact. The model would serve as a support tool for the design and upgrades of treatment systems. In a basic sense, the biofilter model aimed at describing the time evolution of the mass balances for different components, composed of a mass transport term and a conversion term (i.e. the biological reactions). The variables that are of interest to the user (i.e. the model outputs) include those that describe the organic matter content (total COD, soluble COD), the suspended solids (TSS) and the nutrient compounds including nitrogen (TKN, NH_4-N, NO_3-N, and NO_2-N) and phosphorus (PO_4-P and TP). The model could be adapted for inline forecast of water quality to better optimize operations.

The biofiltration stage provides secondary wastewater treatment of the effluent of the primary treatment stage. At the time relevant for this study, the latter consisted of a 3-step treatment using primary settling tanks, followed by aeration basins and finalized by a second step of chemically enhanced primary treatment (CEPT) in which iron salts ($FeCl_3$) are dosed to enhance the coagulation of organic matter in order to increase its removal.

For secondary treatment, the biofiltration lane shown in Figure 4-1 consists of a 3-stage pre-denitrification (preDN), nitrification (NIT) and post-denitrification (postDN) system. The preDN stage puts to good use the 'natural' organic carbon content of the settled water to reduce nitrates to di-nitrogen, but necessitates an input of nitrates by nitrified water through recirculation of the NIT effluent. As indicated in the figure, in case the readily biodegradable carbon concentration is insufficient, an external dosing of methanol (CH_3OH) is possible. The NIT treatment stage functions with active aeration (O_2) of the filters. During the postDN stage, a constant external carbon dosing in the form of methanol is required for denitrification due to the low organic matter content in the effluent of the NIT stage.

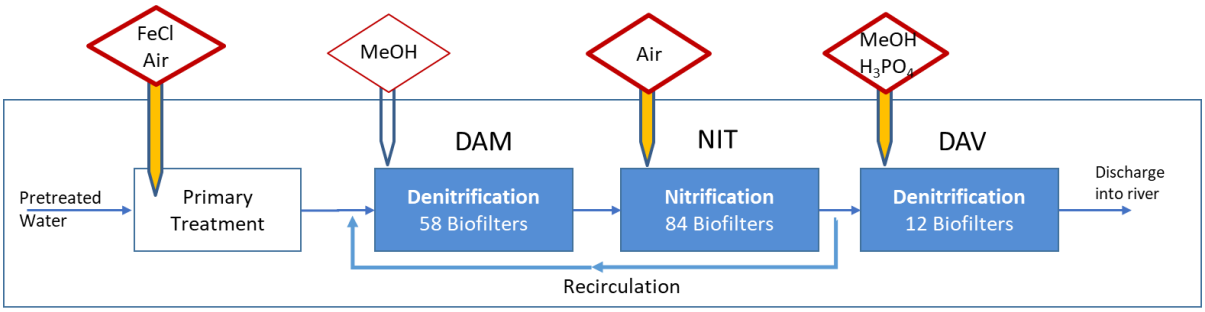


Figure 4-1: Schematic layout of wastewater biofiltration treatment lane at Seine-aval WWTP with indication of chemical dosing and aeration.

This chapter will discuss the outcomes of the calibration and validation of the parameters of the biofiltration model. A particular focus will be on the predenitrification stage, since Zhu (2020) hadn't had the time to calibrate this biofilter model for that treatment stage using long term high-frequent operational data from the Seine-aval WWTP. Moreover, the development of a hybrid model (see Chapter 5) and its application in a continuous model predictive process control system for this predenitrification stage (see Chapter 6) will take place with operational data from the predenitrification stage.

The chapter starts with a description of the datasets and the results of the data-cleaning exercise, followed by the influent characterization and the definition of the parameters for process control. This is followed by the identification of influential parameters through a global sensitivity analysis. For calibration and validation, the performance of the model simulations of effluent quality variables is evaluated by comparing it to observations for the preDN in stage in detail. Finally, the output of the integrated biofiltration model covering the NIT and postDN treatment stages is discussed.

4.1 Description of datasets

Measurement data from inline sensors, analysers as well as daily laboratory analyses and other observational data (e.g. maintenance reports) from the Seine-aval station were accessed in the 'BaSta Virtuel' databank that allows visualizing and exporting data in tables and spreadsheets. This secured databank is managed by the utility SIAAP and contains for each sampling point the meta-data, such as information concerning the sampling frequency, unit of measurement, responsible laboratory staff and dates of installation.

The period chosen for this research runs from December 1st, 2019, until September 31st, 2020. The start of this data range coincides with the installation of in-line probes for COD (Hach Lange, UVAS Sc UV probe) and nitrates (Hach Lange, Nitratax Sc 1mm, UV probe) measurements at the entry and exit of the predenitrification stage in the Seine-aval plant. These sensors became fully operational on November 26th, 2019. The model was calibrated on the first three and a half months available (01.12.2019 – 15.03.2020) which covers 106 days, while for the validation the data from June 1st, 2020 until September 30th, 2020 (122 days) was used. The period in between, from March 16th until end of May does not contain reliable and consistent data because of the Covid-19 restrictions.

The nitrate and nitrite concentrations, which are soluble compounds containing nitrogen and oxygen, are important performance indicators of the preDN treatment. The inline analysers actually monitor the combined compounds of NO_2-N + NO_3-N = NOx-N (mg NOx-N/L) and are located in the outflow channel of the primary settler (see Figure 4-2), where all effluents of the primary settlers come

together. This location is before the inlet of the recirculation of nitrified water that provide nitrates and nitrites to the preDN. In the settled water, NOx-N are present only in very low concentrations. In order to identify the concentration of NOx-N at the entry of the preDN stage, the flow weighted contributions of settled water and nitrified waters were calculated. In the case of settled water, the NOx-N concentration is measured by a sensor. However, as can be seen in Figure 4-2, in the case of the recirculation of nitrified water, the concentration of NOx-N (mg NOx-N/L) is monitored by inline sensors at the entrance to the postDN.

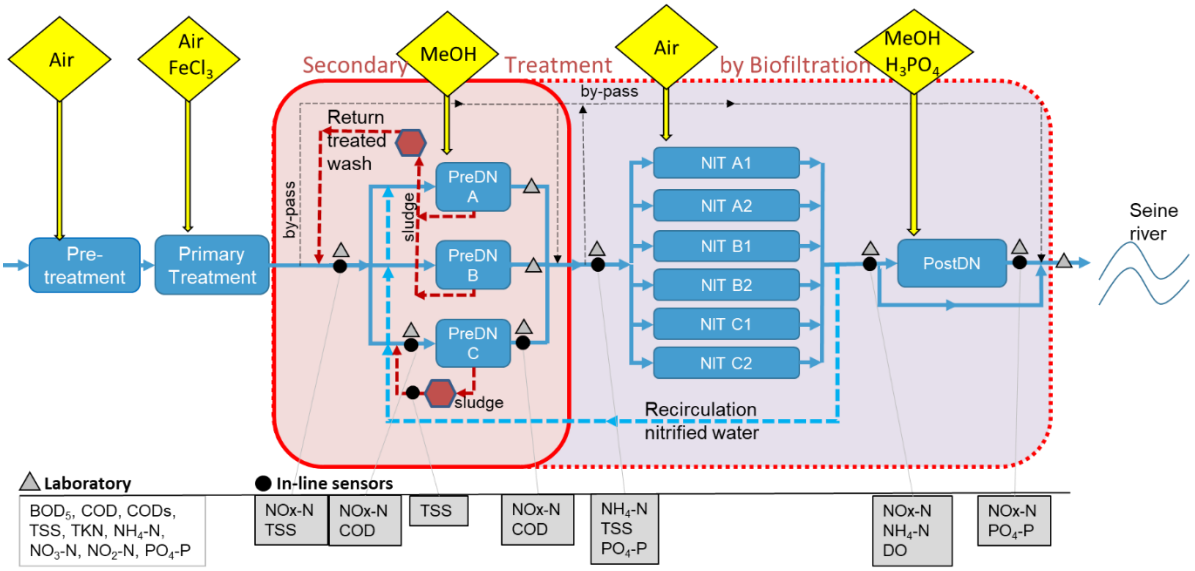


Figure 4-2: Schematic diagram of the principal wastewater treatment lane at the Seine-aval WWTP (SIAAP) with indication of processes (blue boxes) and main dosing actions (yellow diamond shapes). Outlined in red frame are the secondary treatment predenitrification biofilter units (PreDN), the nitrification battery units (NIT) and postdenitrification unit (PostDN). The focus of this study is on the PreDN stage. Also indicated are the sludge treatment units for PreDN (brown hexagons) and location of measurements by in-line sensors (black dots) and laboratory samples (grey triangles).

In addition, as of November 2019, a NOx-N sensor is available at the entrance and exit of preDN battery C. As explained before in chapter 3, the preDN battery C contains 18 BioStyr® filters that receive treated backwash water from the sludge that is removed from the filters through daily retro washing. The composition of the influent water is therefore different than the 'average' water mix treated by batteries A and B. This backwash flow, albeit with a less significant flowrate, has to be added to the overall flow into Battery C. The concentration of TSS (mg/l) is measured at the outflow of the backwash treatment. The backwash water and sludge originating from the filters in battery A and B (20 + 20 Biostyr® filters) is treated in a separate sludge clarifier and this effluent is reinjected into the main incoming stream flowing into the preDN lane, but before it is monitored by the inline sensors for NOx and TSS.

Besides water quality and quantity measurements, data related to process control data is registered in the database, such as the number of filters in operation at each 15 minutes time step, the set points for nitrate concentration in the effluent of the preDN and postDN, the ammonia set point in the effluent of the NIT stage, as well as information about the pumps used for aeration and distribution of water flow (number of hours; energy usage; etc.), the duration of backwashing cycles and the quantity of reagents dosed. These data allow an analysis of the current state of the system. The data could be applied in (automated or by operator action) control strategies on selected process parameters for

which a target value ('set point') is predetermined and for which a modifiable variable is corrected in order to reach this desired set point (e.g. a NOx-N concentration of 1.3 mgN/L in the effluent of preDN).

4.1.1 Operating Conditions during Calibration

Table 4-1 describes the operating conditions of the biofiltration lane during the calibration period based on the information collected from daily flow proportional composite laboratory samples and sensors where available. The daily values at the entrance of the preDN had to be calculated since there is no direct sampling point at this location. This was based on the flow weighted averages and the measurement data available from the exit of the primary treatment phase and the entrance of the postDN, which is assumed to be representative of the nitrified recirculation water quality.

Table 4-1: Working conditions of the predenitrification stage at the Seine-aval WWTP biofiltration lane during the calibration period.

Seine-aval Operating Conditions Entry preDN	**Laboratory data (flow weighted calculated from Settled Water and Nitrified Water)**			**Sensor data**		
01.12.2019–15.03.2020	Nr. of observations	Mean	Standard deviation	Nr. of observations	Mean	Standard deviation
Flow rate (m^3/d)	-	-	-	10 177	2 075 378	462 489
Number of active filters (nb)	-	-	-	10 177	50	10
Flow rate per filter (m^3/d)	-	-	-	10 147	41965	5993
Recirculation rate (%)	-	-	-	10 052	0.8	0.2
BOD_5 (mgO_2/L)	95	50	12	-	-	-
Total COD (mgO_2/L)	95	147	35	9 560	157	72
TSS (mg/L)	95	65	16	-	-	-
TKN (mgN/L)	93	28	6	-	-	-
NH_4-N (mgN/L)	100	20	5	-	-	-
NO_3-N (mgN/L)	100	9	2	9 754	9	3
NO_2-N (mgN/L)	100	0.5	0.2	-	-	-
PO_4-P (mgP/L)	101	0.9	0.35	-	-	-
Ratio COD/NOx (mgO_2/mgN)	100	16	5	9 725	20	17
Ratio BOD_5/NOx(mgO_2/mgN)	100	5	2	-	-	-

During this period, the preDN stage treated on average 2 million m^3/day wastewater of which 1.1 million m^3/d was primary effluent and 0.9 million m^3/d was recirculated nitrified water. The daily average flow rate per filter was about 42 000 m^3/d, which corresponds to a hydraulic load of about 10 m/h (filter surface area of 173 m^2) and a Hydraulic Retention Time of 20 minutes; all considered to be standard operating conditions. As reported in the literature review (chapter 2), the filtration rate should at least be 1.5 m/h to avoid clogging the filter with suspended particulates, but should not be too high as that will reduce the hydraulic residence time. The Carbon to Nitrogen (C:N) ratio, which is an important process factor of the filter's performance, is about 16 for COD/NO_3-N in preDN; this is in accordance with Rother and Cornel (2007), who report a COD/NO_3-N ratio of 15 in an upstream denitrification biofilter responsible for a denitrification efficiency of 80%.

The calculated volumetric loading and removal rates of organic matter and the main pollutants [kg/(m^3*d)] is shown in Table 4-2 for the three treatment stages. The volumetric loading rate for nitrates in the preDN stage is about 0.6 kgN/(m^3*d) and for postDN this is about 2.6 kgN/(m^3*d) during the calibration period. This falls in the range reported by Metcalf & Eddy (2008) of 0.3 to 5.0 kg NO_3-N/(m^3*d) for denitrification filters. The significantly higher volumetric loading rate for postDN is

characteristic for treatment with external carbon dosing as it allows for higher filtration rates of about 17 m/h during the calibration period and 22 m/h for the validation period. Thanks to the continuous addition of methanol in the postDN stage, it achieves a high NO_3-N removal efficiency of 80% and 90% respectively. This is in contrast to the preDN stage, where the nitrified water is mixed with settled water and methanol is only added when necessary.

Figure 4-3 indicates the average values of the water quality variables sampled for the biofiltration lane, both by flow proportional composite samples analysed in the laboratory as by in-line sensors, as well as the process control parameters, such as the set points for nitrates concentrations in the effluent of the preDN and postDN stages. Noteworthy, the average organic matter concentration during calibration that leaves the primary settling phase contains around 90 mg BOD_5/L, a value that is almost five times the admissible discharge concentration, indicated in the blue frame box on the right side of Figure 4-3.

Table 4-2: Performance indicators of Seine-aval WWTP biofiltration lane for the three stages showing average daily values during the calibration (Cal) and validation (Val) periods.

Seine-aval - Biofiltration lane	**preDN**		**NIT**		**postDN**	
	Cal	Val	Cal	Val	Cal	Val
Filtration Rate (flow rate per filter) [m/h]	10.1	9.5	7.2	7.2	16.8	22
BOD_5 Volumetric Loading Rate [kg/(m^3*d)]	3.4					
TSS Volumetric Loading Rate [kg/(m^3*d)]	4.4					
NO_3-N Volumetric Loading Rate [kg/m^3*d]	0.6				2.6	3.1
NO_3-N Removal Rate [%]	65	74			80	90
NH_4-N Removal Rate [%]	-	-	88	93	-	-
COD total Removal Rate [%]	44	54				
TSS Removal Rate [%]	77	65			-	-

About 76% of the BOD_5 will be removed in the preDN stage by the heterotrophic bacterial activity as substrate for the conversion of nitrates (average influent concentration of 8.9 mgN/L) to N_2 gas. The remaining NO_3-N concentration in the preDN effluent is on average 3.1 mgN/L, which gives a 65% removal rate. The downstream denitrification stage (postDN) further treats the NIT effluent. It receives nitrified water with an average concentration of 20 mgNO_3-N/L and brings that down to 4.1 mgNO_3-N/L. The flow is then further mixed with the bypassed nitrified water and the MBR effluent, so that its final concentration is about 12 mgNO_3-N/L on average.

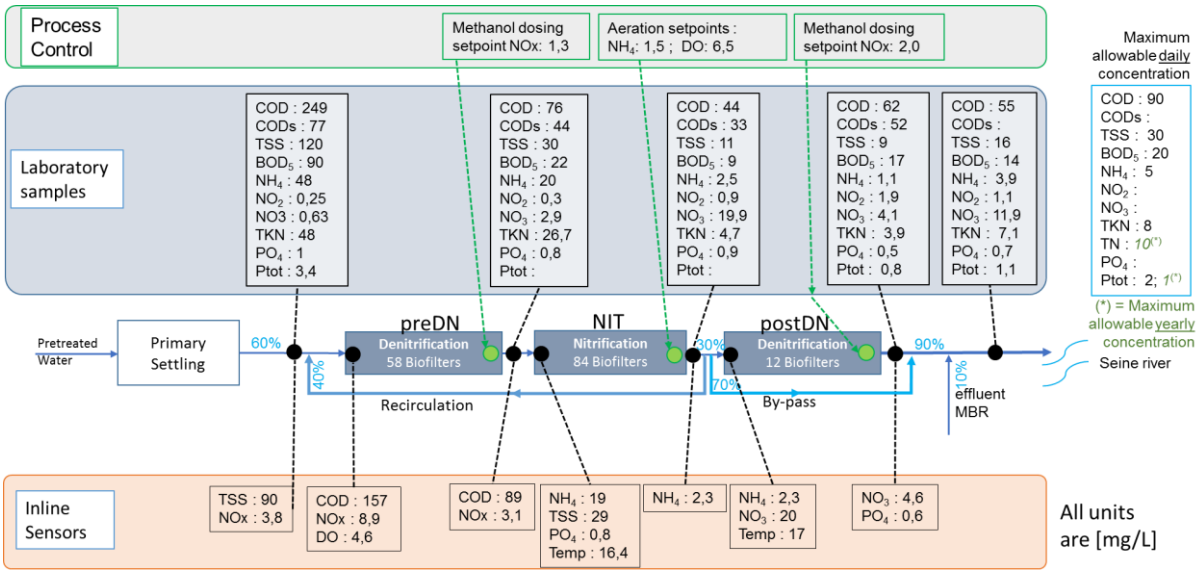


Figure 4-3 : Average values of measured variables at in the biofiltration treatment lane, the Process Control parameters (in green box) and the regulatory admissible discharge concentrations during the calibration period. Also indicated are the average flow distribution percentages (light-blue coloured numbers). All units expressed in [mg/L], except where otherwise indicated.

Concerning the nitrogen concentration, the effluent of the two-step primary settling phase has a daily average of 48 mgN/L, which is roughly 10 times the allowable concentration in the discharge into the river. The recirculation of nitrified water, which carries low values of ammonia (NH_4-N: 2.3 mgN/L) will cause a dilution and thus a lowering of the ammonia concentration at the inlet of the preDN. A small fraction of ammonia is taken up by the microorganisms for their growth in the preDN stage. At the influent NIT stage, the ammonia concentration amounts to 19 mgN/L (sensor) to 20 mgN/L (lab). During NIT, the autotrophic bacteria convert 89% of ammonia to nitrates and nitrites under aerated conditions.

The Total Nitrogen TN, which is sum of TKN and NOx-N, is an important indicator for the nitrogen concentration in the water discharged into the river. For a yearly running average, the TN concentration should be lower than 10 mgN/L. It can be seen in Figure 4-3 that the primary effluent contains a TKN concentration of 48 mgN/L, most of which is ammonia. During the subsequent biofiltration stages, the ammonia concentrations drop after passing through each treatment stage, and with it the TKN concentration. In the effluent of the postDN, there remains a fraction of organically bound nitrogen of about 2.8 mgN/L that is assumed to be unbiodegradable soluble (S_NI) that stays in the water or inert particulate (X_NI) that is captured and exits the system via the sludge line.

4.1.2 Operating Conditions during Validation

Table 4-3 describes the operating conditions of the biofiltration lane during the validation period based on the information collected from daily laboratory samples and sensors where available. The average flow rate per filter during this period (39 000 m^3/d) is slightly lower (7%) compared to that during the calibration period (42 000 m^3/d), but the average recirculation ratio of nitrified water from the exit of NIT to the entry of preDN has increased significantly from 0.8 during calibration to 0.94 during the validation period. Since the recirculation is interrupted during days with high rainfall, this higher percentage is a clear indicator of prevailing dry weather conditions during the validation period from June till end of September. It is also closer to the desired ratio of 1. The higher NO_3-N removal rate

(74% during validation vs 65% during calibration – see Table 4-2 demonstrates that the denitrification process was better performing.

The observed quality variables shown in Table 4-3 have very similar values when compared to the calibration period. The particularity of the summer vacation month of August, when a large portion of citizens leave the city, are discernible with lower overall flow rates.

Table 4-3: Working conditions of the preDN stage at the Seine-aval WWTP biofiltration lane during the validation period.

Seine-aval Operating Conditions entry preDN **01.06.2020–30.09.2020**	**Laboratory data (flow weighted calculated from Settled Water and Nitrified Water)**			**Sensor data**		
	Number of observations	Mean	Standard deviation	Number of observations	Mean	Standard deviation
Flow rate (m^3/d)	-	-	-	11 713	2 011 387	284 056
Number of active filters (nb)	-	-	-	11 713	51	4
Flow rate per filter (m^3/d)	-	-	-	11 713	39 208	5 848
Recirculation rate (%)	-	-	-	122	0,94	0,2
BOD_5 (mgO_2/L)	102	42	19	-	-	-
Total COD (mgO_2/L)	121	147	48	11 703	157	57
TSS (mg/L)	121	68	31	-	-	-
TKN (mgN/L)	101	27	5	-	-	-
NH_4-N (mgN/L)	122	19	5	-	-	-
NO_3-N (mgN/L)	122	9	2	11 703	8	4
NO_2-N (mgN/L)	122	0,5	0,2	-	-	-
Ratio COD/NOx (mgO_2/mgN)	118	16	8	11 616	26	24
Ratio BOD_5/NOx (mgO_2/mgN)	102	4	2	-	-	-

4.2 Data Preprocessing

Data reconciliation involves the preprocessing of the raw measurement data in order to remove biases of in-line sensors, probes and analysers, together with faulty data and outliers, as well as gap filling in order to obtain a complete set of reliable and consistent data. This treatment minimizes errors that affect the model's performance. Data cleaning is a necessary step in data analysis and an important factor to make sure it is accurate, complete, consistent and trustworthy. A first step therein is visualization of data and plausible checks (Rieger et al., 2012).

4.2.1 Visualization

Data visualization, structuring and descriptive statistics allow for sanity checks, together with mass balances to spot inconsistencies and or incoherent data.

Figure 4-4 shows the distribution of water volumes (expressed in flow rates) that arrive at the preDN battery C treatment stage, of which the primary effluent and the recirculation volumes make up the biggest part. The total flow rate to the preDN stage Battery C during the period surveyed averages about 600 000 m^3/d. At the start of December 2019, the proportion of settled water is almost identical to that of the recirculation of nitrified water, both fluxes contribute equally with about 250 000 m^3/d. The return flow from the sludge treatment is the third largest contributor with 70 000 m^3/d, while the by-pass is not used in significant volumes.

However, between January 14th and March 17th, the volume of settled water is significantly higher and averages about 350 000 m^3/d, with high peaks of 500 000 to 600 000 m^3/d late January and beginning

February, caused by heavy rainfall. It is in this period that the bypasses start to function, indicating that the maximum capacity of the battery is exceeded. The volume of recirculation of nitrified water remains relatively stable.

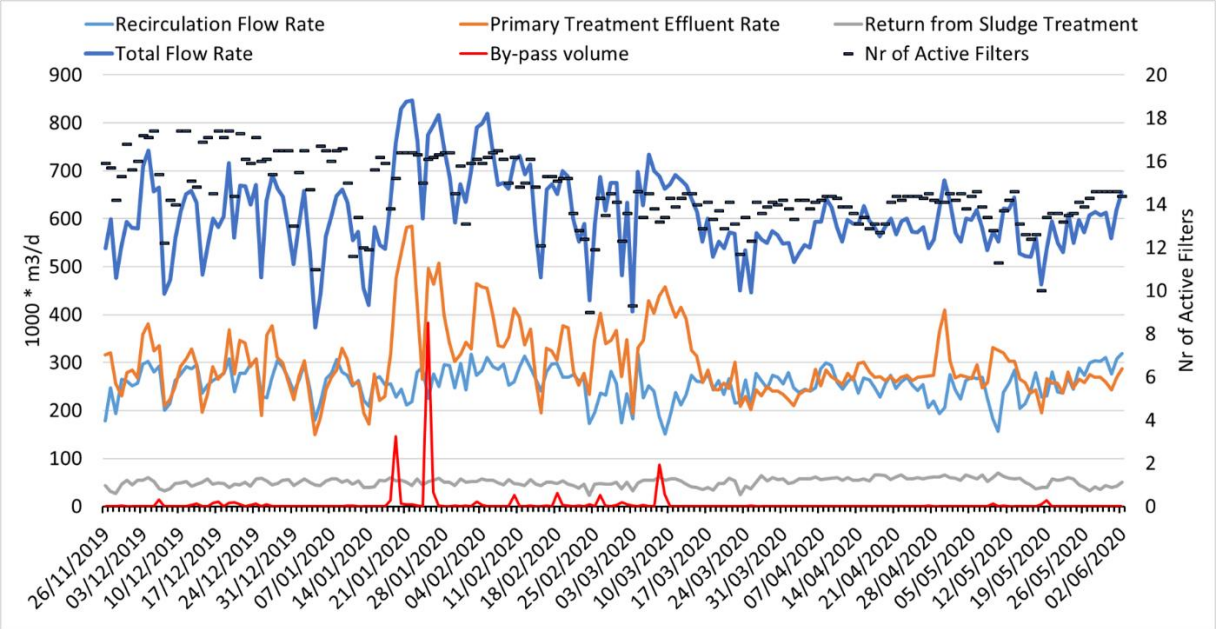


Figure 4-4: Daily influent flow rates per source into the preDN battery C at Seine-aval WWTP (left vertical axis), as well as number of active filters per day (right vertical axis) for the period from end of November 2019 till beginning of June 2020.

The frequency distribution of flow rates in the influent in Figure 4-5 clearly indicates a two-peak distribution; a first one for dry weather around 575 000 m^3/d and a second peak at 675 000 m^3/d under the rainy weather conditions. Also, it can be seen that this period between 1 December and 15 March (112 days) can be considered a wet period with approximate 50 days with flow rates of at least 651 000 m^3/d, which account for 85% of all flow rates registered during this period.

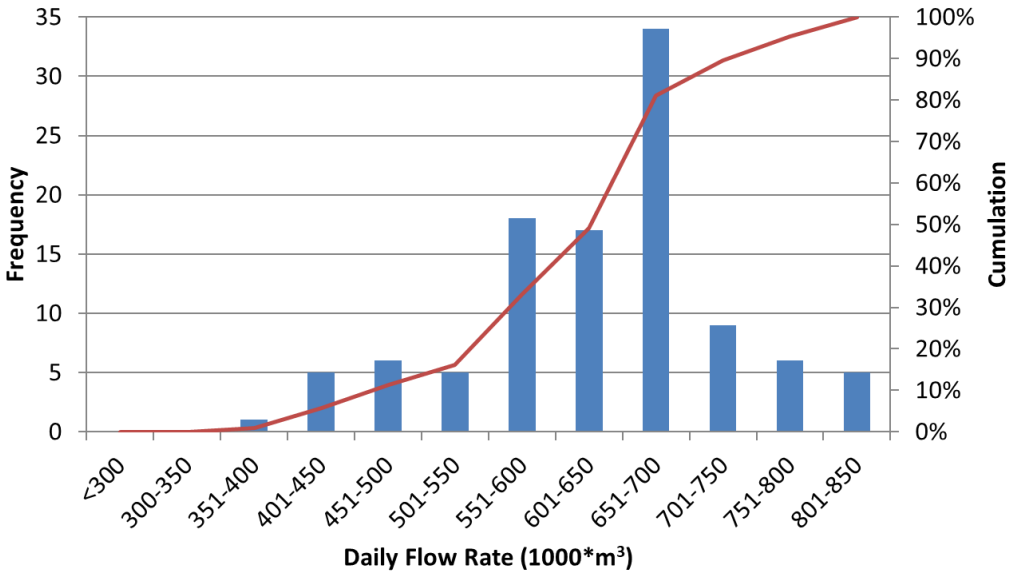


Figure 4-5: Distribution of flow rates in the Influent of preDN battery C.

Another interesting visualization method is to look at the correlation between laboratory and sensor data. Figure 4-6 shows the laboratory flow proportional composite sample concentration of nitrates in the influent of the preDN stage and compares it to the daily averaged values measured by the in-line sensor at the same location. For purposes of statistical reference, the coloured band indicates the range of mean +/- 1 standard deviation (SD), corresponding to 68% of the data range in a normally distributed data set. For NO_3-N, the laboratory data is indicating higher values overall with an average of 10.1 mgN/L; while the sensor has an average value of 9.0 mgN/L. Actually, the sensor data, which measures a combination of NO_3-N + NO_2-N; the outputted signal has been recalculated to NO_3-N and NO_2-N using the laboratory daily ratios.

In the month of December, an overall difference between laboratory results and sensor measurements is noticed with laboratory observations about 2 mgN/L higher than the sensor readings. Then, from January to February, both types of observations have similar daily average values. However, starting from the beginning of February, the sensor signal starts to show larger variability and overall lower average values that fall far outside the 1*SD range. This seems to coincide with the peaks of high settled water fluxes (Figure 4-4) indicating that these sensors might be better at registering short and intense compositional changes of the water than the laboratory composite samples. The laboratory value of 3.5 mgN/L measured on February 1st deviates considerably from the daily average sensor data (11.5 mgN/L); an issue that requires further analysis during data cleaning.

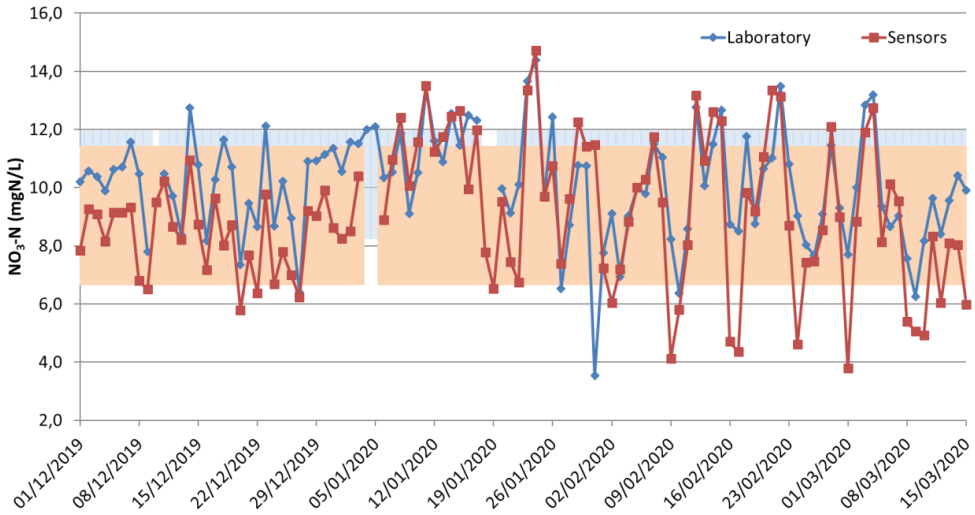


Figure 4-6: Observed concentrations of NO_3-N in the influent of preDN battery C by inline sensors using a decomposed signal (red line), daily average of sensor (red squares) and laboratory composites (blue). The orange zone and blue zone indicate the range of [mean-1SD : mean+1SD].

4.2.2 Data Cleaning

The methodology applied for data cleaning is the method by Alferes and Vanrolleghem (2016), which works in successive steps to: (1) detect outliers, (2) smooth the values and (3) remove faults (see chapter 3 for more info). The main parameters of this method along with their default value and the applied modifications to the data are shown in Table 4-4.

Table 4-4: Parameters of the pre-processing method of Alferes & Vanrolleghem (2016) as applied in this study.

	Parameter	Description	Default Value (Alferes & Vanrolleghem, 2016)	Value applied
1	**Outlier Detection**			
	param.nb_s	Multiplicative factor that drives calculation of the prediction interval.	3	10
	param.nb_reject	Number of consecutive rejected data needed to reinitialization.	100	100
	param.nb_backward	reinitialization	15	50
	param.MAD_ini	Initial value Mean Absolute Deviation	10	10
	param.min_MAD	Minimum Mean Absolute Deviation	0	0
2	**Smoother**			
	h_smoother	Smoother coefficient	5	1
3	**Fault Detection**			
	paramX.range_min	Acceptable value (min)	NaN	case specific
	paramX.range_max	Acceptable value (max)	NaN	case specific
	paramX.moving_window	Range of Window	1000	5
	paramX.reading_interval	Size of Reading Interval	5	10
	paramX.corr_min	Correlation (min)	NaN	-1
	paramX.corr_max	Correlation (max)	NaN	1
	paramX.slope_min	Slope (min)	NaN	-1
	paramX.slope_max	Slope (max)	NaN	1
	paramX.std_min	Variation (min)	NaN	-1
	paramX.std_max	Variation (max)	NaN	1

This method was applied to the high frequent measurement signals available from the flow rates of Settled Water, Recirculation water, Returns of cleaned backwash water, the flow rates of the sludge treatment units, as well as the measured concentrations of COD, NOx-N and DO at the entrance of preDN; the measured concentrations of NOx-N and COD at the exit of preDN Battery C; the concentrations of NH_4-N, PO_4-P and TSS at the exit of the mixed channel preDN; and finally, the NOx-N and NH_4-N concentrations in the recirculation water streams. As an example, the intermediate results for the preprocessing of COD data are discussed below.

Figure 4-7 shows the raw data for the total COD concentration measured by an in-line sensor at the entrance of the preDN stage. This data contains outliers, faulty data and sensor drifts. Clearly discernible is the upper maximum detectable value of 400 mgO_2/l of this sensor.

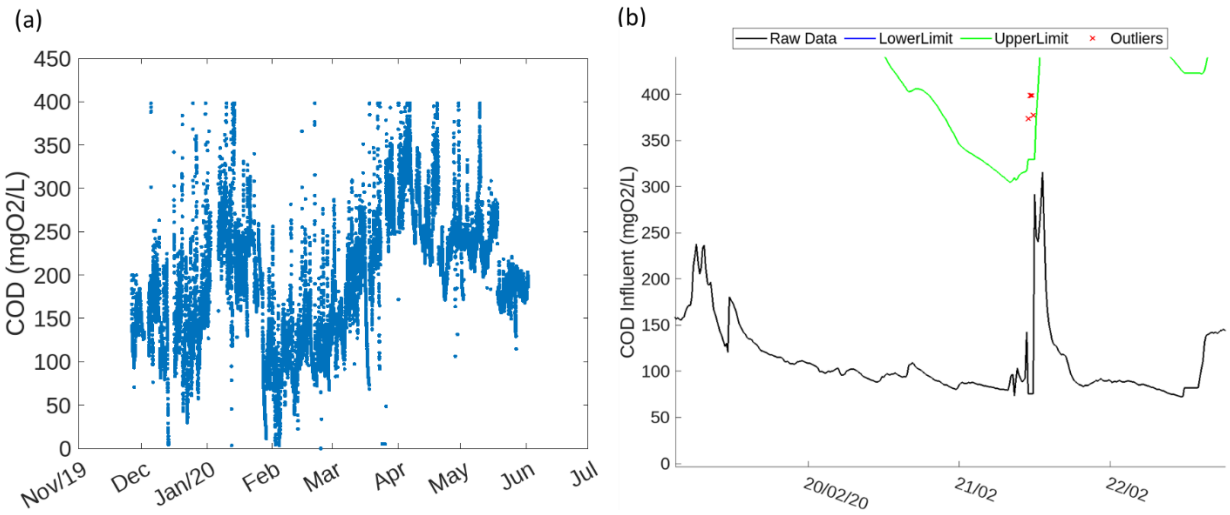


Figure 4-7 a) Raw COD measurement data from in-line sensor at the entrance of the preDN stage. b) Representation of the first step of the data cleaning procedure with indication of detected outlier values (marked by a red X) in a detailed view over a 4-days period.

Step 1: Data Cleaning

During the cleaning of the data, the first step concerns the detection of outliers, marked by the red coloured 'X' in Figure 4-7b, which shows a detailed view over a four day period of the same total COD data. Outliers are the points of data that lie beyond the upper limit (the green coloured line in figure b) and lower limits. The following parameters for the outlier detection method were modified:

1) the multiplication factor for the forecast interval 'nb_s' was set to a value of 15, which makes it a less stringent parameter, necessary due to the large width of sensed values (range from 0 to 400 mgO_2/l);
2) the number of consecutive rejected data points that lead to reinitialize the outlier detection method 'nb_reject' was set to 200 in order to avoid frequent reinitialization;
3) the number of data points used for calculation of forward application 'nb_backward' was set to 100, covering 24 hours.

Step 2: Smoothing

Data smoothing concerns the replacement of the identified outliers from the previous step with new estimated values and prevents the occurrence of data gaps, a prerequisite for the application of the fault detection. The data smoothing was achieved with parameter h_smoother value set to 2, which dictates the number of data points used in the calculation to smoothen a specific point. Low numeric values apply a strict (local) smoothing factor. The outcome of this step is indicated in Figure 4-8 afor the same detailed view as in the previous figure. Data points that fall 'outside' the upper and lower limits as determined by the smoothening algorithm, are replaced by their new smoothed value and are considered to be 'removed'. This will add to the score of the percentage of removed data.

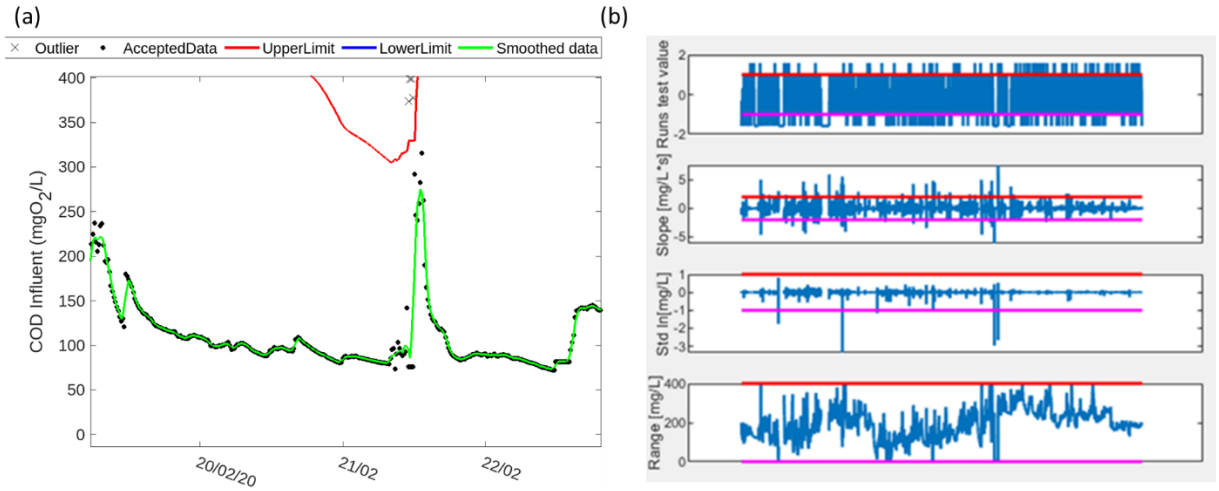


Figure 4-8a) Representation of the second step of the data cleaning procedure with indication of smoothed data (marked by a green line) in a detailed view over a 4-days period. b) Visual representation of the third step with Range, Standard Deviation, Slope and Runs Test upper (red line) and lower (purple line) limits for the COD total Influent data.

Step 3: Fault Detection

In order for the fault detection to function properly, the following parameters have been modified: the size of the 'window' for the moving test 'moving_window' was set to 5 datapoints; the range of acceptable values [1:400]; the minimum and maximum expected slope [-2:2]; reading_interval (10); the minimum and maximum expected correlation [-1:1] and the minimum and maximum variation between accepted and smoothed data 'std' [-1:1]. As can be seen in Figure 4-8b, the result in this third data cleaning step of the COD data indicate several extreme data points that fall outside the defined minimum (pink coloured line) and maximum (red coloured line) values for these parameters, indicating possible faulty data.

A final comparison with the daily laboratory data for each variable, which are assumed to be of a high data quality with a confidence interval of 10%, allows evaluating the performance of the data cleaning method as applied to the high-frequent measurements.

Final Results Data Cleaning

The final accepted outcome of the data cleaning procedure is a dataset that is assumed to consist of corrected, smoothed and fault removed data series, at least for the purpose of process modelling and control. An example is shown in Figure 4-9 for total COD.

This work was performed for each of the variables listed in Table 4-5 that shows the results for both the calibration period (column on the left side) and the validation period (column on the right). During the calibration period, the measurements contain 10 500 data points per variable, of which 4% classify as outliers. A high percentage of outliers was detected in the signals originating from the TSS sensor at the outflow of the preDN for both the calibration period (16%) and the validation period (5%). This might be caused by the monitoring techniques applied, which require regular maintenance of sensors, in particular under the harsh and highly variable conditions of wastewater treatment. The column entitled 'Percent Total Data Removed' shows the total percentage that includes the outliers detected, the data "lost" and replaced due to smoothing and the faulty data-points detected. Again, the TSS

sensor at the outflow of the preDN reports high percentages of removed data with 19% for the calibration period and 13% for the data from the validation period.

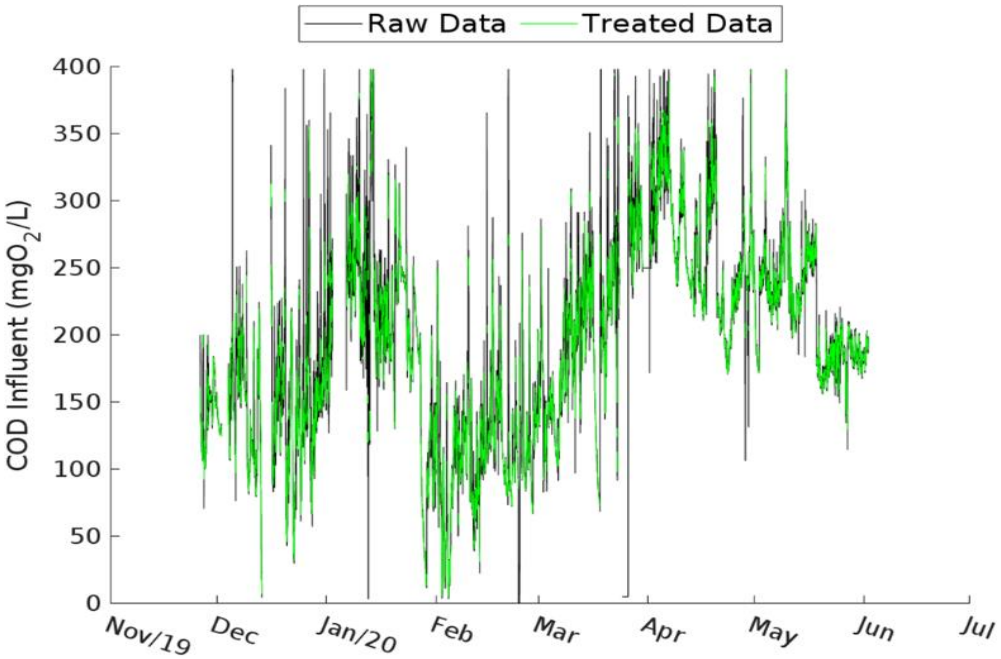


Figure 4-9: Results of the preprocessing of the raw measurement data (black line) of total COD at the entry of the preDN during the calibration period. The green line indicates the final corrected, accepted and smoothed data.

Concerning the data covering the calibration period, on average 16% of the raw data of the variables processed, are not retained, with high percentage losses for orthophosphates (48%) and total suspended solids TSS (33%) measured respectively at the inlet of the preDN and at the outlet of the sludge treatment, recovered during backwashing of the biofilters. In their study, Alferes and Vanrolleghem (2016) confirm typical losses in a wide range of 5 to 50% of data from probes submerged in municipal wastewater treatment streams.

Table 4-5: Overview of raw data processing for the calibration and validation period.

	Calibration			**Validation**		
Variable	N	Outliers (%)	Removed (%)	N	Outliers (%)	Removed (%)
Recirculated water flow	10657	0.5	8.1	8832	0.1	8.4
Settled water flow	10657	0.2	9	8832	0.1	5.3
COD inflow preDN	10036	1.9	11.1	8832	0.32	11.6
NOx inflow preDN	10231	1.3	3.5	8832	0.02	6.7
NOx recirculation	10654	0.2	7.6	8832	0.1	7.1
NH_4 outflow preDN	10573	0.3	6.9	8832	0.2	9.2
TSS outflow preDN	10654	16.1	18.6	8832	4.9	12.9
PO_4 outflow preDN	10626	1.4	48.3	8796	2.2	9.1
Sludge treatment flow	10652	0.06	13	8832	0.4	15.3
Cooling flow	10652	2.4	18.6	8832	0.7	9.3
NH_4 recirculation	10627	0.22	10.9	8831	1.2	19
TSS sludge treatment outflow	10656	0.04	33.4	8832	6.7	33.2
Mean	*10544*	*3.8*	*16.1*	*8829*	*1.4*	*12.3*

Regarding the validation period, on average 1% outliers were detected and 12% of data removed. The highest losses were experienced by the sensors measuring TSS at the preDN outlet (13%) and at the sludge treatment outlet. However, the data treatment for the validation period shows lower percentages than during the calibration for outlier detection (67% less) and for faulty data points removed (24% less). A good example is the PO_4-P sensor data that for the calibration data had 48% of the data removed, while for the validation data only 9% were removed. This is probably thanks to the more consistent and stable operating conditions with lower flow rates during the summer period (validation) than in the winter period (calibration).

This final accepted data series has data gaps, due to the third step of removing faulty data point during the data-preprocessing. This requires an additional data gap-filling technique by interpolation, such as a moving mean window, where missing values are replaced with average interpolated or smoothed values that are different from the observed values. In this study, a linear interpolation has been applied to replace missing data points. The cleaning of raw data by machine learning techniques could further be integrated with an influential generator model (not shown here) that allows gap filling in order to guarantee the quality of the data that feeds the model (Li & Vanrolleghem, 2022).

4.3 Preparation of Model Input

The model requires different input files that describe water quality variables and process control parameters. The input files contain the flow rates and the concentrations of the variables measured in function of the time; and that for each type of water flux: the settled water, the recirculated water and the treated backwash water. Furthermore, influent files are required that characterize the washing water used during the backwashing cycles. Process control variables are those that indicate the recirculation rate, set points and other constraints that can change over time.

4.3.1 Influent Characterization

The measured concentrations of nutrients and pollutants in the influent need first to be converted into state variables to be applicable in the biofiltration model equations. As much as possible, the measured variables serve as the base of this conversion into state variables following a fractionation scheme and by applying fractionation parameters.

The method for characterizing the influent variables is based on the fractionation for an ASM1 model (Henze et al., 2000). The fractionation parameters are shown in table 6 along with the default values from the ASM1 model. The organic matter fractionation model (see Figure 4-10a) was constructed from total COD and soluble COD measurements and/or TSS and converts these into six state variables for particulate matter (X_S: slowly biodegradable substrate; X_BH: active heterotrophic biomass; X_AOB: active ammonia oxidizing autotrophic biomass; X_NOB: active nitrogen oxidizing autotrophic biomass; X_I: particulate inert organic matter; X_P: particulate products from biomass decay) and two state variables for soluble compounds (S_S: readily biodegradable substrate and S_I: soluble inert organic matter). The state variables allow quantifying the evolution of the different organic and nitrogen compounds over time.

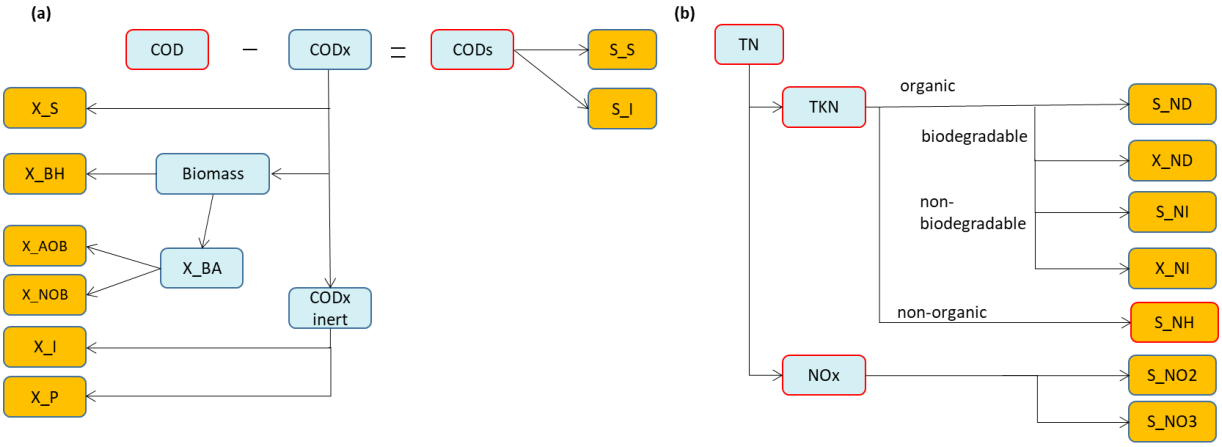


Figure 4-10: Fractionation scheme for a) Organic Matter and b) Nitrogen compounds based on measured variables (red outlined boxes) used to calculate the State Variables (orange boxes). Blue outlined boxes are derived variables.

The nitrogen fractionation (Figure 4-10b) was based on directly measured variables of Total Kjeldahl Nitrogen (TKN), NH_4-N and NO_x. The TKN variable, which combines total organic nitrogen and ammonia, was used to characterize the biodegradable and non-biodegradable nitrogen compounds; S_ND (soluble biodegradable organic nitrogen), X_ND (particulate biodegradable organic nitrogen), S_NI (soluble, non-biodegradable nitrogen), X_NI (particulate, non-biodegradable nitrogen) and S_NH (free and ionized ammonia). Concerning NOx-N, the measured sensor data was converted into nitrate nitrogen (S_NO3) and nitrite nitrogen (S_NO2) using the daily ratio provided by laboratory analyses.

The fractionation parameter values used to estimate the ratio of the components are shown in the lower section of Table 4-6. Values that have been significantly modified compared to the default value (Henze et al., 2000) for settled water, include the presence of biomass (factors 0.1; 0.3; 0.2) for settled, denitrified and nitrified water respectively in order to account for the biomass contribution in the water flowing out from each treatment stage. Both heterotrophic (X_BH) and autotrophic bacteria (X_BA) account for 90% and 10% respectively in the preDN influent.

Table 4-6: Parameters used for fractionation of organic matter and nitrogen in the different fluxes of Settled Water, Denitrified Water and Nitrified Water. Default values are from Henze et al (2000).

Fractionation parameter	Fraction Description	Default value Settled Water	Denitrified Water	Nitrified Water	Settled Water
Organic Matter					
f(S_S_CODs)	Readily biodegradable COD	0.68	0.25	0.15	0.68
f(X_S_CODx)	Slowly biodegradable COD	0.65	0.65	0.05	0.72
f(Biomass_CODx)	Biomass in particulate COD	0.0625	0.3	0.2	0.1
f(X_BH_Biomass)	Heterotrophic biomass	0.9	0.9	0.8	0.9
f(X_AOB_X_BA)	Autotrophic biomass	0.5	0.05	0.5	0.5
f(X_P_CODx_inert)	Inert particulate COD from decay	0.55	0.55	0.7	0.1
Nitrogen					
f(S_ND)	Soluble biodegradable organic nitrogen	0.8	0.8	0.7	0.7
i_X_Biomass	Mass of nitrogen per mass of COD in biomass [gN/gCOD]	0.086	0.086	0.086	0.086
i_X_P	Mass of nitrogen per mass of COD in products formed [gN/gCOD]	0.06	0.06	0.06	0.06

When reconstructing a state variable combining daily measurements with 15 minutes time steps, one must be cautious to avoid negative values, such as can happen in the case of the fractionation of TKN into S_ND and X_ND, which is defined as the subtraction of 1 – (S_ND). In order to prevent such negative values, the values of the state variables have been adjusted with the laboratory observations. This was done by multiplying the 15-minutes measurements with the daily ratio of the laboratory observation and the average of the daily sensor measurement values.

4.3.2 Definition of Process Control Parameters

In the multistage treatment configuration, the model connects the simulated output of the denitrification (preDN) stage with the input to the nitrification (NIT) stage; and the output of the NIT to the input of the postDN. Not only does a 'physical' connection has to be established between the biofilters in the model, also process control settings related to the flow distribution, desired set points of quality variables and chemical dosing are required as input. For more information about the system architecture for this 'three-filters-in-series-model', refer to chapter 3.

The control parameters required for modelling shown in Table 4-7 are calculated based on measurement data for the relevant period and the corresponding treatment phase.

Concerning the preDN stage, the input parameters related to flow distribution are the number of filters available (NpreDN) [number], the rate of recirculation (Ratio_rec) [%], defined as the flow rate of recirculation divided by the sum of the flow rate of primary settling and the return flow, and the maximum allowable filtration flow rate (Vmax_preDN) [m/h] taken from measurement data. In addition, the model determines the minimum number of biofilters that are needed based on the maximum volumetric load (Charge_max_preDN) [kgN/(m^3*d)] of nitrates and nitrites at the entry of the preDN stage, which has been taken from the meta-data stored in the databank. For the methanol dosing, the input parameters are the NO_3 [g/m^3] set point at the outflow (info received from the operator) and the ratio of methanol [g] added per NO_3 [g] removed (ratio_methanol_preDN) [-], taken from the constructor's manual.

Table 4-7: Process control parameters for the preDN stage.

Control Parameter and unit	**Description**	**Value**
NpreDN [-]	Number of filters available	57
Ratio-rec [%]	Recirculation rate	0.6
Vmax_preDN [m/h]	Maximum allowable filtration flow rate	6.5
Load_max_preDN [kgN/m^3/d]	Maximum NO_X volumetric load	580
Set point_NO_3 [g/m^3]	NO_3 setpoint	1.3
Ratio_methanol_preDN [-]	Ratio of methanol [g] added per NO_3-N [g] removed	0.75

4.4 Parameter Estimation

Model calibration involves the identification of parameters that have a strong impact upon the output variables and the estimation of their best fitting values. Parameter estimation is defined as the determination of the optimum values of the model parameters in relation to an objective function (i.e. the RMSE of the difference between observed and predicted values). These optimal values arise in a mathematical description with the aid of experimental data, assuming that the structure of the process

model, in other words the relationships between the variables and the parameters, is known (Dochain & Vanrolleghem, 2001).

The calibration of the dynamic biofiltration model followed the method described by Mannina et al. (2011), which is based on the IWA Modelling Guidelines for Activated Sludge Models for the Use of Activated Sludge Models (Rieger et al., 2012). Models that contain a large number of parameters make it impossible to find manually, by trial-and-error, the best fit between a parameter value and the output variables due to the large variety of possibilities and number of trials required as well as due to the interdependencies among parameters. It is therefore important to identify beforehand which parameters have the most impact and to what degree they influence the values of selected output variables. In order to reduce the number of parameters to be calibrated and to better understand their impact, a global sensitivity analysis is performed (Sin et al., 2011).

Considering the large number of parameters (142) and variables (20) in the model used in this study, as well as the uncertainty of a precise value associated with them, a global sensitivity analysis (GSA) was completed by applying the standardized regression coefficients (SRC) method (Cosenza et al., 2013) instead of a local sensitivity analysis where only local differences in parameter values are investigated.

The sensitivity analysis consisted of the following main steps:

Step 1: Selection of the model parameters relevant for the output variables that relate to the denitrification process evaluated, which are nitrogen and phosphates removal, the consumption of organic matter and the filtering of suspended solids. A pre-selection of parameters based on expert opinion and literature review reduced the number to 89 parameters. Those related to the physical structure (such as, for example, the dimensions of the filter) were fixed at the measured values and thus excluded from the analysis.

Step 2: Execution of Monte Carlo simulations under steady state conditions with a Latin Hypercube Sample (LHS) design of the selected parameters. The results of the simulations are analysed by a multilinear regression. Data normalization, necessary to provide uniformity and accuracy, has been applied by applying the z-score, with a mean $\hat{y}$ and a standard deviation σ for each x value, while maintaining the dataset shape properties.

Step 3: Identification of the parameters that have a significant effect on each of the selected variables by means of an evaluation of the standardized regression coefficient (SRC) of each parameter and grouping them into independent output variables that share the same parameters.

The calibration then continues per independent variable group (conform Mannina et al., 2011), as will be described in paragraph 4.4.3.

4.4.1 Initial Parameter Values

The initial values are based on the previously performed calibration work of Zhu (2020) for the three treatment stages individually. As a reminder, the calibration of the preDN was performed with data from Seine Grésillons (Biostyr, year 2015); the NIT stage with data from Seine aval (Biostyr, 2009) and the postDN stage with data from Seine Centre (Biofor, 2008). Although these plants differ from the Seine-aval WWTP, especially in the volume of water received, the treatment methods applied are similar so that these model parameters can be considered as good initial values.

4.4.2 Identification of Influential Parameters

In order to reduce computation time, the global sensitivity analysis consisted of 10 parallel sessions of 90 simulations, each regenerated with a different seed number in the random generator. A total of 900 steady-state simulations for a period of 30.5 days were run using Monte Carlo simulations and the Latin Hypercube Sampling (LHS) method, which allows for stratified random sampling of the parameter values covering the entire range set requiring a smaller sample size. At the time of 30.5 days, the variation in modelled variable values at the outlet of preDN is assumed to be entirely caused by the chosen parameter values. In this section, the identification and selection of parameters is discussed. A comprehensive evaluation of the performance of the quality of model fits, including the plotting of graphs, will follow in section 4.5.

According to Brun et al. (2002), the sampling range for parameters that are considered well known in literature and describe biofiltration processes that are well understood was set to the range of - 20% to + 20% of their default value, while for lesser-known parameters the range was set to +/- 50% in a uniform distribution. The default values refer to the values of the parameters defined in the previously calibrated model (Zhu, 2020).

Considering the variables at time t = 30.5 days, a multilinear regression analysis allows identifying the existence of a strong relationship between the parameter and the variable, indicated by the standardized regression coefficient (SRC) which varies between the values 1 (strong positive correlation), 0 (no correlation) and -1 (strong negative correlation). The SRC method is applied here in terms of factors prioritization (Cosenza et al., 2013). The absolute SRC scores are divided into three classes (see Table 4-8) to indicate their importance (Bernier, 2014).

Table 4-8: Importance levels of the Standardized Regression Coefficients (Bernier, 2014)

Absolute SRC score	Level of importance of the SRC
< 0.1	low
0.1 < SRC < 0.25	significant
> 0.25	strong

The coefficient of determination (R^2) indicates the fraction of total variance explained in the operational data by the regression model, and is as such a quality indicator. R^2 values larger than 0.7 point to a good validity of the SRC coefficients (Cosenza et al., 2013). As listed in Table 4-9, the variables have a calculated determination coefficient between 0.86 and 0.92, except for the NO_2-N effluent variable, which has a R^2 value of 0.64, slightly below the limit of 0.7, but still higher than the R^2 value obtained during the global sensitivity analysis of the original model (Zhu, 2020). The lower value for nitrite can be explained by the fact that it is difficult to measure correctly.

Table 4-9: Determination coefficients in descending order for the effluent variables as calculated during the global sensitivity analysis of the selected model parameters.

Effluent variable	Description of effluent variable	Determination Coefficient (R^2)	R^2 value original model (Zhu, 2020, p. 438)
SNH	Ammonia NH_4-N	0.93	0.94
TKN	Total Kjeldahl Nitrogen	0.92	0.90
TSS	Total Suspended Solids	0.92	0.87
COD	Chemical Oxygen Demand	0.91	0.80
NO3	Nitrates NO_3-N	0.91	0.82

CODs	Soluble compounds of COD	0.87	0.66
PO4	Orthophosphates PO_4-P	0.86	0.91
NO2	Nitrites NO_2-N	0.64	0.61

Overall, the validity of the SRC as a parameters ('factors') prioritization tool that provides information about the important factors (parameter prioritization) and non-influential factors (parameters fixing) is considered to be accepted.

The SRCs for two important variables groups that relate to the denitrification process by heterotrophic bacteria under anoxic conditions are now discussed in more detail. The results of the global sensitivity analysis for all effluent variables are listed in Annex 3.

NO_3-N Effluent

The outcome for the NO_3–N variable, which is the most relevant variable in assessing the performance of the denitrification process, is shown in Table 4-10. Positive SRC values indicate that an increase of the parameter value will result in an increase of the value of the variable (here: NO_3-N), while a negative SRC indicates that the variable value decreases. In fact, extreme SRC values on either side indicate a large influence.

The influential positive and negative parameters with SRC larger than 0.1 are related to the growth and decay of the heterotrophic biomass, the hydrolysis of substrate and the thickness of the biofilm. Particularly high and positive values are found for the parameter WashRate (SRC +0.36), which controls the fraction of biofilm removed during the backwashing cycle. When the WashRate increases, a larger percentage of the active biofilm is removed so that there is less bacteria at the start of the next treatment cycle remaining to convert the nitrates in the bulk liquid. Therefore, higher NO_3-N values in the effluent are to be found.

Another important positive parameter influencing the NO_3-N variable is the yield correction factor for the bacteria under anoxic conditions (n_hy) with a SRC of +0.31. In ASM1 models (Henze et al., 2000), a single yield (Y_H) value is used in the aerobic and anoxic growth equations of heterotrophic bacteria, but in the absence of dissolved oxygen a lower bacterial yield needs to be applied. Choubert et al. (2009) offer two different values for anoxic and aerobic conditions respectively: Y_H_anoxic = 0.54 gCODproduced/gCODremoved (anoxic growth equation) and Y_H_aerobic = 0.67 gCODproduced/gCODremoved (aerobic growth equation). In the current model, when the n_hy parameter value increases to 1, this will lower its impact and reduce the denitrifying behaviour of the heterotrophs, thereby resulting in higher NO_3-N effluent concentrations. A third parameter, K_S_NO_3 has a SRC of +0.21, which is the half-saturation constant of soluble COD for the bacteria using nitrate, has a similar strong effect; higher parameter values slow down the consumption of nitrates and thereby leave higher concentrations in the effluent.

Table 4-10: Parameters that have been identified by the global sensibility analysis as significant and strong influential upon the effluent NO_3-N variable at the preDN denitrification stage.

Ranking	Parameter	Description	Standardized Regression Coefficient
1	WashRate	Fraction of biofilm removed during washing	0.37
2	n_hy	Anoxic correction factor of heterotrophic yield	0.31

3	n_h1	Anoxic heterotrophic growth rate reduction factor using nitrates	-0.30
4	k_h	Hydrolysis rate	-0.28
5	theta_k_h	Temperature correction factor for hydrolysis rate	0.22
6	theta_b_H	Temperature correction factor for heterotrophic mortality rate	-0.21
7	K_S_NO3	Anoxic half-saturation constant of soluble COD for heterotrophs using nitrates	0.21
8	theta_mu_H	Temperature correction factor for maximum specific heterotrophic growth rate	0.20
9	n_h	Anoxic hydrolysis rate reduction factor	-0.18
10	lami	Filtration coefficient of a clean filter	-0.15
11	b_H	Heterotrophic mortality rate	0.12

On the negative side of the influential parameters, biological growth reduction parameters are found, such as n_h1; when the value of this parameter goes up, the growth of the bacteria slows down, thus converting less nitrates, resulting in higher NO_3-N effluent concentrations. In addition, parameters that control the hydrolysis process – such as k_h (SRC -0.28) and n_h (SRC -0.18) are highly influential. Carbon uptake and the availability of organic carbon are restricted by the rate of hydrolysis of slowly biodegradable (particulate) organic matter (Morgenroth et al., 2002). Higher values for the hydrolysis rate (k_h) will mean more uptake of organic matter by bacteria, which can then convert more nitrates from the water resulting in lower nitrate effluent concentrations. The same influence applies to the reduction factor of the hydrolysis rate to compensate for anoxic conditions (n_h); higher n_h values result in less compensation, and thus higher bacterial yields, thereby lowering the nitrate effluent concentrations. Another important factor can be found in the temperature relationships; parameters theta_b_H (SRC -0.21), theta_k_h (SRC +0.22) and theta_mu_H (SRC 0.20) are temperature coefficients that affect the biological reactions when temperatures in the water rise or drop.

COD soluble Effluent

The Chemical Oxygen Demand (COD) soluble effluent variable, which is the total of readily biodegradable substrate (rbCOD) together with the inert fraction (S_I), can be used as a representation of the biological activity. When the biomass in the filter is highly active, it uses large portions of rbCOD, which lowers the soluble COD concentration in the effluent. Parameters that control the heterotrophic growth rate via a temperature correction (theta_mu_H) and anoxic reduction correction (n_h1) have indeed high SRC scores of +0.27 and -0.23 respectively indicating a strong influence. The parameter K_S_NO3, that controls the anoxic half-saturation constant of soluble COD for heterotrophic biomass using nitrates, has a SRC of +0.14, indicating a significant influence over the COD soluble effluent variable. Table 4-11 lists the parameters with SRCs higher than 0.1.

Table 4-11: Parameters that have been identified by the global sensibility analysis as significant and strong influential upon the effluent soluble COD variable at the preDN stage.

Ranking	Parameter	Description	Standardized Regression Coefficient
1	theta_mu_H	Temperature correction factor for specific maximum heterotrophic growth rate	0.27
2	n_h1	Anoxic heterotrophic growth rate reduction factor using nitrates	-0.23

3	K_S_NO3	Anoxic half-saturation constant of soluble COD for heterotrophs using nitrates	0.14
4	theta_b_H	Temperature correction factor for heterotrophic mortality rate	-0.13
5	k_h	Hydrolysis rate	0.13
6	mu_H	Maximum specific heterotrophic growth rate	-0.12
7	theta_k_h	Temperature correction factor for hydrolysis rate	-0.11
8	n_h	Anoxic hydrolysis rate reduction factor	0.11
9	n_hy	Anoxic correction factor of heterotrophic yield	0.11

TSS Effluent

When taking the modelled Total Suspended Solids (TSS) concentration in the effluent of the preDN stage into consideration, the physical filtration of a filter is the principal process involved. This is represented by those parameters that describe filter characteristics. The global sensitivity analysis did indeed identify the following parameters as the most influential with a SRC larger than 0.1 (see Table 4-12): the filtration coefficient ('lami'), which is the filtration coefficient of the filter bed of a cleaned filter after each washing cycle, the fractionation parameter ('F_TSS_COD') that determines the concentration of suspended solids (TSS) in the water based on the concentration of particulate COD, the y-coefficient in the Ives equation ('yy'), the WashRate, the z-coefficient in the Ives equation ('zz') and the biofilm density ('rho'). The parameters 'lami', 'yy' and 'zz' belong to the filtration 'Ives-equation' and each describe a phenomenon influencing filtration efficiency, which is directly negatively affected by an accumulation of particulates on the filter bed.

Table 4-12 : Parameters that have been identified by the global sensibility analysis as significant and strong influential upon the effluent TSS variable at the preDN stage

Ranking	Parameter	Description	Standardized Regression Coefficient
1	lami	Filtration coefficient of a clean filter	-0.44
2	F_TSS_COD	Fraction of TSS from particulate COD	0.22
3	yy	Second term in Ives equation	-0.20
4	WashRate	Fraction of biofilm removed during washing	-0.14
5	zz	Third term in Ives equation	0.14
6	rho	Biofilm dry density	-0.13
7	xx	First term in Ives equation	0.07

4.4.3 Grouping of Parameters to be calibrated per Process

Following the procedure described by Mannina et al. (2011), in order to improve parameter estimation, the global sensitivity analysis (GSA) results were structured in a sequential order of priority relating to the process evaluated. Here, the processes considered are nitrogen and phosphates removal, the consumption of organic matter and the filtering of suspended solids.

The first category of variables focuses on the particulate components (represented by TSS) in order to cover the physical properties of filtration; thereafter the focus was on the soluble substrate variables for the biological reactions of the denitrifying microorganisms using organic matter (soluble COD) and nitrates (NO_3-N). Group 2 thus contains both soluble COD and nitrates combined, because they both share many of the same and interlinked parameters given the metabolic kinetics of the biomass that describe the use of soluble COD (the substrate for the biomass) to convert nitrates into nitrogen gas.

Group 3 (NH_4-N), group 4 (NO_2-N) and group 5 (PO_4-P) contain individual variables that are of a lesser concern to denitrification, but could still provide valuable information about the parameter settings.

4.4.4 Calibration of selected Parameters

The calibration per variable group consisted of Monte Carlo Simulations with 30-day dynamic simulations that were each time preceded by steady-state simulation for a 30-day period to initialize the model. The parameter ranges used in the Monte Carlo-based optimization (Sin et al., 2011) were sampled by the LHS method with five times the number of parameters as the total number of Monte Carlo runs.

Evaluation of statistical scores for mean error (ME), mean absolute error (MAE) and the root mean square error (RMSE) leads to find the optimal parameter value. In this study, optimal values are determined by minimizing the RMSE. The modelling guidelines maintain that an error range of 10 to 15% is acceptable for ASM models (Rieger et al., 2012), which coincides with the confidence interval of measurements. Furthermore, Rittmann et al (2018) present a framework for good practice in modelling biofilm reactors, in which they indicate reasonable ranges of values for several variables. However, in the absence of a widely accepted and consistent protocol with specific criteria for biofilm models, statistical error values of 15% are considered acceptable (Rittman et al. al., 2018). Once the best model parameter values are found for one group of parameters, they are fixed and the calibration continues with the parameters belonging to the following group. Finally, a validation of the model parameters is performed over the full period to allow for a visual inspection and a statistical analysis of model performance and for the manual fine-tuning of parameters.

An overview of the model parameters that were recalibrated is given in Table 4-12 together with the default values taken from the original model (Zhu, 2020) and their relative differences. The modified parameters cover the full range of processes and characterize filtration, backwashing, the transfer of material within the biofilm by diffusion, the nature of the biofilm and the biological kinetics.

Table 4-12: Overview of the estimates of the influential parameters of the mechanistic model recalibrated during this study.

Description of parameters	Symbol	Unit	Original value (Zhu, 2020)	Range assigned	New value	Difference
Composition						
Mass of nitrogen per mass of COD in biomass	i_X_B	gN/gCOD	0.098	20%	0.109	+11%
Mass of phosphorus per mass of COD in biomass	i_X_BP	gP/gCOD	0.0055	50%	0.0046	-16%
Mass of phosphorus per mass of COD in products formed by bacterial decay	i_X_UP	gP/gCOD	0.006	50%	0.008	+33%
Filtration						
Filtration coefficient of clean filter	lami	m^{-1}	0.4	50%	0.43	+8%
Empirical constant y in the filtration coefficient equation	yy	-	0.83	50%	1.06	+28%
Kinetics						
Anoxic half-saturation constant of soluble COD for heterotrophs consuming nitrites	K_S_NO2	$gCOD/m^3$	13.8	20%	16.06	+16%
Anoxic half-saturation constant of soluble COD for heterotrophs consuming nitrates	K_S_NO3	$gCOD/m^3$	28.6	20%	41.37	+45%

Mortality rate of heterotrophic biomass	b_H	d^{-1}	0,7	20%	0.73	+4%
Ammonification rate	k_a	m^3/ (gCOD*d)	0.075	20%	0.065	-13%
Hydrolysis rate	k_h	gCOD/ (gCOD*d)	2.5	50%	2.6	+4%
Maximum specific growth rate of heterotrophic biomass	mu_H	d^{-1}	5.18	20%	4.74	-8%
Anoxic growth rate reduction factor of heterotrophic biomass consuming nitrite	n_h2	-	0.14	50%	0.19	+36%
Factor for reducing the rate of growth in anoxia of heterotrophic biomass by consuming nitrate	n_h3	-	0.47	50%	0.31	-34%
Anoxic Heterotrophic Yield Correction Factor	n_hy	-	0.78	50%	0.72	-8%

Compared to the values of the original model, the parameters change on average by +21%. This could indicate a difference in conditions for the bacteria and their metabolisms between the current calibration period and the one carried out previously in the study by Zhu (2020), which used data from another but similar WWTP. Important modifications are proposed to the parameters describing the activity of the heterotrophic biomass under anoxic conditions, among these the reduction factors of the growth rate of heterotrophic biomass reducing nitrites (n_h3: -34%), that of reducing nitrogen monoxide (n_h2; +36%) and the half-saturation constant of soluble biodegradable COD for heterotrophic biomass consuming nitrate (K_S_NO3: +45%). Overall, 9 modified parameters belong to the category of biological kinetics and 2 parameters to filtration; what is expected in a model of biofilm in a filter. See Appendix 3 for a complete list of all model parameters.

Realism of simulated Thickness of Biofilm and Boundary Layer

Two other indicators of the performance of a biofilm model are the level of realism of the simulated thickness of the biofilm and the Mass Transfer Boundary Layer (MTBL), which separates the biofilm from the bulk liquid. The MTBL thickness is a representation of the resistance of transport and transfer of particulate and soluble matter from the bulk liquid towards the biofilm (Rittmann et al., 2018).

Earlier biofilm models set the thickness of the biofilm and the boundary layer to a constant value and used a variable biomass density to account for the varying biomass amounts in the biofilm; considering these to be a structural model parameter. However, the biofiltration model used in this study (Zhu, 2020) contains a variable thickness of each of the five biofilm layers that is recalculated per time step in function of the mass of particulates that are 'entrapped' in it, and in function of microbial growth. In addition, the thickness of the MTBL is also time-variable to be able to describe the changing resistance on the diffusion gradient from the bulk liquid into the film. The thickness of the boundary layer is calculated according to the Ohashi equation (1981) as the quotient of the geometrical characteristics of the filtering media and the Sherwood Number that represents the mass transfer ratio, which itself is a function of the flow rate (Reynolds Number) and the fluid characteristics and diffusion coefficients (Schmidt Number) (Morgenroth, 2008).

Comparison of the average biofilm thickness in this model with the typical values mentioned in literature is a good practice. Thickness should remain within acceptable limits and should decrease in function of the height of the reactor column in an upward flow directional biofilter. The highest organic loads are to be found at the entrance of the filter; where a thicker biomass will develop (Vigne, 2010).

According to literature review (chapter 2), the acceptable values reported are 80 – 120 µm for the biofilm thickness (Rittman et al., 2018) and around 100 µm for the MTBL thickness (Morgenroth, 2008). In the previous calibration work by Zhu (2020), the average simulated thickness was 95 µm and 97 µm for the biofilm and boundary layer thickness respectively in the predenitrification stage.

In this study, the averaged simulated biofilm thickness was 83 µm with a minimum average of 38 µm (the moment just after backwashing) and a maximum average thickness of 139 µm for the first section of the filter column; values that fall in or are close to the acceptable range. The average MTBL boundary layer thickness was 84 µm; which is in the same order of values reported in literature (100 µm) and about 13% less than the average MBTL thickness in the original model by Zhu (2020). This could indicate an overall lower resistance to the transport and transfer of compounds into the biofilm.

4.5 Simulation of Denitrification by Biofiltration

This paragraph discusses the results of the simulations of effluent water quality by the mechanistic dynamic biofilter model for the predenitrification stage using data of primary effluent and recirculation of nitrified water from the calibration period from 01.12.2029 until 15.03.2020.

4.5.1 Calibration Results Denitrification Stage

Table 4-13 presents the values of statistical criteria for the simulations of the preDN effluent components during the calibration period. The relative mean error ME/ȳ values indicate that the bias of the model is low compared to the average values observed, in particular for the total COD, soluble COD and TSS variables, in particular when keeping in mind the laboratory measurement error that typically have a 10% error range.

In the case of NO_3-N, the ME/ȳ indicator has larger values (-0.47 and 0.32 respectively), indicating that the model overestimates the biological conversion rates, so that too much nitrates get converted in the model and the effluent contains too low NO_3-N concentrations. This could be related to inaccuracies with the estimation of dissolved oxygen (DO) concentrations in the influent water. In fact, high DO has an inhibitory effect on denitrification by favouring aerobic metabolism of heterotrophic bacteria (Henze et al., 2008). However, the data of the DO probes have a high percentage of outliers and suppressed data (28%), which during the data-preprocessing are replaced with average interpolated or smoothed values that may be lower than observed.

Table 4-13: Statistical scores for calibration and validation of the mechanistic model for the modelled variables at outlet preDN.

		Calibration					Validation					
Variable	Unit	N	Mean obs.	ME/ ȳ	MAE /ȳ	RMSE/ ȳ	N	Mean obs.	ME/ȳ	MAE /ȳ	RMSE/ ȳ	Janus coef
NO_3-N	gN/m^3	10656	2.89	-0.47	0.67	0.86	8015	2.62	-1.23	1.29	1.46	1.53
NO_2-N	gN/m^3	10656	0.31	0.32	0.45	0.71	8015	0.38	0.34	0.5	1.27	1.36
totCOD	gO_2/m^3	10656	89.07	-0.01	0.14	0.20	7218	64.84	-0.53	0.55	0.6	2.20
solCOD	gO_2/m^3	10656	43.3	0.04	0.14	0.21	7218	43.92	-0.35	0.39	0.45	2.13
TSS	g/m^3	10656	21.2	-0.11	0.23	0.39	7399	20.12	-0.36	0.46	0.63	1.54

Figure 4-11 shows the calibration results for the effluent nitrate concentration for the month of December 2019 in more detail. In Figure 4-11a, the effluent nitrate concentration modelled at high frequency (blue line plot) follows the values measured in the laboratory (the brown triangles) and corresponds to the observations. The modelled average daily values amount to 4.7 mg NO_3-N/L on average for this period, while the average nitrate concentration observed in the effluent was 5.3 mgNO_3-N/L for the month of December. The evaluation in terms of statistical scores for this period shows an ME of 0.67 mgNO_3-N/L, an MAE of 1.08 mgNO_3-N/L and an RMSE of 1.36 mgNO_3-N/L, all values less than 25% of the observed average. The parity plot (figure b) of daily observations against simulations indicates the correlation between the two. One can conclude that the model is capable of simulating the conversion of NO_3-N to N_2; especially for effluent values in the range 4 – 6 mgN/L. Higher values [6 – 10 mgN/L] tend to be underestimated by the model. The R^2 value of 0.41 for the model predictions indicates a correlation with a high uncertainty.

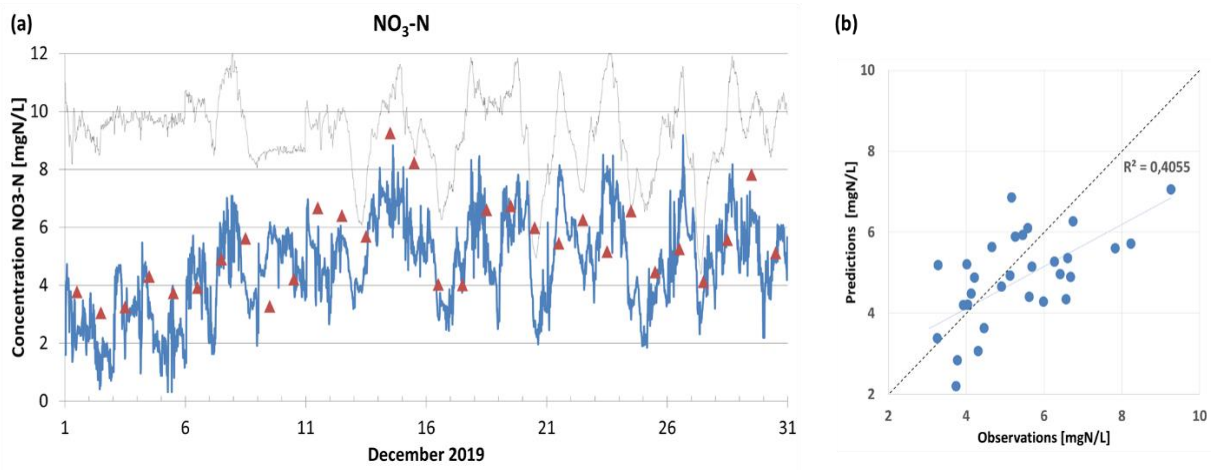


Figure 4-11: a) Calibration results of the NO_3-N concentration in the effluent of the preDN (brown triangles: daily composite laboratory observations, blue line: simulation results, black line: NO_3-N influent). b) Parity plot showing observed relative to daily average simulated nitrate concentrations.

Concerning the results of the modelled nitrite (NO_2-N) concentration in the preDN effluent, the modelled values are on average (0.12 mgN/L) approximately twice lower than the average of the observations (0.26 mgN/L). This difference is reflected in the statistical scores: ME (0.14 mg N/L), MAE (0.14 mgN/L) and RMSE (0.18 mgN/L). Simulating nitrite concentrations is difficult because nitrite is an unstable compound that quickly oxidizes to nitrate. In addition, the low levels here observed often approach the detection limits of the probes. This metrological difficulty is therefore reflected in the quality of the model's results.

Figure 4-12 shows the results of total COD (a), soluble COD (c), TSS (e) and NH_4-N (g). The simulations, in particular for total COD and soluble COD, vary around and with the same amplitude as the observations most of the time. This is confirmed by the parity plots for COD total (b) and COD soluble (d), which show points of daily average modelled versus daily composite sample observations in close proximity to the optimum with R^2 values of 0.67 and 0.65 respectively.

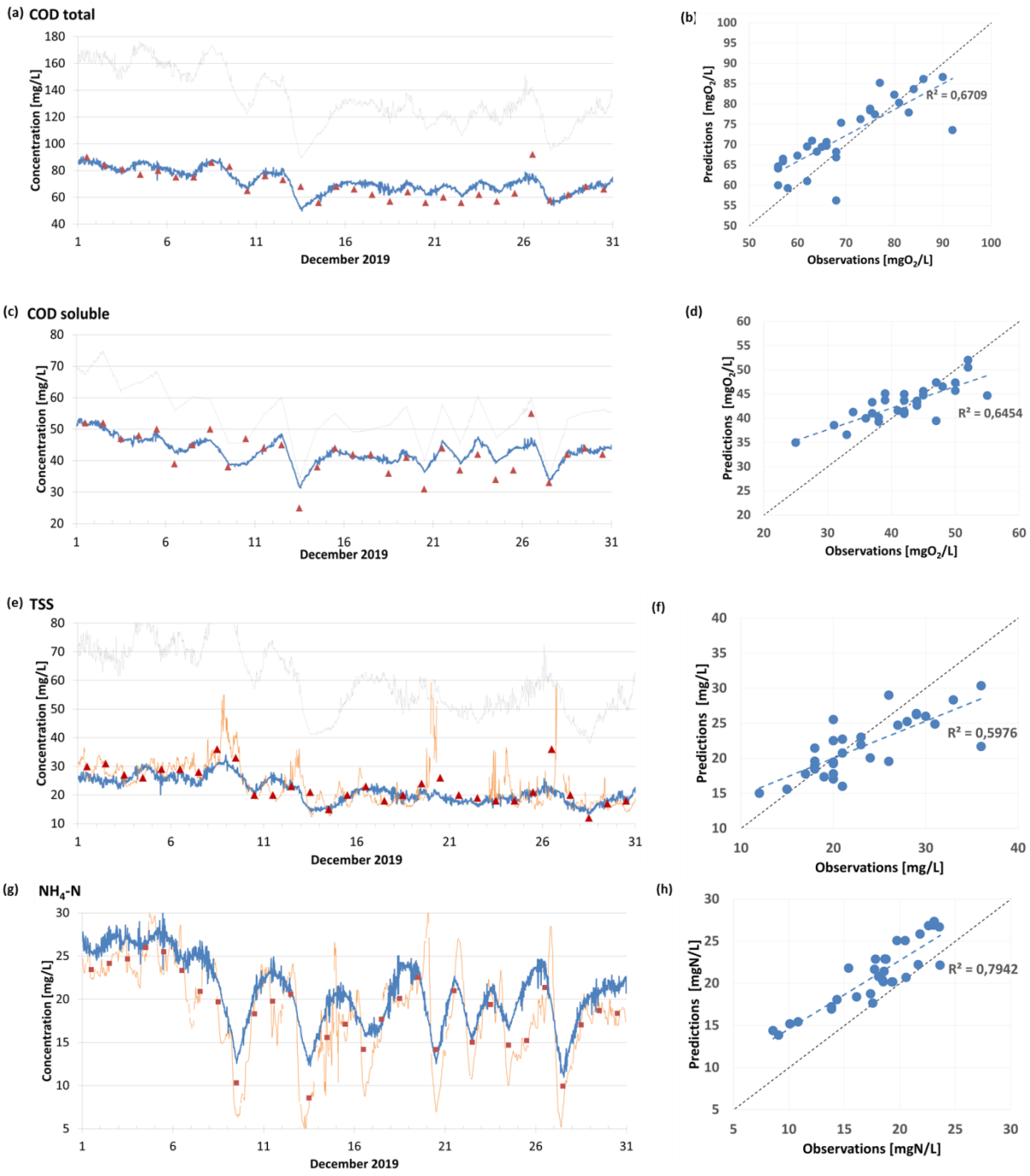


Figure 4-12: Calibration results at the outlet of the preDN for concentrations of a) total COD, c) soluble COD, e) TSS and g) NH_4-N. Parity plots shown on the right side are for b) total COD, d) soluble COD, f) TSS, h) NH_4-N. (Brown triangles: composite sample laboratory observations, brown line: sensor observations, brown squares: daily average of sensor observations; blue line: simulation results, black line: influent.

Comparison for TSS (c) with the high-frequent measurements shows less precise results, sometimes with significant deviations (on days December 8, 19, 26). This could be due to the TSS in-line analyser, which requires intensive maintenance to keep it clean and calibrated, or to specific conditions during these periods of time that did not correspond to the rest of the calibration period. The ME and MAE for TSS are 2.3 mg/L and 4.9 mg/L respectively to the sensor data with an observed mean of 21.2 mg/L. The laboratory results however give a much better fit when compared to the average daily simulated TSS effluent with R^2 value of 0.60 (f).

Finally, for ammonia (NH_4-N), the simulations follow the trend and amplitude of the observed values, but are consistently overestimated with a ME of 2 mgN/L. Although ammonia is not transformed during the denitrification stage, it is assimilated by the biomass for its growth and can be 'produced' by hydrolysis/ammonification of organic nitrogen, such as protein, urea, and certain components of cells (like nucleic acids). When bacteria die, the nitrogen they contain is liberated as NH_4-N into the bulk liquid. The model is able to simulate this ammonia uptake and subsequent release well, resulting in a high R^2 value of 0.79 (figure h).

4.5.2 Validation Results Denitrification Stage

This paragraph discusses the results of the simulations for the validation period using as influent data those from the period 01.06.2020 until 30.09.2020. A good validation is defined by a good generalization of the model parameters. In other words, when a model can satisfactorily simulate effluent concentrations measured in a period not used for calibration. The Janus coefficient allows evaluating the model performance outside the calibration data range. With an expected optimal value of 1, the model is nevertheless considered valid with a Janus coefficient of 2 (Hauduc et al., 2015), meaning that the RMSE during the validation period should not be more than twice the one during the calibration period.

The results for the validation period (Figure 4-13) show the concentrations of the variables NO_3-N (a), total COD (c) and soluble COD (e) at the preDN effluent for the modelled values (blue coloured line) and observations (brown triangles). The predicted NO_3-N concentrations are generally overestimated and show significant deviations from the observations, in particular from the beginning of August, which results in a weak parity diagram (Figure 4-13b). Explaining this may be the lower flow arriving at the plant and the drastic drop in or a change in the composition of organic matter load during the month of August due to a reduction in (human) population in the city during the summer period. This could provoke a reduction in the denitrification rate.

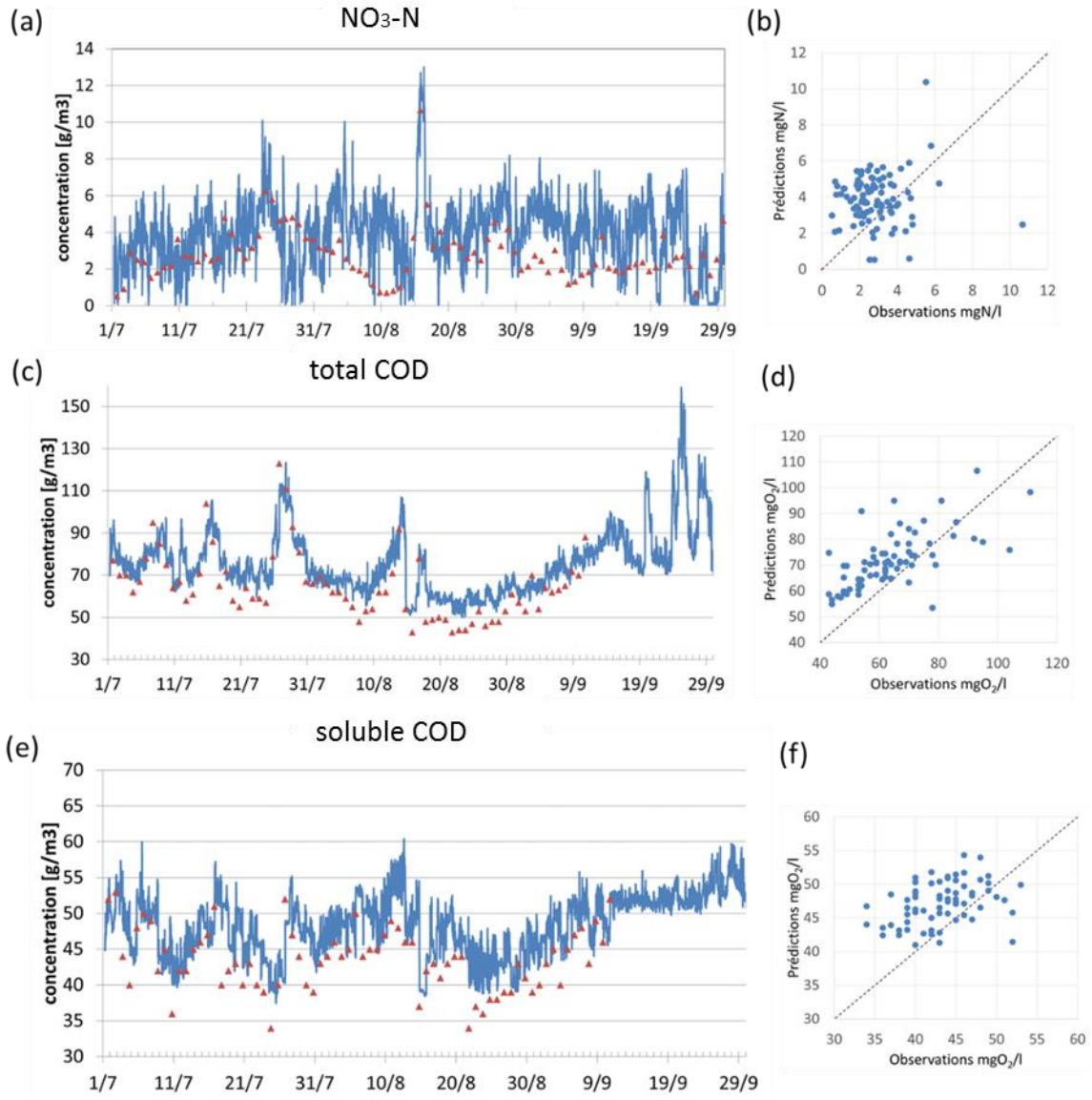


Figure 4-13: Validation results at the preDN effluent for a) the NO_3-N concentration; c) total COD and e) soluble COD. Parity plots show the modelled values in function of the observations for b) NO_3-N, d) total COD and f) soluble COD. Brown triangles: laboratory observations, blue line: simulation results.

The estimates of total and soluble COD are generally quite satisfactory, but are systematically overestimated (figures c to f). Similarly, the actual effluent contains fewer nitrates than simulated by the model, in particular during the period from the end of July to mid-August 2020. This overestimation of COD and nitrates at the preDN effluent can be explained by an insufficient metabolic activity of the biomass in the model during the validation period.

Many other factors can introduce simulation errors, such as the precision of the measured values (especially for in-line sensors which show large uncertainties) and the fractionation of the measured data of total COD and soluble COD into the state variables used by the model. Experiments to characterize wastewater fluxes, such as by respirometry methods, will allow capturing changing conditions and (regular) updating of the fractionation parameters (e.g. Mesta et al., 2023).

Evaluation of the Janus coefficient (see Table 4-13), which serves as an indicator of the performance of the model during validation in comparison to calibration, shows values in the range [1.36: 2.20] with

an average of 1.75 indicating good model performance. The nitrogen components (NO_3-N and NO_2-N) have a coefficient of around 1.5, which in the case of denitrification, is very relevant and an indicator of well-modelled biological behaviour. The Janus coefficient for total COD (2.20) and soluble COD (2.13) indicate a validation error that is however not far from the acceptable coefficient range between 1 and 2. TSS has a Janus coefficient similar to nitrates and nitrites reflecting the good description of the filtration by the model.

4.5.3 Integrated Biofiltration Model

The integrated model of the biofiltration wastewater treatment lane allows not only to simulate preDN effluent quality, but also to simulate the effluent of the nitrification (NIT) and postdenitrification (postDN) stages. Since the results for the preDN stage have already been discussed above, this paragraph continues the comparison of the modelled effluent variables for organic matter and nitrogen for the sub sequential NIT and postDN stages.

Results for NIT stage

The modelled output of the Nitrification stage is shown in Table 4-14 and Figure 4-15 for the nitrates, nitrites, ammonia, TSS, COD total and COD soluble concentrations for the month of December 2019. The simulations of ammonia (Figure 4-15e) fit the observations with a relative mean error ME/$\bar{y}$ of -0.02 and a mean absolute error MAE/$\bar{y}$ of 0.3, which is affected by short-term, but large peaks occurring on December 19, 21 and 26. Also indicated in the figure is the ammonia set point in the effluent, which fluctuates almost daily between 0.5 – 1.5 mgNH_4-N/L due to the operator testing the impact of different NH_4-N set points.

The modelled nitrate concentrations (Figure 4-15a) also correspond well with the observations. The modelled daily values amount to 22.1 mgNO_3-N/L on average, while the average nitrate concentration observed was 19.3 mgNO_3-N/L (laboratory) and 16.3 mgNO_3-N/L (inline analyser). The statistical scores calculated are a ME of -3.4 mgNO_3-N/L, an MAE of 3.4 mgNO_3-N/L and a RMSE of 4.4 mgNO_3-N/L, values that fall between 17 – 23% of the observed mean.

Table 4-14: Overview of statistical scores for different effluent variables during a 30-day period (December 2019) at the NIT stage.

		Calibration							
Variable	Unit	N	Mean obs.	ME	ME/$\bar{y}$	MAE	MAE/$\bar{y}$	RMSE	RMSE/$\bar{y}$
NO_3-N	gN/m^3	30	19.31	-3.37	-0.17	3.45	0.18	4.39	0.23
NO_2-N	gN/m^3	30	0.84	-4.67	-5.54	4.67	5.54	4.89	5.8
totCOD	gO_2/m^3	30	42.59	-10.89	-0.26	11.24	0.26	11.83	0.28
solCOD	gO_2/m^3	30	32.03	-3.2	-0.1	3.2	0.1	3.37	0.11
TSS	g/m^3	30	9.93	-3.73	-0.38	3.82	0.38	4.88	0.49
NH_4-N	gN/m^3	30	2.34	-0.06	-0.02	0.69	0.3	0.79	0.34

There is, however, a complete overestimation of the concentration of nitrites by the model (Figure 4-14b) which results in a large relative mean error ME/$\bar{y}$ (-5.5) and mean absolute error MAE/$\bar{y}$ (5.5) due to the low average observed nitrites concentrations in the effluent (Table 4-13). Part of this

difference could be explained by the instability of NO_2-N after sampling and the difficulty to measure it in the lower range. There might also be an accumulation of nitrites in the modelled bioreactors that gets released into the bulk liquid and results in higher NO_2-N concentrations in the modelled effluent.

Furthermore, Table 4-14 presents an overview of the values of the statistical criteria calculated from simulations of the NIT effluent during the calibration period. The relative ME/$\bar{y}$ values indicate that the bias of the model is relatively low compared to the average values observed, in particular for the variables NH_4-N (ME/$\bar{y}$ = -0.02), soluble COD (ME/$\bar{y}$ = -0.1) and NO_3-N (ME/$\bar{y}$ = -0.17), albeit all negative values indicating an overestimation of the effluent concentrations. In the case of total COD, this bias becomes stronger with a ME of -10.9 mgO_2/L. This is partly explained by the contribution in total COD of particulates organic matter during the same period, which is overestimated in larger proportions as indicated by the TSS ME of -3.73mg/L.

The parity plot (Figure 4-15l) exhibits a high correlation with a R^2 value of 0.93 and indicates the model's overestimation as a systematic error (or a bias). A possible explanation is that the non-biodegradable soluble COD (S_I) that forms part of the soluble COD is actually not correctly fractionated during influent characterization. Indeed, the variable S_I is not directly measured, but is often taken as a portion of the value of soluble COD in the effluent (Henze et al., 2000). However, it is known that S_I is in fact produced by organisms when they metabolize COD (about 5-10% of degraded COD is converted to S_I), and this means that the effluent COD is in fact composed of the S_I present in the effluent of the nitrification stage plus the S_I produced in the postDN stage on the basis of the degraded external COD. Moreover, due to the constant dosing of externa carbon in the postDN stage, the total soluble COD in the effluent is higher than it would be without the dosing. Thus, the soluble COD in the nitrification biofilter effluent could be lower than the effluent COD of the postdenitrification reactor.

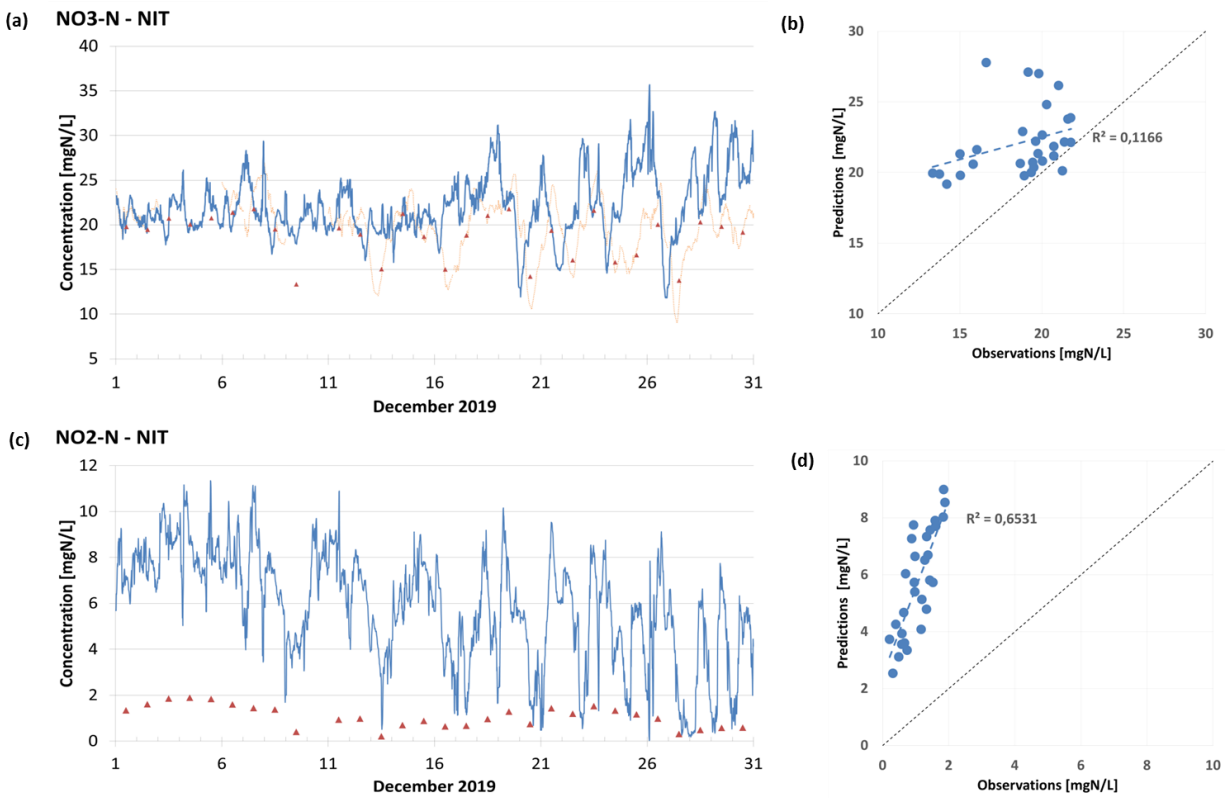
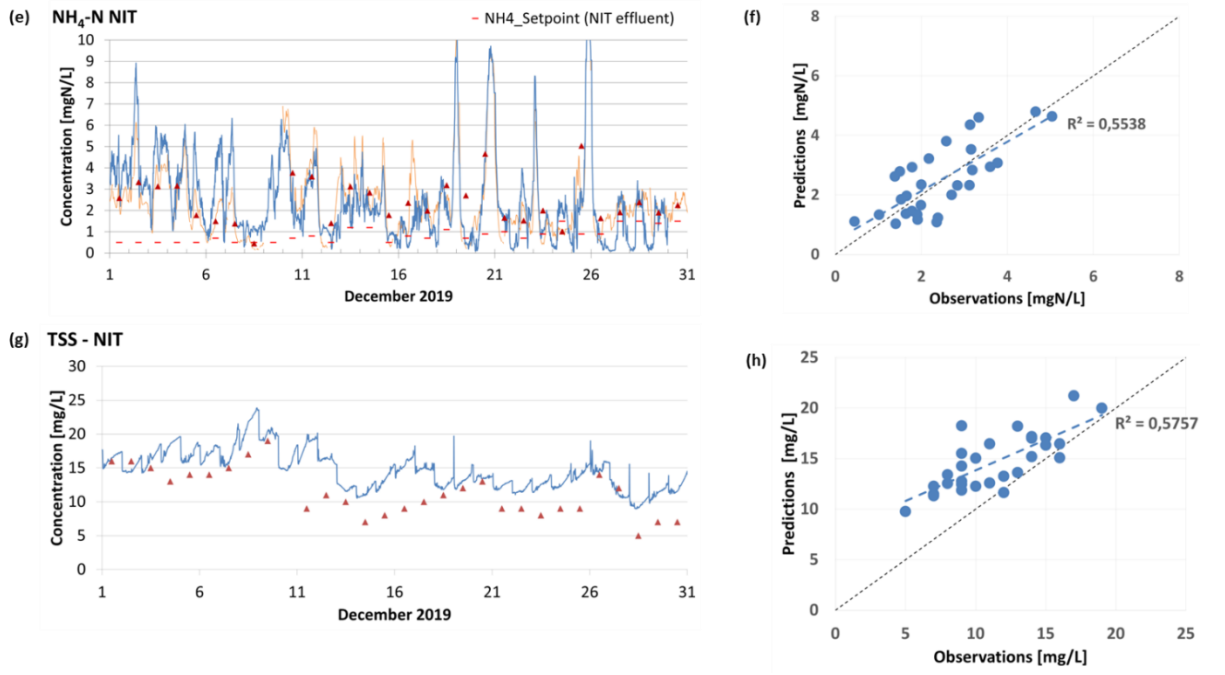


Figure 4-14: Calibration results at the outlet of the NIT stage for a) effluent NO_3-N; b) Parity plot for NO_3-N; c) NO_2-N effluent; d) Parity plot for NO_2-N. Brown points: composite sample laboratory observations, brown line: sensor observations, blue line: simulation results.



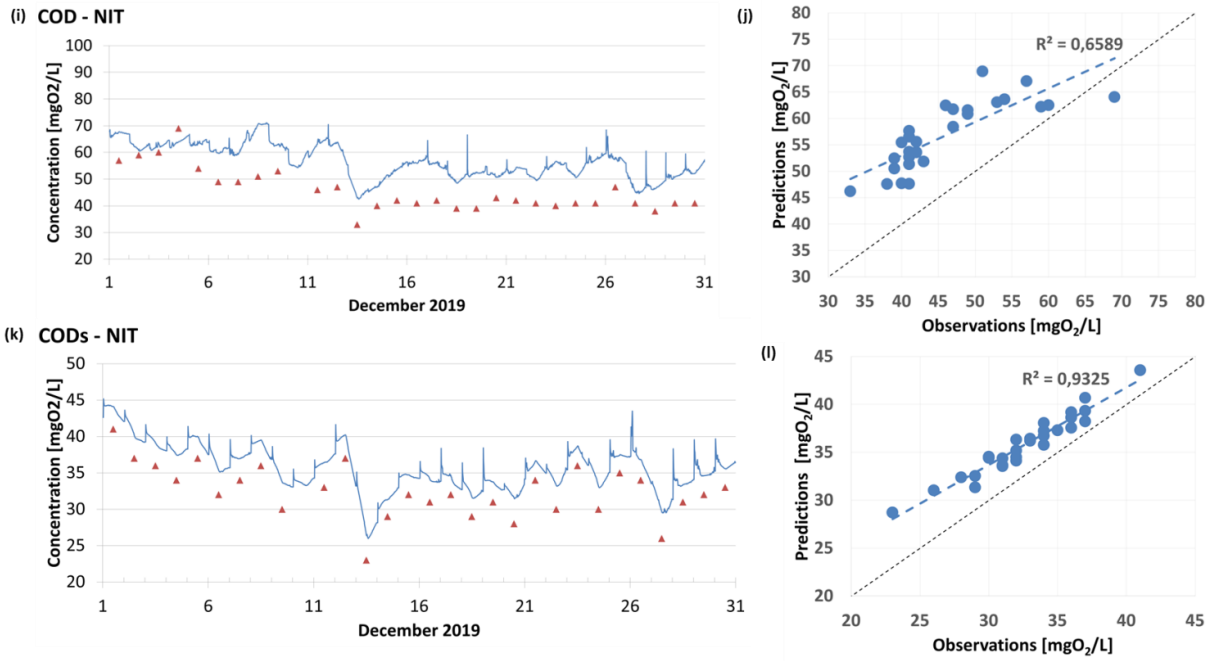


Figure 4-15: Calibration results at the outlet of the NIT for e) NH_4^+-N; g) TSS; i) total COD and k) soluble COD. Parity plots on the right side. Brown points: composite sample laboratory observations, brown line: sensor observations, blue line: simulation results.

Results for postDN stage

The final treatment stage of the biofiltration lane, the downstream denitrification postDN stage, treats the effluent of the nitrification NIT stage by means of a battery consisting of 12 submerged Biofor© filters. An external source of carbon is added to the water in the form of methanol (CH_3OH), which serves as the main source of readily biodegradable substrate for the biomass.

The modelled effluent quality of the postDN is shown for December 2019 in Figure 4-16 for the variables nitrates and nitrites, and in Figure 4-17 for ammonia, TSS, total COD and soluble COD. Concerning nitrate effluent simulations, when compared to the effluent sensor signal, the simulated amplitude is higher than that measured and shows peaks of up to 11 to 12 mgN/L that remain unexplained. A very low R^2 value is witness to the lack of correlation. The simulation of nitrites in the effluent is also overestimated and shows a systematic bias.

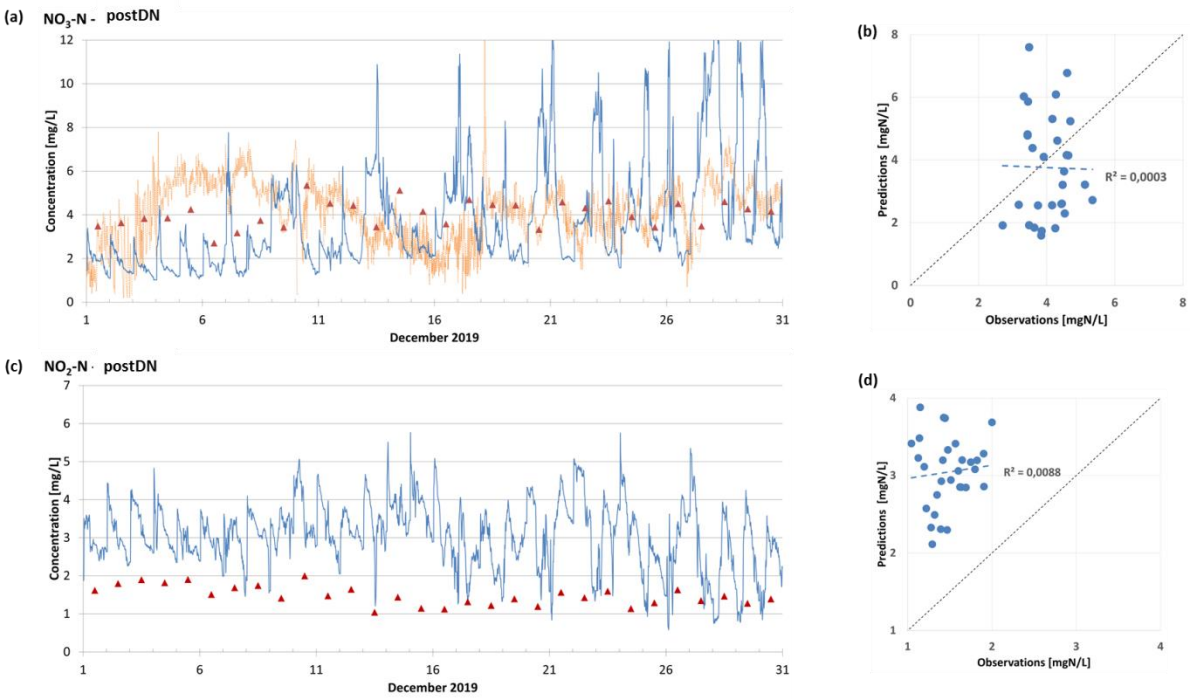
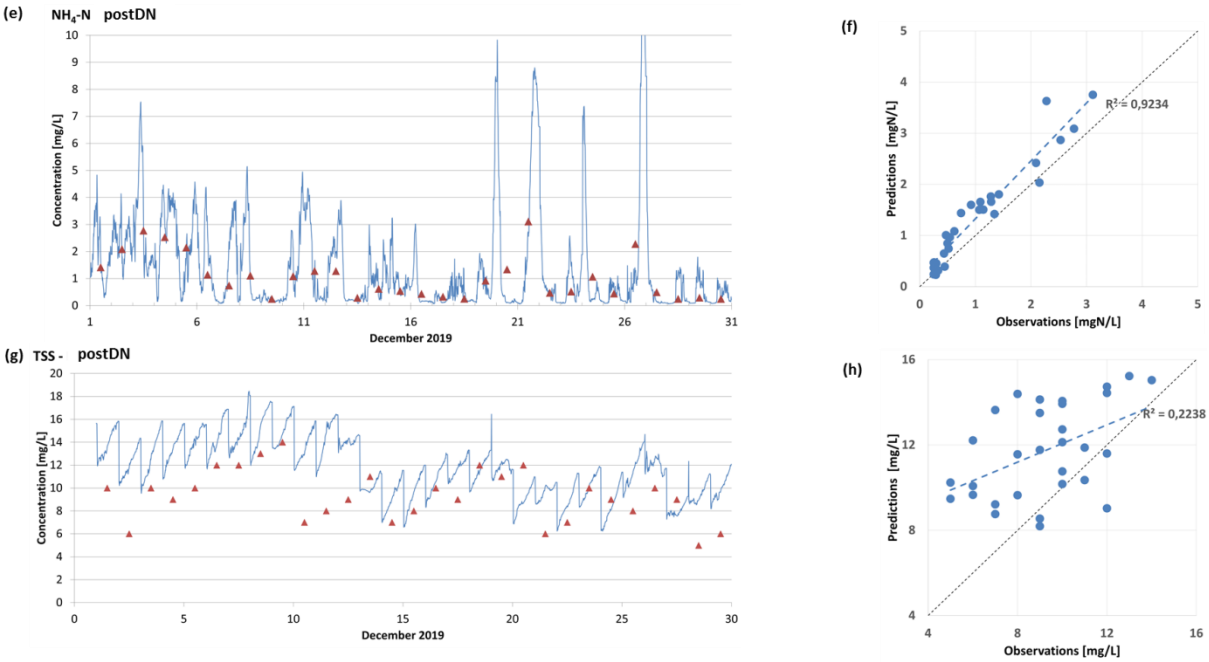


Figure 4-16: Calibration results of the effluent of the postDN for a) NO_3-N; b) Parity plot; c) NO_2-N; d) Parity plot. Brown points: composite sample laboratory observations, brown line: sensor observations, blue line: simulation results.

The model simulations for ammonia shown in Figure 4-17e fit the measurements quite well with a R^2 value of 0.92 for the averaged values. Estimations for COD total (mean modelled: 85.4 mgO_2/L; mean observed 54.5 mgO_2/L) and COD soluble (mean modelled: 65.8 mgO_2/L; mean observed 46.1 mgO_2/L) are about a factor 1.5 too high, in the case of COD total with a scale factor of 1.6 and 1.4 for COD soluble. The TSS concentration modelled follows the trends observed, although it starts too high.



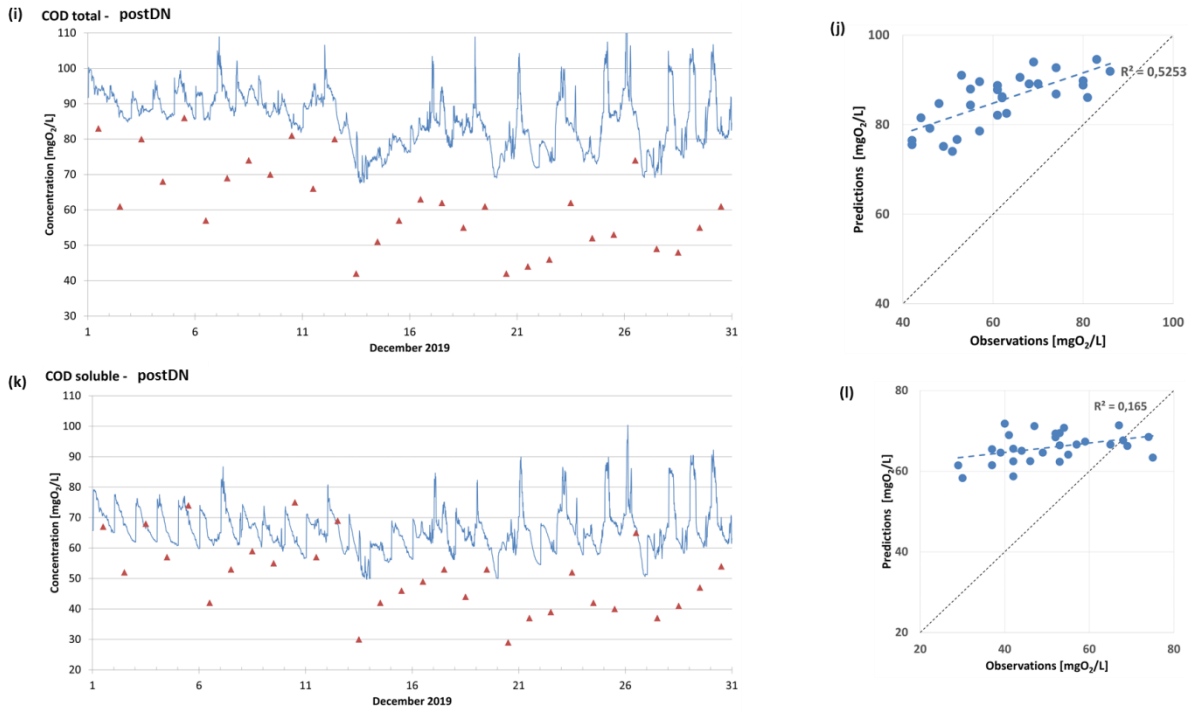


Figure 4-17: Calibration results of the effluent of the postDN for a) NO_3-N; c) NO_2-N; e) NH_4^+-N; g) TSS; i) total COD and k) soluble COD. Parity plots for each of the variables on the right side. Brown points: composite sample laboratory observations, brown line: sensor observations, blue line: simulation results.

Table 4-15 provides an overview of the errors for the postDN effluent variables for December 2019. The negative ME scores for all variables but the NO_3-N confirm that the model in general overestimates the effluent concentrations at the postDN. This indicates a lack of metabolic activity of the biomass in the model. This could be due to the applied procedure of methanol dosing, of which the process control was integrated in the model is in accordance with the constructor's guidelines. It might be that since then, some instructions or procedures have been modified (e.g. different set points or ratio of methanol dosing), causing a bias in the model results.

Table 4-15: Overview of statistical scores for calibration of postDN.

			Calibration						
Variable	Unit	N	Mean observed	ME	ME/ȳ	MAE	MAE/ȳ	RMSE	RMSE/ȳ
NH_4-N	gN/m^3	31	1.0	-0.3	-0.34	0.4	0.36	0.4	0.45
NO_3-N	gN/m^3	30	3.6	0.3	0.08	1.5	0.42	1.8	0.49
NO_2-N	gN/m^3	30	1.7	-1.6	-0.94	1.6	0.94	1.6	0.98
totCOD	gO_2/m^3	30	55	-24	-0.43	24	0.43	25	0.47
solCOD	gO_2/m^3	31	46	-15	-0.32	16	0.35	19	0.40
TSS	g/m^3	30	8	-2.5	-0.32	2.9	0.36	3.4	0.43

Moreover, the dosing of methanol creates conditions in the filters that favour the growth of methanol-oxidizing microorganisms (methylotrophic denitrification), that are characterized by slow kinetics and lower yields compared to that of heterotrophic denitrification with more traditional carbon sources (Lu et al., 2014). This results in less thick and slow growing biomass, thereby moderating the transformation of nitrates into N_2. This slower growth rate also means that at the start of the significant methanol dosing sessions, a longer adaptation time is required before a critical biomass starts to be present. A recalibration of the relevant model parameters might be necessary.

4.6 Conclusions and Recommendations

The treatment of wastewater by activated biological submerged filters (biofiltration) is a complex process with short hydraulic retention times and remains more difficult to model than conventional systems, such as activated sludge. ASM mechanistic models (Henze et al., 2000) are well developed and calibration protocols widely accepted. However, these ASM models cannot be applied directly to biofilm models due to the importance of transport and transfer processes and biological conversions taking place from water to biofilm and within biofilm, as well as the influence of regular backwashing on biomass presence (Rittmann et al., 2018).

Therefore, biofilm models contain a large number of additional parameters that require a more elaborate calibration process and need longer computation times.

In general, models are sensitive to the quality of the data provided. This highlights the importance of the quality of data represented by the "data-pipeline", a process that describes the necessary steps to turn raw data into useful information in order to make intelligent decisions. The data must be of high quality in order to derive useful information from them (Therrien et al., 2020). High frequency data (e.g. 15-minute intervals) from in-line sensors are exposed to large variations and amplitudes, in particular in harsh wastewater treatment conditions, and may show signs of drift, bias and other measurement errors.

A data preprocessing procedure based on the detection of outliers, smoothing and fault detection (Alferes & Vanrolleghem, 2016) was used to reduce the impact of measurement errors. The analysis led to an average of 4% of all measured variables detected as outliers in the calibration data, whereas 16% of all data points had to be removed due to data faults and gaps created by outliers. The highest percentage of outliers was detected in the TSS sensor measurements (16.1%), while the highest percentage of removed data was identified in the PO_4-P sensor (48%). The data cleaning method for the validation data identified on average 1.4% data as outliers, which is 2 to 3 times lower than during calibration. The percentage of removed data points for the validation data was about 12%, 25% less than during calibration. This could be explained by calmer conditions during the validation period resulting in less fouling and other sensor errors, as well as increased sensor maintenance by the operators.

In this study, a previously developed mechanistic model (Zhu, 2020) was adapted using high –frequent (15 minutes intervals) measurement data (2019 – 2020) from the secondary treatment by biofiltration on a full-size municipal wastewater treatment plant (Seine-aval, SIAAP). The calibration method involved a sensitivity analysis in order to identify the most influential parameters on the response of the model describing the denitrification procedure in successive stages. Parameters characterizing filtration and washing, material transfer to biofilm by diffusion, the nature of the biofilm and biological reactions were identified as having the most influence.

The simulation of the calibrated model for the calibration period showed that the model managed to simulate the effluent NO_3-N, NO_2-N, NH_4-N, COD, COD and TSS with reasonable performance for the upstream denitrification preDN stage, having average errors between 10 – 30% of the observed mean. A similar performance is modelled for the effluent of the nitrification NIT stage, with a good estimation of effluent soluble COD, ammonia and nitrates (ME/$\bar{y}$ < 17%). The model predictions for the effluent

of the downstream denitrification postDN stage however are less satisfactory. This could be due to the presence of methylotrophic bacteria that require an adaptation of kinetic model parameters, which fell outside the scope of the study. The model's performance could also benefit from an additional calibration period that includes the summer months to cover different operating conditions.

Concerning the model performance for the preDN stage with validation data unknown during calibration, the Janus coefficient for nitrate, nitrite and TSS effluent was found to be between 1 and 2, indicating a valid model. The Janus coefficients for total COD and soluble COD effluent in the preDN stage are slightly higher than 2, but are still very acceptable.

Overall, this work demonstrates the potential of mechanistic biofilm models to describe the behaviour of denitrifying and nitrifying biofilters. However, the accuracy of the simulations of a mechanistic biofilm model could be further improved by adding an intelligent data-based layer to it in a hybrid concept, which will be the topic of the next chapter.

5 Hybrid Modelling of Nitrogen Removal by Biofiltration: A Full-scale Study

Keywords: *Hybrid models, data-driven models, mechanistic models, biofiltration, wastewater treatment.*

Abstract

The digitalization of the wastewater sector is proceeding fast with the application of data-based monitoring, model-based predictive process control and other tools for data-driven decision-making.

In this chapter, a parallel hybrid model is presented for the simulation of nitrogen removal by submerged biofiltration of a very large-size WWTP. This digital twin consists of a Machine Learning model and a mechanistic model that are combined to produce accurate predictions of water quality variables as the processes' output. The models are calibrated/validated using detailed and quality-controlled operational data collected over a period of 3.5 months in 2020. The mechanistic model applied is a modified Activated Sludge Model 1 that describes the biological, physical and chemical processes taking place in a biofilm reactor. The mass balance equations are based on the domain knowledge of these processes. A three–layer feedforward Artificial Neural Network with a ReLu activation function that aims to reduce the mechanistic model's residual error then corrects its output.

The results show how the hybrid model outperforms and significantly reduces the size of the mechanistic model's prediction errors of the effluent nitrate concentration from a relative mean error of 12% for the mechanistic model to 2% for the hybrid model during training. While the error on nitrate simulations increases to 8% during hybrid model testing, this is significantly lower than the error of 110% for the mechanistic model. The overall good performance of the hybrid model allows for future applications of hybrid biofilm models in an operational context, such as for process control with a Hybrid Model Predictive Controller (the subject of chapter 6).

5.1 Introduction

Improving the performance and efficiency of wastewater treatment plants (WWTPs) involves strengthening the control of the processes that take place there. Optimization of energy performance, reduction of the use of reagents, of greenhouse gas emissions and of overall environmental impact of operations, the improvement of the recovery of resources and a stable effluent quality are nowadays the objectives for the wastewater treatment sector. An important field of research aims to study the potential of the application of artificial intelligence (AI) techniques to achieve this aspired improvement in the performance of WWTPs (Schneider et al., 2022).

Wastewater treatment includes several successive physical, biological and chemical processes, that are non-linear and non-stationary, which complicates the optimization of their management (Henze et al., 2008). To facilitate and improve treatment and ensure the achievement of performance in accordance with regulatory obligations, in recent decades, the instrumentation, control and automation (ICA) of WWTPs has developed significantly, such as elsewhere in many other industries (Ingildsen and Olsson, 2016; Corominas et al., 2018). This trend can be seen particularly in the areas of flow monitoring and water quality indicators, and the development of mathematical models for simulation and scenario analysis (Garrido-Baserba et al., 2020).

Since the 1990s, mechanistic dynamic models describing the suspended growth process of biomass in activated sludge systems, such as the ASM1 and ASM2 models (Activated Sludge Models, Henze et al., 2000) have acquired a solid and valid basis within the scientific community. These models contain a large number of parameters and require a significant calibration phase (Rieger et al., 2012). Yet, in the case of treatment by biofiltration, which is an active and compact 'fixed-culture' biomass process that is widely applied in urbanized regions of, for instance, Paris, model adaptation is necessary in order to describe the processes that take place within a biofilm (e.g. mass transport by diffusion, kinetics of several types of biomass and the effect of backwashing). Mechanistic dynamic biofilter models are therefore considered more complex and are still under full development. Additionally, there are not yet widely accepted protocols or guidelines available to the modelling community to guide the development or use of biofilm reactor models (Rittmann et al., 2018).

Moreover, the greater complexity of the mechanistic models used makes their set-up, calibration and validation much more laborious, time consuming and data dependent. This can become a constraint limiting the application of the model (Vanrolleghem et al., 2005). These models also require relatively high computing power, which makes them less suitable for high frequency applications where model updates are needed in real time with a short time horizon of action (Schneider et al., 2022).

Quite otherwise, data-driven models that use AI techniques to find patterns in data are computationally very fast, making them very interesting for real-time applications. These models describe the system only based on information extracted from the provided data and have strong interpolation features, but they are less reliable for making predictions outside of training conditions (Newhart et al., 2019). Data-driven models typically require larger data-sets than mechanistic models. Related to the inherent characteristics of wastewater treatment processes, which are non-linear and show a large time dependency, frequent retraining of a data-driven model remains necessary (Torfs et al., 2022).

Hybrid modelling combines the use of knowledge-based models with data-based models and is able to cumulate the advantages of each approach (Von Stosch et al., 2014). The hybrid model concept combines the strengths of these two modelling techniques: a mechanistic model that incorporates relevant knowledge about the processes, as well as a data-driven model that increases the predictive power of the model by including information on lesser-known sub processes at reduced computing power. The training and testing of machine learning models is much faster than the calibration and validation of mechanistic models. In recent literature, hybrid models describing activated sludge WWTPs are reported to improve model performance and support process monitoring and control (Torfs et al., 2022). The main advantage of a hybrid model is a higher benefit/cost ratio for solving complex problems, which is a key factor for process systems engineering (Schneider et al., 2021).

The hybrid concept has been the subject of research for a decade in other industrial sectors, where AI techniques have already been introduced and adopted. The wastewater sector is experiencing a lag firstly due to the complexity of the processes to be modelled, the high level of investments and the long life expectancy of the infrastructures, and additionally requires a specific adaptation of the methods to take into account the uniqueness of the wastewater treatment process, such as the interdependent, nonlinear, and nonstationary nature of WWTP data (Corominas et al., 2018). On the other hand, there lies great potential in applying AI and its machine learning (ML) methodologies associated with the analysis of "big data" (Garrido-Baserba et al., 2020).

A critical element in their successful use remains the quality of the data and the way it is processed and applied ("the data pipeline"); in order to be able to make intelligent decisions, raw data must be collected, cleaned, structured and made accessible to mathematical models in order to derive knowledge and understanding, so that it leads to actionable insights (Therrien et al., 2020). The application of AI models to clean and generate long, high-frequency input datasets (i.e. influent generator for the WWTP models) is a promising complementary research (Li & Vanrolleghem, 2022).

In this chapter, a hybrid 'digital twin' is developed to improve wastewater treatment modelling by combining a mechanistic dynamic simulation model with an AI data-based model using high-frequent operational data. The hybrid model reduces the size of prediction errors. This increases the confidence in the model's performance and allows the application of hybrid biofiltration models as a decision-making tool for managers and operators in the wastewater industry.

5.2 Model Development

A hybrid model (HM) in the context of wastewater processing involves an interaction between a mechanistic model (MM), which serves as the fundamental basis for the prediction of effluent water quality variables, and a data-driven model (DDM) that supports the former. The mechanistic model specifies the basic dynamics of the relevant process variables and preserves process knowledge ('the First-Principles'). It has good extrapolation capabilities. The data-driven model learns unknown relationships from the data and helps improving simulation quality, especially for interpolation purposes. The DDM accounts for lesser known and non-linear characteristics of wastewater treatment that are not explained by the mechanistic model. These capacities can compensate for dynamics missed by the mechanistic model (Schneider et al., 2022).

The hybrid model developed here has a parallel configuration, where the DDM and the MM model both produce an output that is subsequently combined, by addition of the two outputs, into a modified, more accurate result that improves the fit of data. The objective of the data-driven component is to calculate the residual error of the knowledge-based model. It extracts relationships from the data inputs with which it can estimate the error of the MM-based simulation of the system. However, it does not learn explicitly the underlying process dynamics, which could form a problem when trying to make simulations in situations that fall outside the conditions with which the model was trained (Anderson et al., 2000). To overcome this limit in extrapolation power, the data-driven component needs to have sufficiently large and representative data that capture the diversity of operational conditions (Quaghebeur et al., 2022). The parallel structure provides a significant advantage for conditions where the MM contains a large number of parameters necessary to describe complex processes, such as in biological wastewater treatment applications (Lee et al., 2002).

5.2.1 Hybrid Model Architecture

The hybrid model architecture adopted here is identifiable as a cooperative parallel hybrid structure (Schneider et al., 2022) where the data-driven model has been trained to calculate the error in the simulations of the mechanistic model (DDM-estimated error). The output from the mechanistic model is then corrected by adding or subtracting the DDM-predicted error. During the DDM training phase, which corresponds to a supervised learning, the residual error, which is defined as the difference between the values of the mechanistic model simulations and historical data or observations (residual error), is known to the DDM.

The key element for a hybrid model lies in the information content of these residuals: a well-trained machine learning model can pick-up 'hidden' dynamic information and use this to correct the simulation of the mechanistic model (Lee et al., 2002).

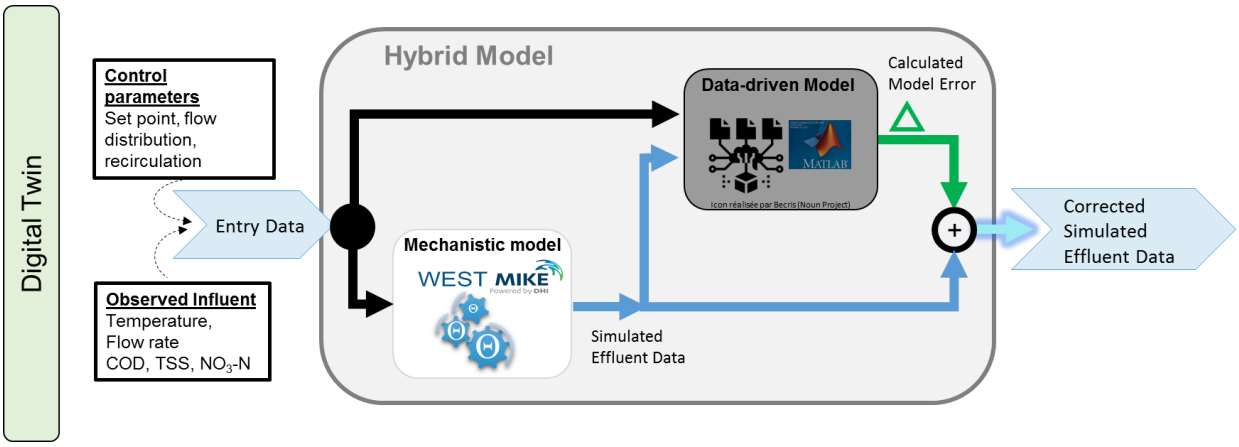


Figure 5-1: Schematic representation of the developed hybrid model (grey box) where the simulations (blue arrow) of the knowledge-based mechanistic model (white box) and the estimated residual error (green arrow) by the Data-Driven Model (black box) are combined to produce corrected values of effluent quality variables (for more details, see text).

The architecture of the developed hybrid model applied in this study (Figure 5-1) is a parallel structure where the data-based model is trained to predict the difference between the mechanistic model simulations and the measurements (the residual error Δ). In this configuration, the mechanistic model first simulates the effluent water quality variables, sludge concentrations, greenhouse gas emission and flow rates, as well as the intermediate state variables of the model. Selection of features is based on data availability followed by an analysis of importance using an F-test method that allows reducing the dimensionality of the data. The selected simulated variables are then passed to the data-based model (blue coloured arrow in the figure). In addition, the data-driven component receives measurement data of the influent quality and flow rates, as well as information on the principal treatment parameters, such as the effluent set points for nitrates, as specified in the MM (black arrow). The DDM uses this information to estimate the residual errors of the mechanistic model simulations Δ (green arrow), which are defined as the most probable expected model errors based on the training differences between previously observed (measured data) and simulated data. Finally, the two results are added together to obtain corrected values of the effluent variable (light blue highlighted arrow). In this approach, a machine learning model is trained for each of the output variables, where for each variable, a corresponding subset of process input variables or features is used.

During the training phase of the data-driven model, observations of the water quality variables to be modelled are used to calculate the difference between observed and simulated variables. This residual error of the variables as simulated by the mechanistic model is then used by a Bayesian optimization algorithm to train the machine learning model so that it learns to generate the proper errors. It serves as labelled responses in a supervised machine learning training environment.

5.2.2 Mechanistic Model Structure

The mechanistic model applied here is designed as a ('pseudo') 1-dimensional dynamic biofilm reactor model describing a 'fixed-culture' submerged upward flow biological filter. The model was previously developed (Zhu et al., 2019; Zhu, 2020) based on an earlier model (Bernier et al., 2014) that uses the

ASM1 biokinetics (Henze et al., 2000). It consists of the combination of sub-models for the hydraulic representation of the filter media, the transfer and transport of soluble and particulate substrates to/from and within the biofilm, the kinetics of bacterial growth and decay, and backwashing to avoid clogging of the filter. The model simulates the biological and physicochemical conversion of carbon, nitrogen and phosphorus. In total, 22 biological conversions are accounted for by describing the evolution of substrates by a mass balance.

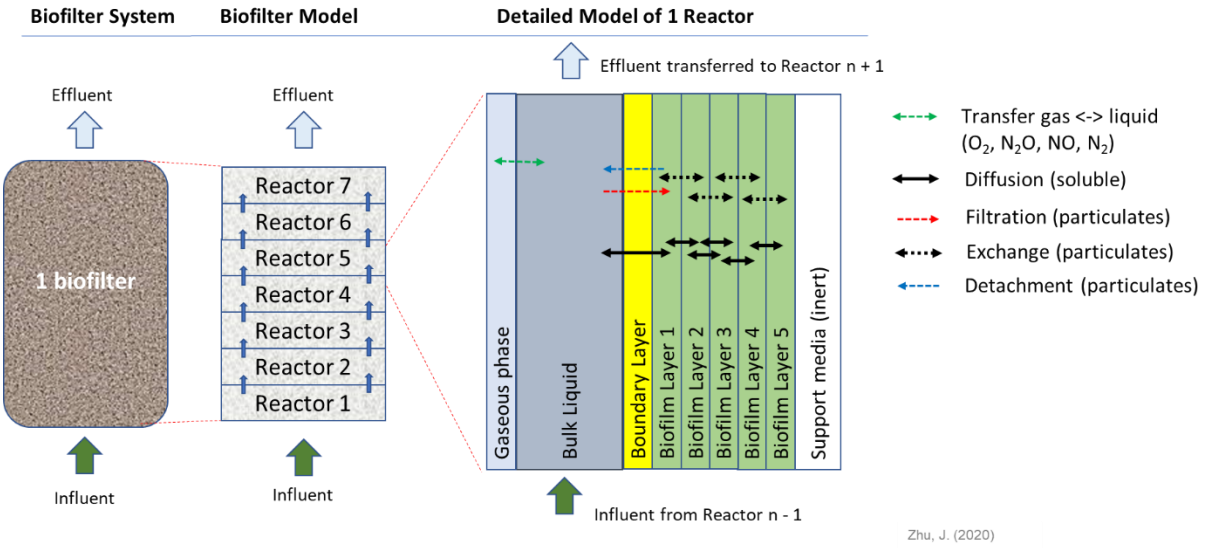


Figure 5-2: General concept of the biofiltration model developed by Zhu (2020) that simulates the behaviour of 1 biofilter, divided over seven reactors. For each reactor there are compartments for gas, bulk liquid, the Mass Transfer Boundary Layer and the biofilm, divided into 5 layers. The horizontal arrows indicate the main physical phenomena modelled.

The performance of biofilters is related to the filtration of suspended particles and the biological activity of purifying biomasses (Henze et al., 2008). Figure 5-2 shows the approximation of the vertical hydraulics of a filter into seven successive completely mixed reactors. A number that was selected to provide enough level of modelling detail that does not incur a too high computational cost (Zhu, 2020). The volume of each mixed reactor is divided between the biofilm compartment (in orange), a bulk liquid phase (blue), a gaseous phase (grey) and the inert media (white). The Mass Transfer Boundary Layer (yellow) simulates the transfer resistance of soluble components from the liquid phase to the biofilm by diffusion, which is a representation of the stagnant liquid film between the biofilm and the liquid phase (Boltz et al., 2011). To better simulate the impact of this boundary layer, the model considers a variable thickness of the boundary layer for each soluble component as a function of the media diameter and the soluble component's Sherwood number based on the equation of Ohasi et al. (1981). This allows simulating with more precision the impact of different hydraulic conditions near the surface of the biofilm that directly affect the diffusion resistance (Zhu, 2020). The biofilm compartment is split into five homogeneous layers, a number applied previously by Vigne et al. (2010) that was found to not take too much computational time and still accurately simulate the concentration gradients of soluble substrate (Zhu, 2020). The model assumes a dynamic biofilm volume with a variable thickness of each biofilm layer up to a predefined maximum allowable thickness, based on the mass of particulates present in each layer, the surface area of liquid to media and a constant biofilm density.

The physical processes considered in the model (the arrows in Figure 5-2) in each reactor are the transfer to and transport within the biofilm of soluble components by diffusion, the transfer and

transport of particulate components by filtration, attachment and detachment and exchange between the layers of the biofilm; as well as a flux of four gaseous components (O_2, N_2, NO, N_2O). These gazes are considered as soluble components in the biofilm layers and transported by diffusion (Zhu, 2020). In addition, the model redistributes the particulate components between the successive mixed reactors during treatment and especially during backwashing by representing continuous media mixing at different intensities for normal operation and backwashing. Mixing occurs between the same layers of the bioreactor directly above and below it. The model allows estimating energy consumption related to aeration and pumping, as well as greenhouse gas emissions (Zhu et al., 2018).

5.2.3 Data-Driven Model Structure

In a hybrid model structure, different data-driven models selected among the machine learning (ML) methods can be applied, such as artificial neural networks (ANN), support vector machines (SVM), ANFIS, non-linear regression and decision tree-based methods. These models can be categorized based on the learning processes applied, i.e. supervised, unsupervised and reinforcement learning (Schneider et al., 2022). In this study, supervised learning is applied.

A lot of attention is given in recent research to the application of a variety of neural network structures and training algorithms to wastewater systems (i.e. Lee et al., 2002; Shirkoohi, 2022; Li & Vanrolleghem, 2022) since neural networks typically have good predictive accuracy. In this study, the selection of the optimal regression model includes the exploration of data features, validation schemes and model performance. The performance of ML models depends heavily on the data quality and quantity. Since no domain knowledge is included (the ML-model parameters are not related to knowledge of physical processes), these models have been criticized for a lack of interpretability (Quaghebeur et al., 2022). In a hybrid concept nevertheless, where the mechanistic model is based on domain knowledge, the data-driven component is only allowed to correct the former within certain boundaries, such that the impact of interpretability remains controlled. The data-driven part is here inherently coupled to a model structure with physical meaning, which allows evaluating the specific dynamics of the data-driven part as a measure of model structure error of the mechanistic model (Schneider et al., 2022).

ML models are sensitive to overfitting when the model learns too well and 'memorizes' the associated data patterns to the output in such a way that the error on the training set becomes very low, but the error on a testing set with unseen data becomes very high. This can be prevented by applying generalization learning methods, such as cross-validation, regularization and pruning during training (Shirkoohi, 2022).

In this study, a cross-validation method was selected before the start of model training so that the models can be compared with the same validation scheme. Cross-validation is a resampling procedure used to evaluate machine learning models that uses smaller subsets of the original dataset in order to avoid an overfit or underfit of the model during training. The k-fold cross-validation technique partitions the data into k-subsets ('folds') and trains a model for each subset using the corresponding data. Finally, it calculates the average validation error over all subsets. This method makes efficient use of all the data and still gives a good estimate of the predictive accuracy of the trained model, although it can take a longer time to execute because the model is trained repeatedly on the subsets.

5.3 Materials and Methods

5.3.1 Case Study Site

The research project was carried out within the Seine-aval (SAV) WWTP located in Achères (Yvelines), France, managed by SIAAP. This large size plant receives wastewater generated by nearly 6 000 000 population equivalents every day, i.e. a volume of approximately 1 250 000 m^3/d in dry weather, which corresponds to an average daily flow of 14.5 m^3/s. The reference flow (maximum flow considered) is 2 300 000 m^3/d (26.6 m^3/s) (Zhu et al., 2018).

The wastewater treatment chain consists of a physico-chemical treatment step (screening, grit-oil removal and primary settling enhanced by the addition of coagulant-flocculant reagents) followed by biological filtration carried out over three stages for the removal of organic matter and nitrogen in a configuration of predenitrification (preDN), nitrification (NIT) and postdenitrification (postDN), as shown in Figure 5-3. The typical flow rate to the biofiltration train during the period studied (December 2019 – October 2020) is 1 100 000 m^3/d.

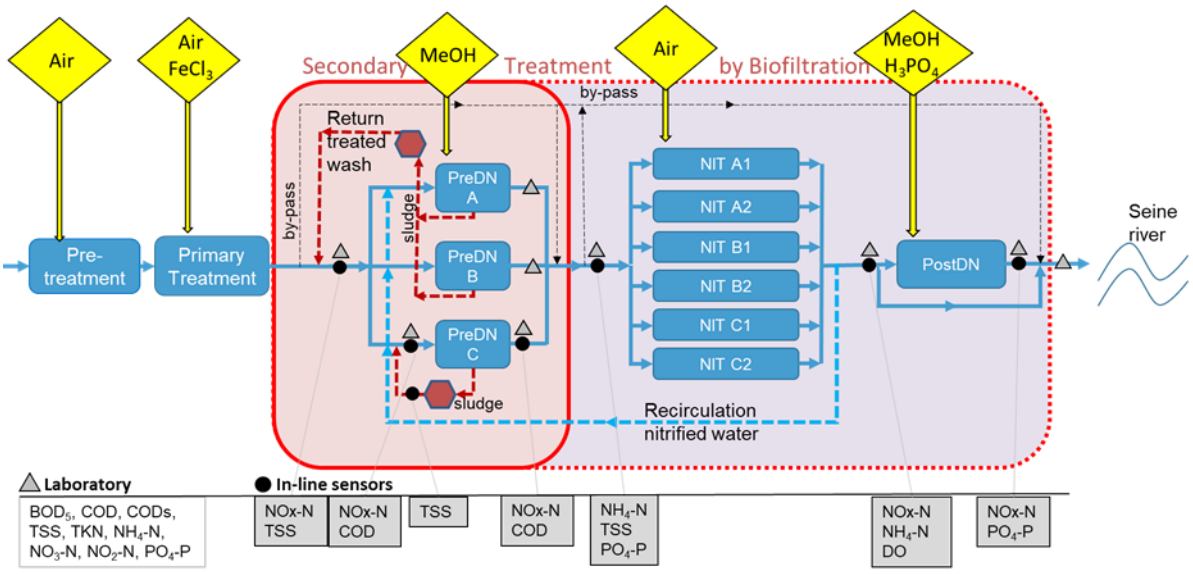


Figure 5-3: Schematic diagram of the principal wastewater treatment lane at the Seine-aval WWTP (SIAAP) with indication of processes (blue boxes) and main dosing actions (yellow diamond shapes). Outlined in red frame are the secondary treatment predenitrification biofilter units (preDN), the nitrification battery units (NIT) and postdenitrification unit (postDN). The focus of this study is on the preDN stage. Also indicated are the sludge treatment units for preDN (brown hexagons) and location of measurements by in-line sensors (black dots) and laboratory samples (grey triangles).

The first stage of biofiltration (preDN) consists of 58 Biostyr® type biological filters (173 m^2 * 3.5 m height/biofilter; granular material: polystyrene 4.5 mm in diameter; three batteries: A, B, C) for the denitrification of nitrates (NO_3-N) and nitrites (NO_2-N) by heterotrophic bacteria using the organic matter (OM) present in the settled water. Nitrified water from the NIT outlet is recirculated to the preDN inlet. The second stage includes 84 Biostyr® type biological filters (173 m^2 * 3.5 m height/biofilter; granular material: polystyrene 4.0 mm in diameter; six batteries: NIT A1, A2, B1, B2, C1, C2). The final section (postDN) consists of 12 Biofor®-type biological filters (147 m^2 * 3m height/biofilter; granular material: clay 4.5 mm in diameter; 1 battery) that denitrify nitrified effluent by the addition of an exogenous carbon source (here in the form of methanol CH_3OH).

The average upflow velocity through the Biostyr filters is 10 m/h for an average NO_3-N removal of 70% (preDN) and 7.2 m/h to remove 90% of the NH_4-N (NIT). The average upflow velocity of the Biofor filters (postDN) is 19 m/h with an NO_3-N reduction of 85% over the period studied for dry and rain weather events combined.

The biological filters are backwashed daily on a fixed schedule by reversing the flow using treated water stored on site and accompanied by the injection of large air bubbles. During washing, which lasts 30 minutes per day per filter, the trapped particles and part of the biomass are removed in order to restore the filtration capacity. Additional short backwashing sessions of 10 to 15 minutes can be applied to a filter when the pressure loss reaches a determined value.

5.3.2 Data Input

The datasets used in this study cover the period from December 2019 to October 2020 and consist of measured data from sensors and online analysers of flow rates (Krohne, MIK electro-magnetic flow rate meter and Vegabar 80 for electronic differential pressure measurements) and water quality variables for organic matter (Hach Lange, UVAS Sc UV probe), TSS (Hach Lange, Solitax Sc, infrared) and nitrogen (NH_4-N and NO_x-N) (Hach Lange, Nitratax Sc 1mm, UV probe) with a 15 minutes time itervals, supported by daily composite laboratory samples collected at key points in the system (see Figure 5-3).

The calibration period covers December 2019 until March 15, 2020, which coincides with the start of the first lockdown (Covid-19). The validation period runs from June to September 2020, a summer period characterized by low flows during the holiday month of August, conditions very different from the calibration period, making for an ambitious validation challenge.

Operational data collected in raw format were first cleaned and smoothed to remove outliers and erroneous data using the method described by Alferes and Vanrolleghem (2016). Datasets for model input require gap filling of missing data points, for which the moving median method was applied. Furthermore, data input for the ML models is normalized using the z-score, with a mean $\hat{y}$ and a standard deviation σ, and scaled in the range of 0 to 1 to provide uniform data across all features that still have the original shape properties.

5.3.3 Model Performance Criteria

The evaluation criteria used to evaluate the model performance are the mean error (ME) between observed and simulated data, the mean absolute error (MAE) and the root mean square error (RMSE), as well as their relative values. In an ideal case, the errors should be close to 0, but values of up to 15% - 20% are considered acceptable. The Janus coefficient, which has an optimal value of 1 and is statistically considered acceptable up to a value of 2, makes it possible to compare the validation results with those of the calibration (Hauduc et al., 2015). Janus coefficients above 2 indicate a weak generalization of the model and an issue of over-fitting.

Model calibration for the mechanistic and hybrid models took specifically into account the RMSE, which provides a quantitative measure of how well the model fits into the data average, although hit is more sensitive to outliers or large errors. The ME and MAE were supplementary indicators used for indication of a model's bias and of the error's size on average.

5.3.4 Mechanistic Model Calibration

The mechanistic biofilter model was calibrated with high-frequent (15 minutes intervals) operational data from December 2019 to March 2020 in accordance with the method described by Mannina et al. (2011), which extends and follows the modelling guidelines for IWA Activated Sludge Models (Rieger et al., 2012). The applied method consisted of the following: First, the selection of the output variables of interest: TSS; Soluble COD and NO_3-N; NH_4-N; NO_2-N. Second, the identification of the parameters that have a significant effect on each of the selected variables by performing a global sensitivity analysis using the well-established method of standardized regression coefficients (Cosenza et al., 2013). Overall, 89 parameters were retained. Third, the evaluation of statistical scores that indicate model performance, based on the RMSE value.

The model calibration, which was performed in the WEST (DHI) environment, is described in more detail in chapter 4.

5.3.5 Data-Driven Model Training

Under a supervised learning approach, the response variable is provided to the model as a label during the training phase so that the settings of the hyper-parameters can be adjusted in order to have the model output match the desired outcome. For a neural network for example, these steps include setting the number of layers and neurons, setting the weights and defining the activation function. Training and testing of a ML model follows a general pattern of importing the data, organizing and pre-processing it, exploring data characteristics, building the model, and evaluating performance.

The selection of the optimal DDM type included training with linear regression models, regression decision tree models with a fine, medium and coarse structure; support vector machines and neural networks with one, two and three hidden layers. A five-fold cross validation was applied that splits the data into five subsets (folds). One randomly chosen subset was first used for model validation, while the other four were used for training. This process was repeated five times, so that each subset was used once for validation. The average error across all five partitions is then reported as the average validation error. The standardized variables (features) included the type of day (weekday or weekend), type of weather (dry or rainy), temperature, flow rates for settled water and nitrified water, observed concentrations of TSS in the settled water, nitrates and total COD in the influent of preDN, as well as operational parameters such as the NO_3-N set point in the effluent and the methanol dosing rate. In addition, the mechanistically modelled concentrations of NO_3-N, total COD or TSS at the preDN outlet were included, depending on the desired prediction error. During the training phase of the ML models, the residual error of the mechanistic model with respect to the measured data was transmitted to allow supervised training. This was done for each univariate output to cover the residuals of the NOx concentrations (NO_3-N + NO_2-N), total COD and total suspended solids TSS.

The optimal structure of the DDM was selected by comparing the performance of each model expressed as the RMSE, averaged among the five validation folds, looking for the smallest (least number of connections or weights) and most interpretable model that still provided sufficient accuracy. Due to the time-dependent character of the data, the data are not randomly divided into a training and testing set, but split chronologically where roughly the first 80% of the records in the time series are marked for training and the remainder 20% are set aside for testing.

Optimization of the selected model was done by changing the hyper-parameters, such as the activation function, the number of neurons, the weights and biases. Parameter optimization is an iterative

process, although the interpretability in human terms of a certain change of a hyper-parameter might not be very high. Additional techniques, such as feature importance analysis, are then helpful to better understand the model. In addition, due to the random initialization of weights in a neural network, gradient-descent based methods might be unable to move away from local minima and the process remains difficult to reproduce identically each time with the exact same model performance.

The F-test was applied for feature selection in order to identify the variables that allow to best fit the data. The F-test uses a filter type feature selection algorithm that measures feature importance based on the characteristics of the features, such as feature variance and feature relevance to the response. It examines the importance of each feature individually and allows defining the likelihood that an observed improvement of a fit to data is worth the use of the extra feature. It tests the hypothesis that the estimated response values are determined from datasets with the same mean against the alternative hypothesis that the dataset means are not identical. Larger scores (including infinites) indicate greater importance. Effectively, it measures to which degree features add any new explanatory power. The 'cost' is indicated by the loss in the degrees of freedom due to the added number of features. In other words, the optimum model has less features and a higher degree of freedom and still maintains a high performance (Helsel & Hirsch, 1992).

The ML models were trained and tested on the operational datasets available covering the period 01.12.2019 to 15.03.2020, corresponding with the calibration/validation set of the mechanistic model. The data were split 80% – 20% for training and testing purposes in chronological order to respect the time-dependent character of the observations. A 5-fold cross-validation protocol was used for validation as explained above. The training dataset contained 8 000 data points for each of the response variables NO_3-N, TSS, COD total, with a 15-minute interval covering the period 1.12.2019 to 28.02.2020. The testing of the ML model was performed with the remainder of the data (1 300 points) covering the period 1.3.2020 to 15.3.2020, which are the last two weeks of the dataset used for calibration of the mechanistic model. In this way, a fair comparison can be made about the impact of the ML model on the predictions of the mechanistic model.

The raw data processing, the development and the training of the DDM were all performed in the MATLAB R2022a (MathWorks) environment using the *Machine Learning* and *Deep Network Designer* toolboxes.

5.4 Results & Discussion

The configuration of the hybrid model implies an interaction between a mechanistic model (MM) and a Data-driven model (DDM), as the mechanistic simulated output is used by the DDM to calculate the residual error during training. It is to his end that the simulation output of the mechanistic model creates the reference situation for the DDM and directly affects the final outcome of the hybrid model (HM). For that reason, this paragraph commences with a quick overview of the mechanistic model's performance.

5.4.1 Mechanistic Model Performance

The mechanistic biofilter model (Zhu, 2020) was first calibrated with recent operational measurement data of a predenitrification (preDN) stage measured in a large-size municipal wastewater treatment plant (Seine-aval, SIAAP) for nitrate, nitrite, ammonia, total COD, soluble COD and TSS concentrations with a high–frequent interval of 15 minutes, supported by daily composite laboratory results.

Simulation results from the calibrated MM for effluent nitrate concentrations are presented in Figure 5-4a, which show that the observed trend is well simulated, in particular for the period starting from the end of January until mid-March. However, the model predictions are overestimated by 2 to 3 mgN/L during the period end of December until end of January.

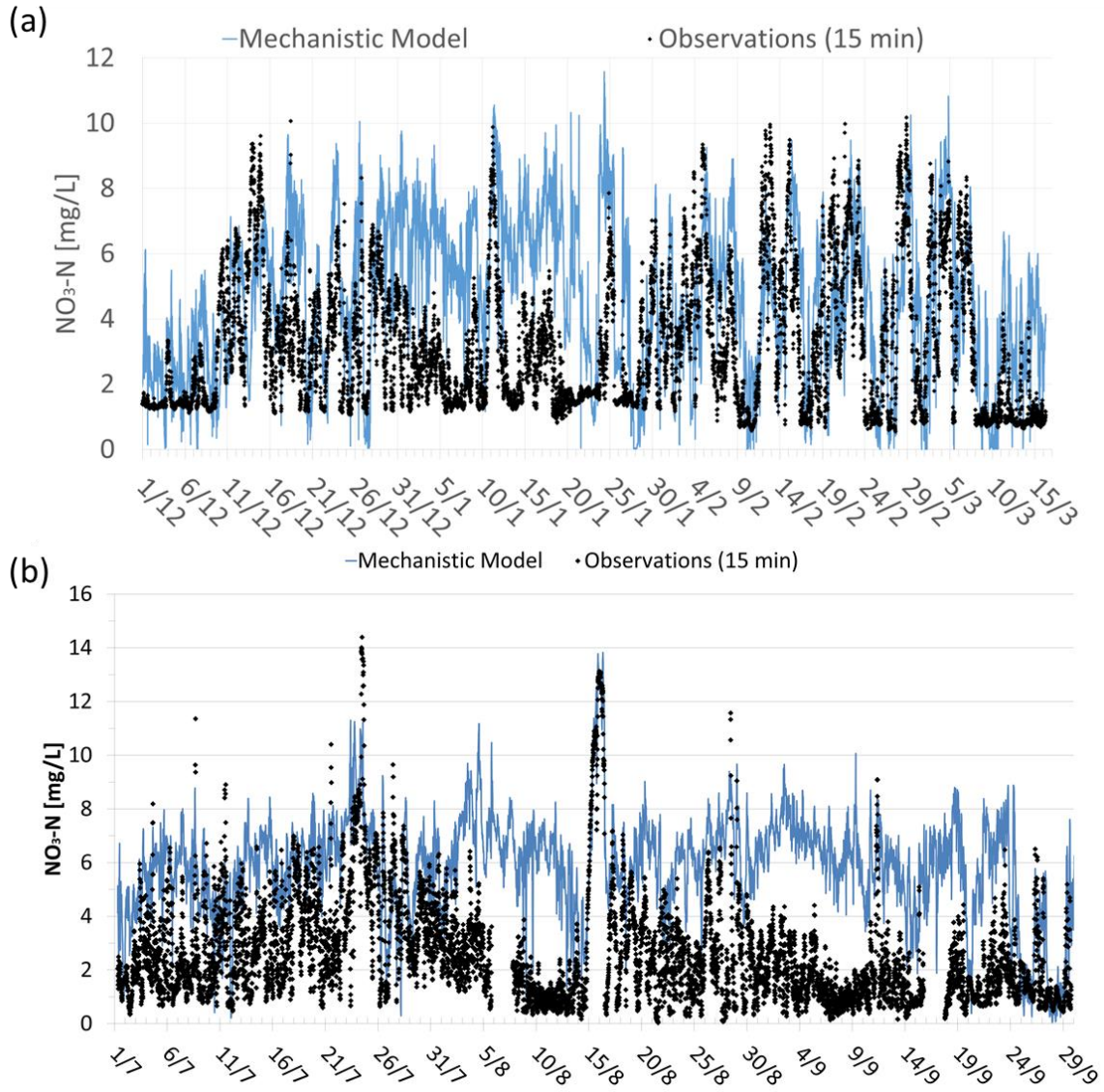


Figure 5-4: Simulation results of the mechanistic model MM simulations for the NO_3-N concentration at the outlet of the preDN stage for a) calibration, b) validation (black points: sensor observations, blue line: simulation results).

The MM is able to reproduce the daily average performances of the denitrifying filter for the calculation of the NO_3-N, NO_2-N, NH_4-N, total COD, soluble COD and TSS concentrations, with relative mean errors (ME) between 10 – 30% of the observed mean. The model results for the validation period (Figure 5-4b) are showing a significant overestimation of the sensor observations with a ME of -3,22 mgN/L, except for the short periods related to observed peak values, such as on the 25th of July, 16th of August and 12th of September. Similar trends of overestimation are observed for the total COD and TSS simulations. In general, the MM is able to simulate daily average concentrations and to follow trends observed for the nitrate, as well as for ammonia, total COD, soluble COD and TSS concentrations in the effluent of a denitrifying filter, albeit with a significant bias towards overestimation, in particular during the validation period. For more details on the mechanistic model performance, see chapter 4.

5.4.2 Selection of Data-Driven Model

Using the training data set that consisted of the feature variables and the response variable, which serve as the label for supervised learning, multiple DDMs were trained. This included fine, medium and dense decision trees, neural networks, supervised vector machines and linear regression models. The performance of each model was evaluated on the training set, in particular in terms of the RMSE criterion using a five-fold cross validation scheme. After training, the selected model was tested with the testing dataset.

Table 5-1 provides an overview of the 10 best scoring DDMs for the estimation of the residual errors between the mechanistic model simulation results and the observations of nitrate concentrations in the effluent of the preDN system. The results are ranked according to the testing RMSE values, and the selected model is highlighted. As can be seen, neural networks provide a good performance with the lowest RMSE score for both testing and training data of NO_3-N. The model with the lowest RMSE value during testing is a 3-layer neural network with 12, 25 and 1 neurons respectively for the input, middle and output layer. The 2nd to 5th positions are taken by decision tree models, indicating that models with a simpler structure would have a similar performance. However, considering the training RMSE values, where ANN models are listed at first, second and third places, it becomes clear that ANN models have a stronger performance. Overall, the neural network models consistently outperform the other model types, such as Decision Trees and Supported Vector Machines.

Table 5-1: Overview of the 10 best scoring DDM models on the training and testing data set for the estimation of the NO_3-N residual error.

Model type	Hyperparameters	Selected features	RMSE training	RMSE testing
Neural network	**Neurons: 12:25:1**	**12**	**0.058**	**0.170**
Decision tree	Minimum leaf size: 12 Surrogate splits: off	9	0.063	0.191
Decision tree	Minimum leaf size: 4 Surrogate splits: off	9	0.059	0.193
Decision tree	Minimum leaf size: 4 Surrogate splits: off	12	0.059	0.194
Decision tree	Minimum leaf size: 4 Surrogate splits: off	10	0.059	0.194
SVM kernel	Regularization strength: auto	12	0.062	0.246
SVM kernel	Regularization strength: auto	10	0.062	0.274
SVM kernel	Regularization strength: auto	9	0.059	0.295
Neural network	Neurons: 9:25:1	9	0.059	0.297
Neural network	Neurons: 10:25:1	10	0.058	0.365

An overview of the 10 best scoring data-driven models for the estimation of the total COD residual error is presented in Table 5-2, ranked by RMSE values on testing data. The best performing model is an ANN with a 3-layer structure identical to the ANN selected for the prediction of the NO_3-N residual error. The RMSE values are of the same order of magnitude as with the NO_3-N model error predictions. As was the case for nitrate, during training ANN models take up first and second places when ranked according to RMSE values. The third best fit is accounted for by a Support Vector Machine model; however, it has a poor performance with testing data.

Table 5-2: Overview of the 10 best scoring DDM models on the training and testing data set for the estimation of the total Chemical Oxygen Demand (COD) residual error.

Model type	Hyperparameters	Selected features	RMSE training	RMSE testing
Neural Network	**12:25:1**	**12**	**0.043**	**0.158**
Decision Tree	Minimum leaf size: 12	12	0.046	0.163
Decision Tree	Minimum leaf size: 12	12	0.046	0.163
Decision Tree	Minimum leaf size: 4	12	0.044	0.165
Decision Tree	Minimum leaf size: 4	10	0.044	0.165
Neural Network	12:10:1	12	0.049	0.172
Neural Network	10:10:1	10	0.049	0.180
SVM Kernel	Regularization strength: auto	12	0.044	0.195
Neural Network	10:25:1	10	0.043	0.248
SVM Kernel	Regularization strength: auto	10	0.043	0.266

Table 5-3 lists the models simulating the TSS residual error. As can be seen, here too an ANN outperforms the other model types; even though the RMSE testing values are on average two orders of magnitude larger than those of the NO_3-N and total COD models, indicating a lower overall performance of the TSS residual error prediction. Both for the training and testing data sets, decision tree models take up prominent places in the RMSE ranking. These model types are easier to interpret and more transparent, but also prone to overfitting as indicated by the RMSE values for the test data that are higher than that of the selected ANN.

Table 5-3: Overview of the 10 best scoring DDM models on the training and testing data set for the estimation of the Total Suspended Solids (TSS) residual error.

Model type	Hyperparameters	Selected features	RMSE training	RMSE testing
Neural Network	**12:10:1**	**12**	**0.054**	**0.297**
Linear Regression	Robust option: Off	12	0.074	0.330
Decision tree	Minimum leaf size: 36	12	0.052	0.335
Decision tree	Minimum leaf size: 36	10	0.052	0.335
Decision tree	Minimum leaf size: 12	12	0.045	0.342
Decision tree	Minimum leaf size: 12	10	0.045	0.342
Decision tree	Minimum leaf size: 4	12	0.041	0.350
Decision tree	Minimum leaf size: 4	10	0.041	0.350
SVM Kernel	Regularization strength (Lambda): Auto	12	0.047	0.388
Neural Network	12:25:1	12	0.045	0.391

5.4.3 Optimization of the selected Data-Driven Model

Optimal tuning of the number of features and the model's hyper-parameters was done to avoid under-fitting or over-fitting conditions over both training and testing periods. This was achieved by selecting the simplest model that still has a low RMSE score by removing features that are not contributing to the precision of the DDM (see 'Feature Selection' below) and by changing the values of hyper-parameters of the model, such as the number of neurons in the hidden layer in the case of a neural network (see 'Hyper-parameter & Model Selection' below).

NO_3-N DDM Feature Selection

Not all features contribute equally to the response variable; some can actually disturb the model output and decrease performance. Model improvement is achieved by removing features that have a low predictive power. For an optimal configuration, it is necessary to retain only those features that contribute significantly. The selection of which features to retain is based on a feature-ranking algorithm.

Here, the F-test was applied for feature selection in order to identify the variables that best fit the data and to be able to remove any features that are redundant. For each validation fold, the features that had a zero or near zero F-test score were removed, such that only the highest ranked features with a significant contribution (F-test above a value of 200), were retained. Figure 5-5 shows that 10 out of 12 features were identified to be influential to some degree, of which five have an infinite importance on the model outcome. The most influential variables are considered to have a F-test score of at least 200, and those are water temperature, COD influent concentration (COD_IN), NO_3-N recirculation concentration (NOx_REC), TSS concentration in settled water (TSS_SW), NO_3-N concentration in effluent modelled by the mechanistic model (NOx_OUT_MOD), the flowrate of settled water (Q_SW) and the weather type (Weather). Of these features, all but one are measured variables that serve to characterize the influent water quality; the exception being NOx_OUT_MOD which is a simulated output of the mechanistic model.

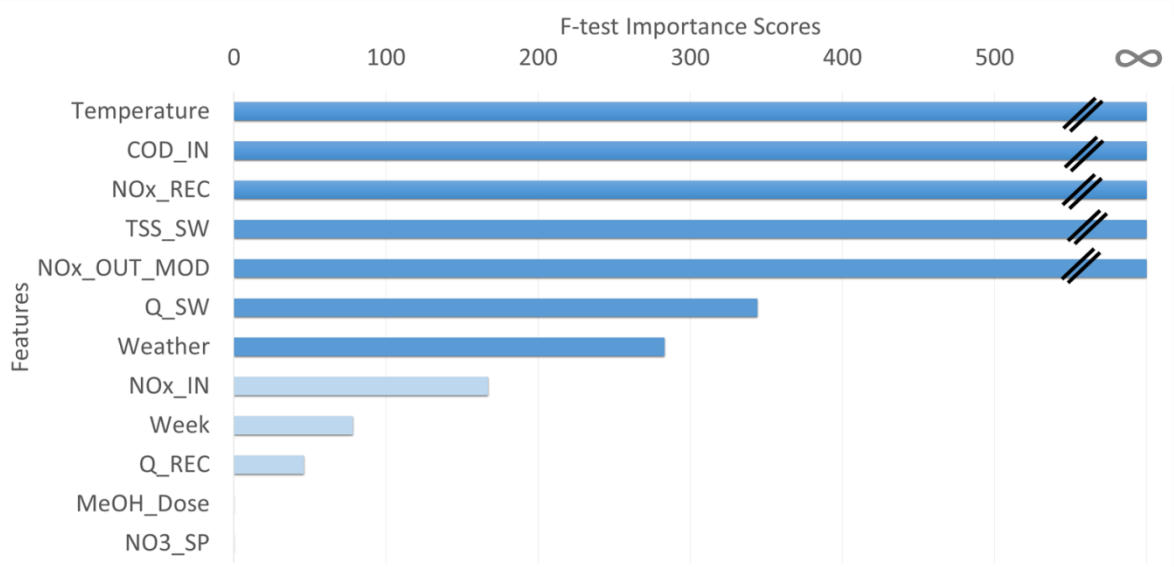


Figure 5-5: Features used by the ML model and their ranking by F-test scores for the NO_3-N residual error.

The variables NOx_IN, which is the measured NOx-N concentration in the preDN influent, and Q_REC (the measured flow rate of the recirculation of nitrified water [m3/d]), are both identified as less influential. Their intrinsic information content is actually included in variables that have been classified as influential, namely the NOX_REC, which accounts for the significant source of nitrates and nitrites in the influent; and the Q_SW (the settled water flow rate). The latter is proportional to the flow rate of recirculated nitrified water; if Q_REC increases, then the Q_SW decreases.

NO_3-N DDM Hyperparameters & Model Selection

Another important step in DDM model development is the identification of the hyperparameters that allow optimizing the model to best fit the data. A hyperparameter is a parameter that describes a configuration that is external to the model itself, like the learning rate, the number of estimators and the type of regularization. Hyperparameter tuning was performed using Bayesian optimization with a loss function calculating the mean squared error (MSE). An acquisition function determines the next set of hyperparameter values to try.

The hyperparameters that have been optimized for the neural network are 1) the number of fully connected layers with a range of 1, 2, and 3 fully connected layers; 2) the first layer size with range [1:300]; 3) the second layer size with range [1:300]; 4) the third layer size with range [1:300]; 5) the activation function, a search among ReLU, Tanh, None, and Sigmoid; 6) the regularization strength (Lambda). The optimization algorithm searches among real values log-scaled in the range [1e-5/n,1e5/n], where n is the number of observations.

The optimal network configuration was determined by simulations with a smallest configuration of an input, hidden and output layer and by varying the number of nodes in the hidden layer.

The best performing ML model was selected from the lowest RMSE value on calculating the NO_3-N error between the MM and the measurements for both the training and the testing data. As indicated in Table 5-1, during training the ANN model that applies 12 features has a score similar to the ANN model using 10 features; however, during testing the 12 features model has a lower RMSE of 0.17 compared to the 10 features model (RMSE: 0.37). It was therefore decided to select the three-layer feedforward neural network with 12 neurons in the input layer, 25 neurons in the middle layer and one neuron in the output layer with a Rectified Linear Unit (ReLU) activation function.

Total COD DDM Feature Selection

The results of the feature selection by applying the F-test for the data-driven model estimating the residual error for total COD are shown in Figure 5-6. Two features were identified as highly influential for improving prediction performance; these are the observed COD influent concentration (COD_IN) and the observed TSS concentration in the settled effluent. Of very high to high importance are the measured settled water flowrate (Q_SW) and water temperature, the simulated total COD effluent concentration by the MM (COD_OUT_MOD), the measured recirculation flow rate containing nitrified water (Q_REC) and the measured NO_3-N concentration in the recirculated water (NOx_REC). The observed nitrate influent concentration (NOx_IN) receives a low F-test scores that falls under the minimum value of 200 and thus is considered to be of lesser significance. As was the case with NO_3-N, the feature describing the type of day (WEEK), that of the methanol dosing (MeOH_Dose) and the nitrate set point (NO3_SP) received a low score.

For total COD model prediction performance, the weather type that indicates dry weather conditions or rainfall receives a very low impact score. The main significance in daily operations of ‘weather’ type is that it controls the recirculation flow rate of nitrified water from the effluent of the NIT stage to the entrance of the preDN stage, which during rainfall is practically halted. This allows increasing capacity for the denitrification filters to treat more effluent of the primary settling phase. Features that characterize the recirculation flow rate (Q_REC) and the contribution of settled water (e.g. Q_SW and TSS_SW) have high to very high and infinite scores, so that the weather feature is effectively compensated for.

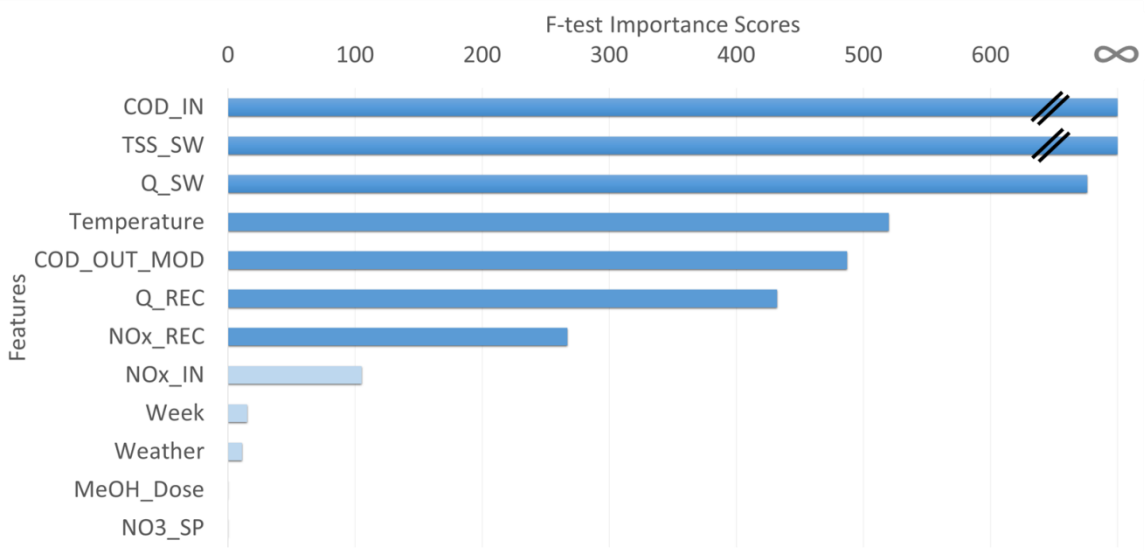


Figure 5-6: Features used by the ML model and their ranking by F-test scores for the total COD residual error.

Total COD DDM Hyperparameters & Model Selection

The procedure applied to optimize hyperparameters is identical to the one applied for the NO_3-N model. The best performing ML model was selected from the lowest RMSE value on calculating the total COD error between the MM and the measurements for both the training and the testing data. The selected model is a three-layer feedforward artificial neural network with 12 neurons in the input layer, 25 neurons in the middle layer and one neuron in the output layer with a Rectified Linear Unit (ReLU) activation function; identical to the ANN selected for the nitrate model.

TSS DDM Feature Selection

Figure 5-7 shows the results of the ranking of features by the F-test method for the TSS data-driven model. Those features that received scores of infinite and very high importance are the mechanistically modelled TSS effluent concentrations (TSS_OUT_MOD), the measured settled water flowrate (Q_SW) and the measured water temperature. The first two features directly affect the input of TSS into the filters; the temperature is a controlling feature of the biological activity and metabolism rate of the biomass and its ability to capture and absorb particulate organic matter.

Of significant importance in the prediction performance of the 'TSS- model' are those features that describe the measured total COD and TSS concentrations in the influent (COD_IN) and settled effluent (TSS_SW) respectively. The recirculation flow of nitrified water has considerate influence as well as it effectively indicates the contributively proportion of settled water in the total inflow to the preDN stage. Finally, the indicators for type of day (week), weather, the amount of methanol dosing (MeOH_Dose) and the nitrate effluent set point (NO3_SP) are considered to have an insignificant contribution.

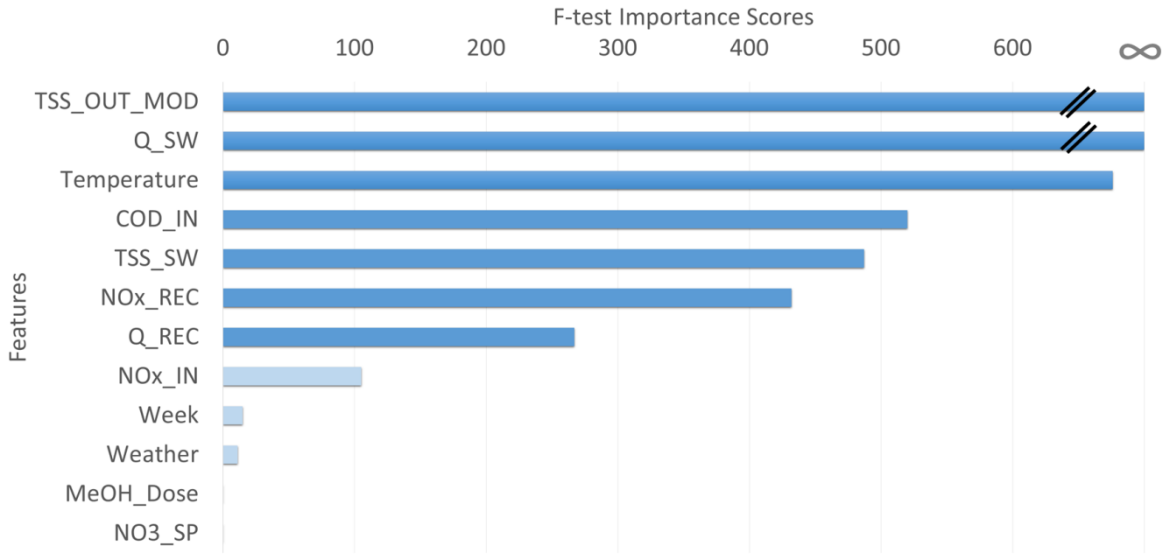


Figure 5-7: Features used by the ML model and their ranking by F-test scores for the TSS residual error.

TSS DDM Hyperparameters & Model Selection

Here too, the steps described earlier have been applied to optimize hyperparameters. The best performing ML model was selected from the lowest RMSE value on calculating the TSS error between the MM and the measurements for both the training and the testing data. The selected model is a three-layer feedforward artificial neural network with 12 neurons in the input layer, 10 neurons in the middle layer and one neuron in the output layer with a Rectified Linear Unit (ReLU) activation function. This ANN model contains 10 neurons in the hidden middle layer; which makes it different from the neural networks selected for the nitrate and total COD model.

5.4.4 Training and Testing of Hybrid Model

For the calculation of the hybrid model's output, the DDM calculations of the model error for the effluent variables of NO_3-N, COD and TSS are added to the MM simulations of the corresponding effluent variables. The results with the training and testing data are now presented for each of the variables.

5.4.4.1 Effluent NO_3-N

Figure 5-8 compares the high-frequency observations with the nitrate concentration simulations for the training (figure a) and testing (figure b) periods by the hybrid and mechanistic models respectively. The outputs of the hybrid model are much closer to the observed values than to those of the mechanistic model over the two periods.

Interesting to notice are the model simulated values lower than 1.0 mgNO_3-N/L (e.g. between January 20th and 30th) that are lower than the measured data that never seem to drop below 1 mg NO_3-N/L, the appropriate detection range. Conversely, the simulations seem to correspond with the laboratory analysis results for composite samples that are indicating daily averages below 1 mg NO_3-N/L on occasion. It seems that the detection limit of the sensors, that falls around 1.0 mgNO_3-N/L, is causing the measurement values to stay above 1.0 mgNO_3-N/L makes the comparison with the simulation void.

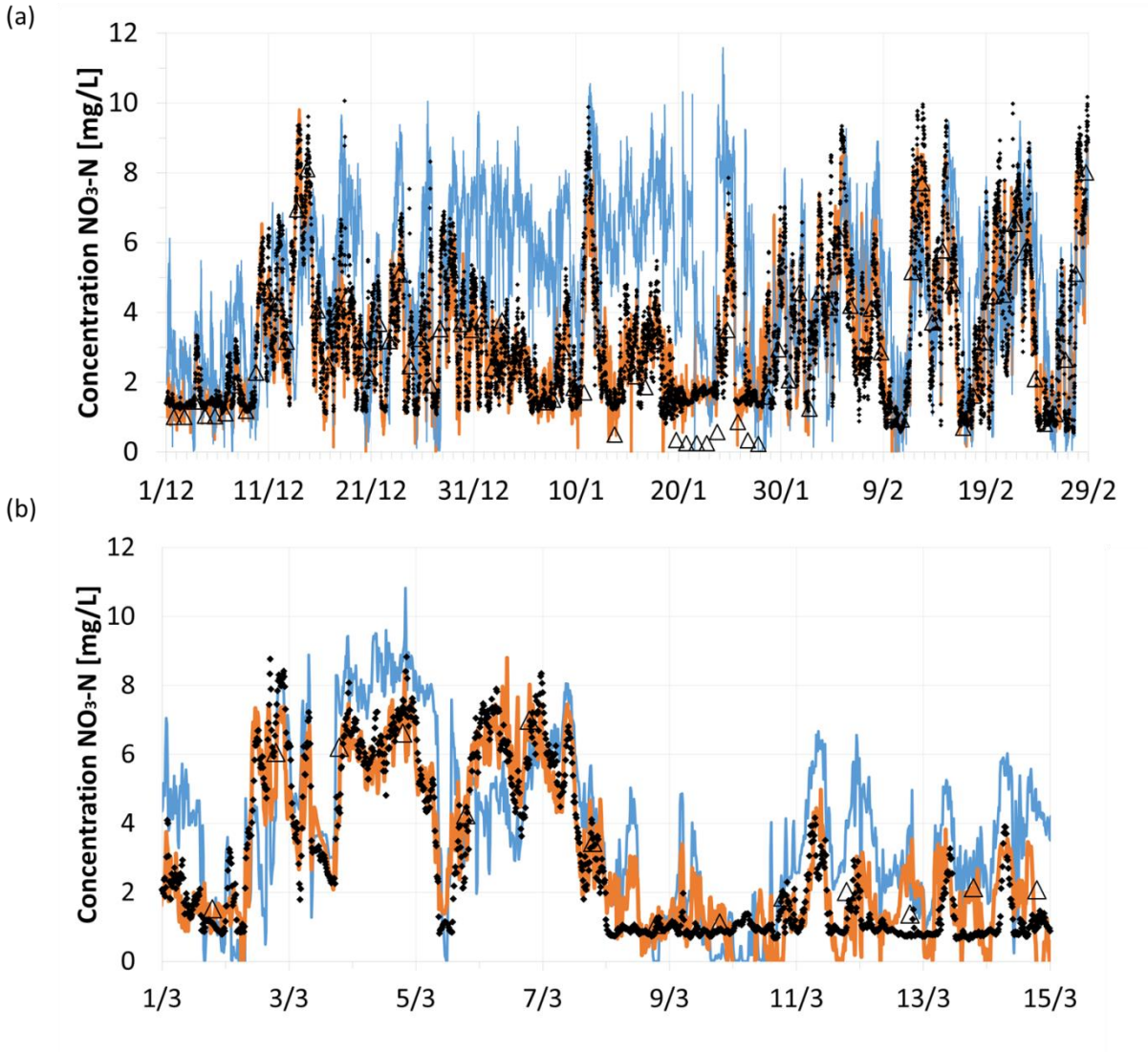


Figure 5-8: Results of the training (a) and testing (b) of the hybrid model (orange line) for the NO_3-N concentration at the outlet of the preDN. Blue line: mechanistic model simulation; orange line: hybrid model results; black points: high-resolution observations; black-outlined triangles: laboratory composite observations.

Figure 5-8b shows that the hybrid model simulates well the trends and variability of the nitrate concentration over the testing period. Towards the end of the testing period however, the relative improvement of the HM seems to be lower. Starting from March 8th a period occurs in which the HM has difficulties to compensate for the residual error of the MM simulations, and this especially in those cases when the MM predicts low NO_3-N concentrations of 1 mgN/L or less. The daily composite laboratory samples however show average values between 1.0 and 2.0 mgNO_3-N/L, which conceal the high variability in the 15-minutes values that can temporarily reach extreme high and low values.

This could be due to different reasons, such as a lack of good quality measurements (due to the measurement detection limit of the nitrate sensor) or less reliable simulations by the MM model in the lower range of nitrate and nitrite concentrations. A lower performance with the test data could indicate that the data from the training period does not contain a sufficiently wide range of information to train the DDM. Still, the Janus coefficient of 1.0 has the optimal value and does not indicate an error of over- or underfitting.

The statistical performance indicators are listed in Table 5-4 for the simulations of the effluent variables at the preDN outlet with the hybrid model (HM) and are compared to those with the mechanistic model (MM). In this table, the performance of the models is relative to the observed data from the calibration/training and validation/testing periods; MM-calibration (01.12.2019 – 15.03.2020); MM-validation (01.07 – 30.09.2020); HM-training (01.12.2019 – 29.02.2020) and HM-testing (01.03 – 15.03.2020).

As can be seen for the nitrate effluent, the HM outperforms the MM with relative mean errors that are much smaller during training and testing for the HM compared to the calibration and validation of the MM. The evaluation criteria in the Table 5-4 show also in more detail that for nitrate the performance of the HM is good, even though the RMSE is significantly higher, since that is affected by the weights given to larger errors in its calculation (e.g., in the RMSE, the errors are squared before they are averaged). During the testing phase, the quality criteria show similar values, which is to be expected. The Janus coefficient, with an optimal value of 1.0, indicates that the RMSE values are near identical, such that the model can be considered well-validated for nitrate simulations. Overall, even though the relative MAE and relative RMSE are ranging from 24% to 32%, the scores remain globally acceptable given the highly dynamic conditions of the process and considering the measurement errors of the sensors.

Table 5-4: Statistical performance indicators for the mechanistic (MM) and hybrid (HM) models during calibration/training and validation/testing with high-frequent data for the NO_3-N, total COD and TSS effluent variables simulated at the outlet of the preDN process. Shown are the relative errors compared to the observed mean. N = Number of observations; ME = Mean Error; MAE = Mean Absolute Error; RMSE = Root Mean Square Error; ȳ= observed mean.

			Calibration (MM) or Training (HM)					Validation (MM) or Testing (HM)					
Variable	**Model**	**Unit**	**N**	**Mean obs (ȳ)**	**ME/ ȳ**	**MAE /ȳ**	**RMSE/ ȳ**	**N**	**Mean obs (ȳ)**	**ME/ȳ**	**MAE /ȳ**	**RMSE/ ȳ**	**Janus coef.**
NO_3-N	MM	gN/m^3	10656	2.89	-47%	67%	86%	8015	2.62	-123%	129%	146%	1.5
NO_3-N	HM	gN/m^3	8736	2.94	-11%	25%	34%	1440	3.23	8%	24%	32%	1.0
COD	MM	gO_2/m^3	10656	89.07	-1%	14%	20%	7218	64.84	-53%	55%	60%	2.2
COD	HM	gO_2/m^3	8541	89.14	<1%	6%	9%	1440	87.03	2%	15%	19%	2.1
TSS	MM	g/m^3	10656	21.20	-11%	23%	39%	7399	20.12	-36%	46%	63%	1.5
TSS	HM	g/m^3	9212	28.58	-5%	12%	18%	1440	23.54	-33%	32%	39%	2.3

5.4.4.2 NO_3-N Daily Values

In contrast to high-resolution data with a time interval of 15 minutes, which can capture short-term fluctuations in wastewater conditions and are therefore more appropriate for real-time control operations, daily average values may be more appropriate for analysis purposes because they provide a noise-reduction and provide a better indication of overall performance. Wastewater characteristics can fluctuate rapidly over short time intervals due to various factors, such as changes in weather conditions or sudden changes in upstream activity. These fluctuations can introduce fast dynamics into the data, making it difficult to identify overall trends and patterns. By using daily average values, the noise is reduced, and trends can be better observed. In addition, daily average values can provide a better representation of the overall performance of the wastewater treatment process over an

extended period of time, rather than the short-term fluctuations when evaluating the model performance scores on high–frequency sensor data.

Figure 5-9 plots the daily averages of nitrate effluent concentrations for the simulations by the mechanistic and the hybrid model as compared to the daily averaged inline measurements for the training period. Plotted as well are the daily composite laboratory samples. It is noticed that the HM simulations are much closer to the daily averaged inline observations, with a R^2 value of 0.95, as compared to the daily averages of the MM with a R^2 of 0.21. The MM tends to overestimate the nitrate effluent concentrations, especially in the period between 24.12.2019 to 25.01.2020. The hybrid model is capable to correct for this and reduce the overall relative mean error to 2% under training conditions (see Table 5-4).

The daily laboratory analysis results follow in general the daily average values of the inline sensor measurements, with a few notable exceptions (e.g. a short period from 19 to 22 of January, 1, 12 and 28 February). The HM is here not able to compensate, since the data-driven model was trained on the high-resolution measurements and is thus unaware of the discrepancy between laboratory and sensor observations.

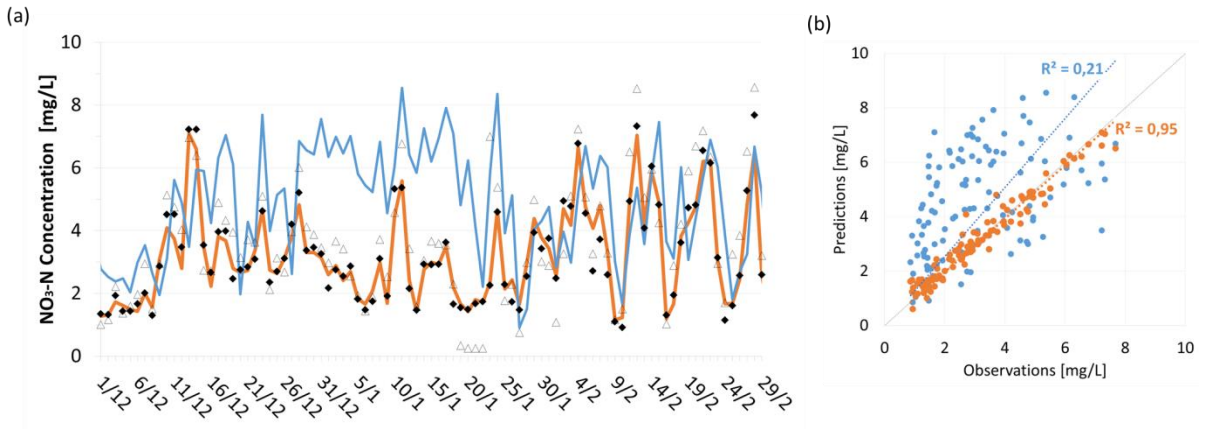


Figure 5-9: (a) Performance of hybrid model (orange) and mechanistic model (blue) evaluated for NO_3-N effluent daily average data for the training of the HM at 15 minute interval; orange line: connected points of the daily average of the HM simulations; black diamonds: daily average of the inline observations; black-outlined triangles: daily laboratory samples. (b) Parity plot of the simulated data of the MM and HM compared to the observed data using daily averaged inline values.

Concerning the results of the testing period with the daily averaged values (Figure 5-10), the maintained overall high performance of the HM compared to the daily averages of the sensor measurements with a R^2 of 0.97 is noticed in figure b. This could be explained by the fact that the daily averages are to a much lower degree affected by the high-frequency dynamics around the 1 mgNO_3-N/L that falls around or under the detection limit of the monitoring equipment, as is the case with the high-resolution data with a time interval of 15 minutes, shown in Figure 5-9.

The daily averages of the simulations with the knowledge-based model are following the trend of the operational data, but beginning on March 8th they start to show an overestimation. This is confirmed in the parity plot in Figure 5-10b, where the mechanistic model only has a R^2 value of 0.21 for the same data period.

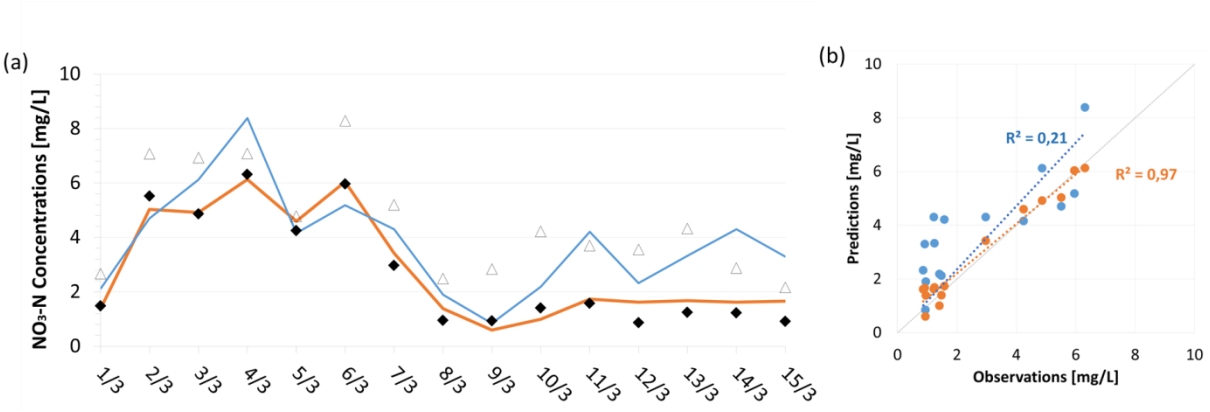


Figure 5-10: (a) Results of the test of the MM (blue) and HM (orange) models for the NO_3-N concentration at the outlet of the preDN using daily averaged values of the 15 minutes simulation results. Blue line: connected points of daily average of the mechanistic model simulations; orange line: connected points of daily average of the hybrid model simulation results; black diamonds: daily average of the online observations; black-outlined triangles: daily laboratory samples. (b) Parity plot of the simulated data of the HM and MM models compared to the observed data using daily averaged values.

5.4.4.3 Effluent COD

Concerning the effluent COD variable shown in Figure 5-11, the developed hybrid model is able to simulate the effluent variable much better than the mechanistic model. This is particularly true for the training period (Figure 5-11a), where the neural network was capable to correct for the dynamics missed by the mechanistic model, for instance on January 19th. The relative error (see Table 5-4) ME/ŷ is less than 1%, showing no bias, and the MAE and RMSE are all below 10% of the observed mean for this dataset.

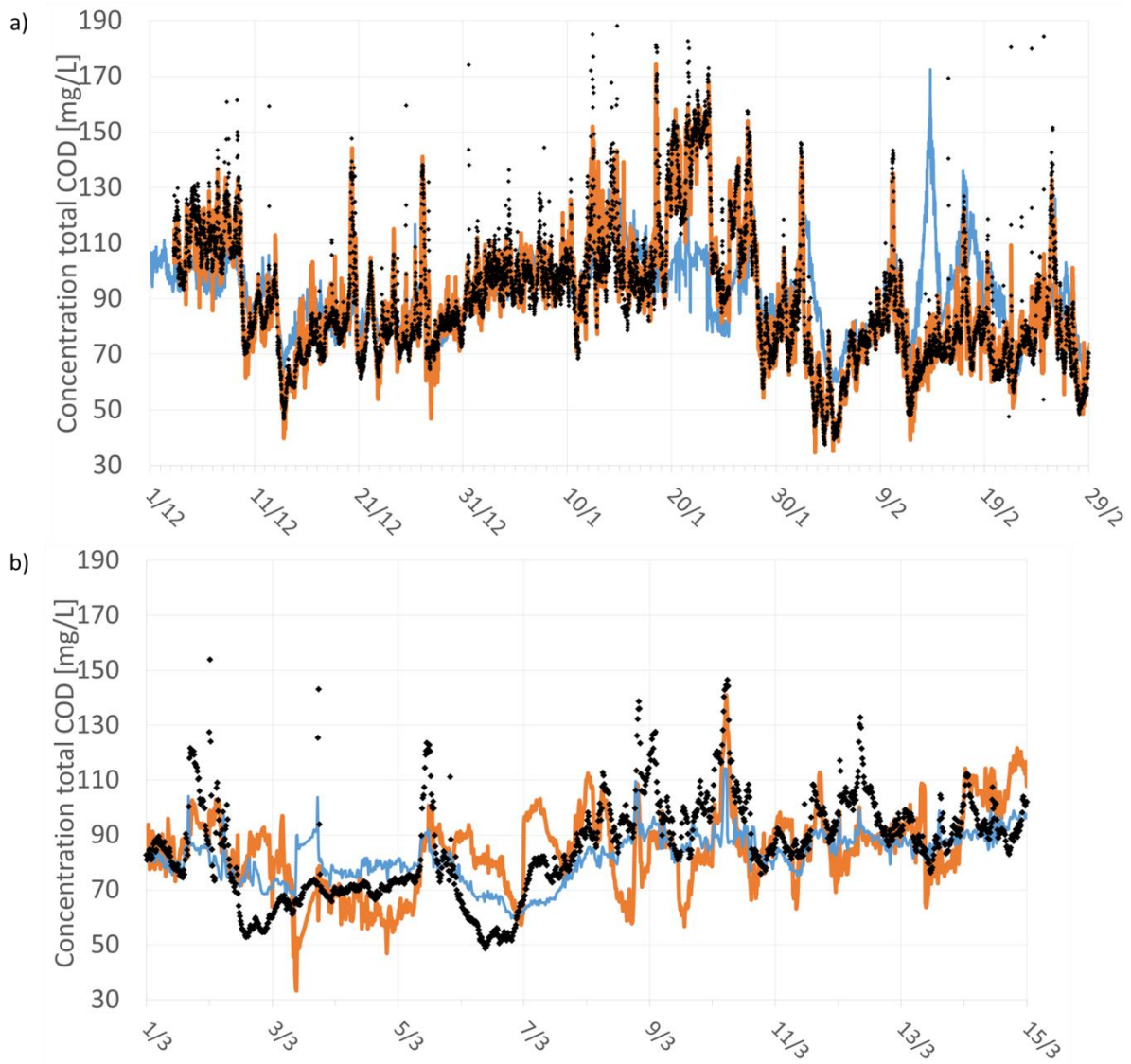


Figure 5-11: Results of a) the training and b) testing of the hybrid model for the total concentration of COD in the effluent with high-frequent data at the outlet of the preDN stage. Blue line: simulation of the mechanistic model; orange line: hybrid model simulation results; black points: high frequency observations.

However, the scores are not as good when the hybrid model is confronted with unknown high-resolution data during the testing phase (Figure 5-11b). The ML model does not seem to be able to describe the amplitudes of the variation very well and at some moments simulates dynamics in the opposite direction (e.g. on the 14th of March). The performance of the hybrid model with this dataset is not much better than the one of the mechanistic model. The HM tends to apply a too large correction, thereby simulating values that are too low compared to the measurement data. Nonetheless, the error of the hybrid model is still lower than that of the mechanistic model for the total COD in the effluent during the same period, indicating that the ML model component can compensate for missed dynamics.

Regarding the daily averaged values of the simulations (Figure 5-12), the hybrid model is capable to very closely simulate the effluent total COD concentrations at the preDN stage during the training phase with a R^2 value of 0.98 (not shown). This is also true for those short periods of time where the mechanistic model has a difficulty in simulating the correct COD concentration, such as between 17 –

22 February. The MM doesn't follow the trend, while the HM is fully capable to compensate for that residual error for these days.

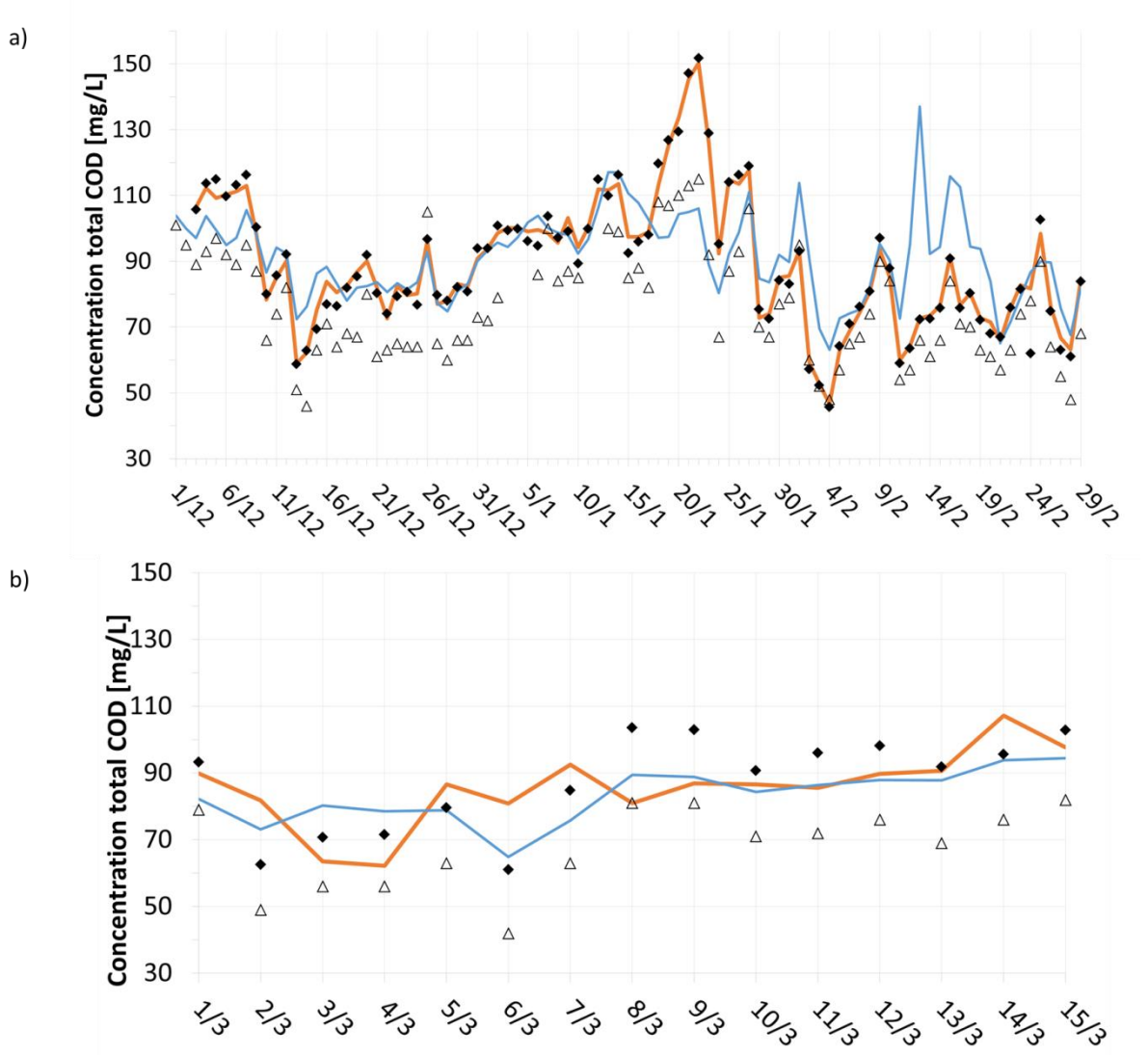


Figure 5-12: Performance of hybrid model (orange) and mechanistic model (blue) during (a) training and (b) testing periods evaluated for daily average data for the effluent total COD simulations at the preDN stage; orange line: connected points of the daily average of the HM simulations; blue line: connected points of the daily average of the MM simulations; black points: daily average of the inline observations; black-outlined triangles: daily laboratory samples.

The performance of the HM during the testing period (Figure 5-12b) is not only lower compared with that of the training data, which is of course expected, but also compared with the MM performance for the same time period with regards to the RMSE of 16.6 gO_2/m^3 (HM) versus a RMSE of 13.3 gO_2/m^3 (MM). However, the ME of the HM (1.5 gO_2/m^3) is still significantly lower than the ME of the MM (3.9 gO_2/m^3) indicating that the HM is less affected by bias. Notwithstanding the fact that the contribution of the HM remains an added value, there are some instances during the testing period where its simulation performance is quite modest. Such is the case between March 8th and 12th, where the HM does not correct the MM sufficiently. Effectively, both model's predictions remain about 10% below the daily averaged observed values.

5.4.4.4 *Effluent TSS*

For the effluent TSS variable shown in Figure 5-13, the hybrid model is again able to simulate the training and testing data with a higher precision than the mechanistic model. The errors listed in Table 5-4 indicate relative errors for the training of -5%, 12% and 18% for ME/ŷ, MAE/ŷ and RMSE/ŷ respectively for the HM; values that are half of those indicators for the MM (-11%, 23%, 39%). In fact, the mechanistic model has difficulties to simulate the high TSS observations while at other times the model output is much higher than the observations (i.e. period 9 – 20 February). This might be related to a high variability in the suspended solids content of the influent that is a mix of settled influent and recirculated nitrified water. At some moments, plant operators can decide to lower or stop the recirculation to the entrance of the preDN stage; thereby increasing its capacity; but this can have short-term effects with a sudden flow that is high concentrated with suspended solids reaching the preDN stage. On the other hand, it might be that the MM cannot well simulate the filtering aspect of the biofilter at such short-term intervals.

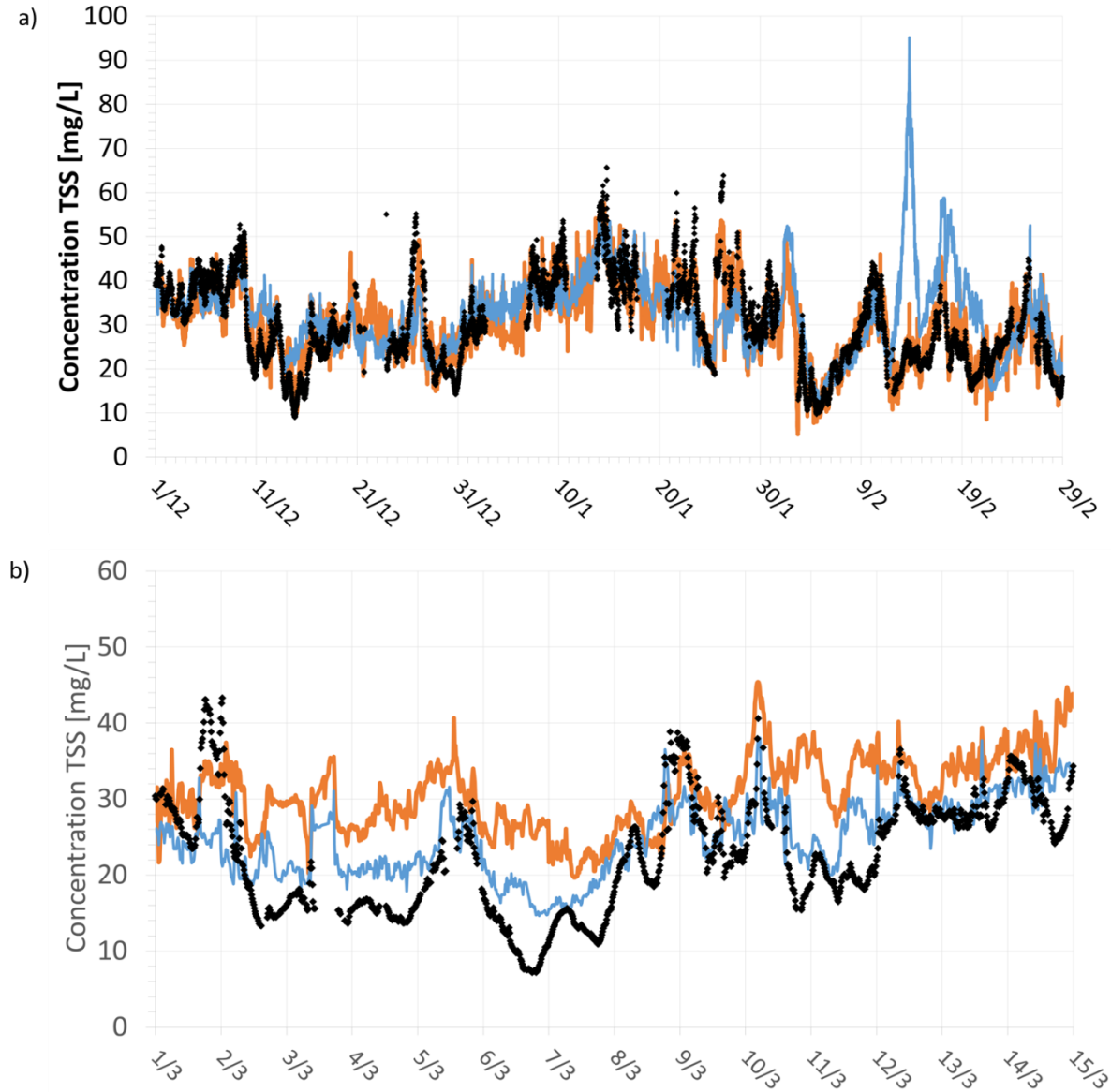


Figure 5-13: Performance of hybrid model (orange) and mechanistic model (blue) with high-frequent data during (a) training and (b) testing periods of the TSS concentration at the outlet of the preDN stage. Blue line: simulation of the mechanistic model; orange line: simulation of the hybrid model results; black points: high frequency observations.

For the testing dataset (Figure 5-13b), the DDM component is unable to adequately calculate the residual of the mechanistic model error and overcorrects the TSS concentration for most of the time. This is reflected in the ME for the testing period (-7.76 g/m^3), significantly larger compared to the ME for the training period (-0.34 g/m^3), indicating the model's tendency to overestimate the effluent TSS concentrations and to underestimate the system's filtration capability. More research is needed to better understand the dynamics that are at play here. It could be that different conditions (variations on daily and monthly basis) during testing period that were not observed during the training period caused a lack in the performance of the DDM. A larger training period, spanning at least a year to account for seasonal dynamics, might increase the DDM's reliability.

5.4.4.5 Daily values of effluent TSS predictions

The daily averaged results for the effluent TSS simulations (Figure 5-14) show that during the training period, the HM is able to well compensate for the residual error of the MM. This is also visible during the period of February 10th to 20th, where the MM experiences a large deviation from the observations. The DDM component of the HM is well trained to compensate and predict the residual errors observed during this period.

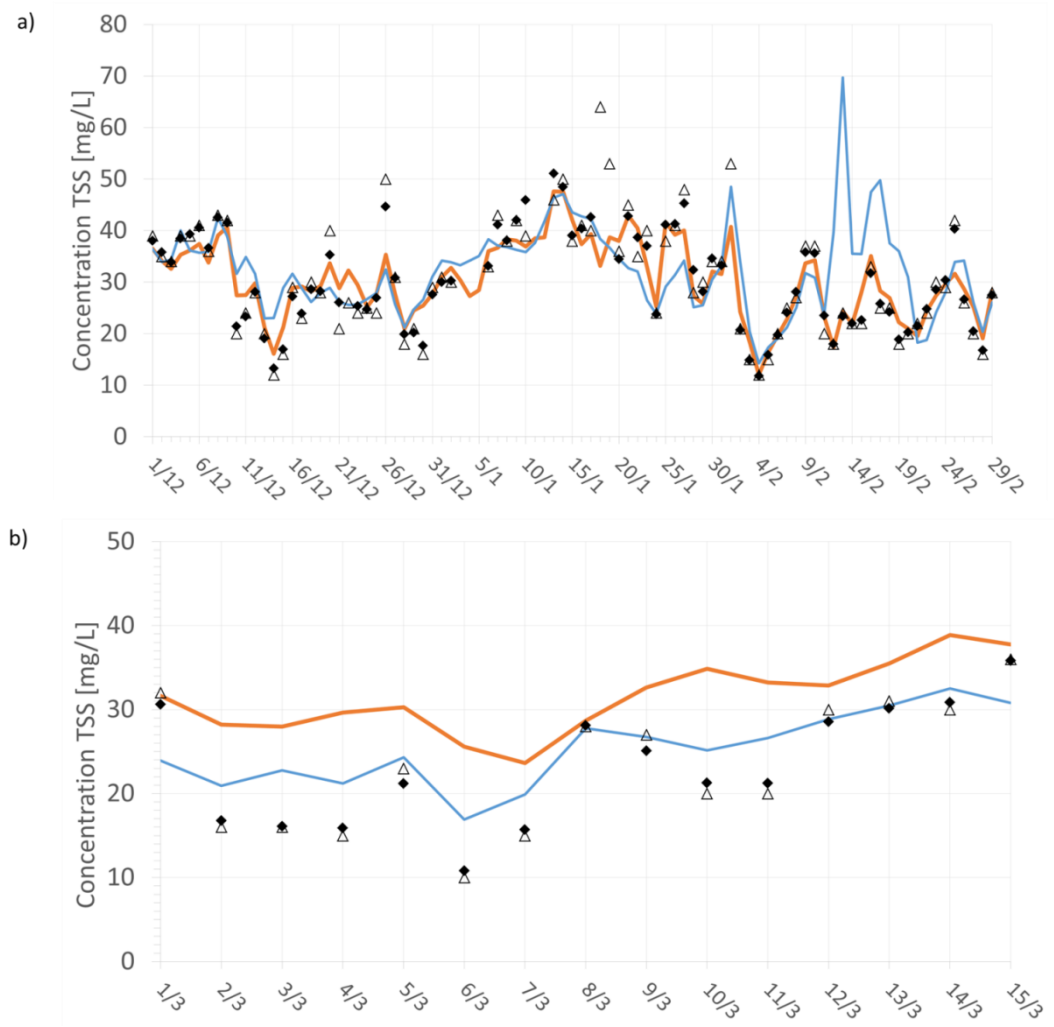


Figure 5-14: Performance of hybrid model (orange) and mechanistic model (blue) during (a) training and (b) testing periods evaluated for daily average data for the TSS effluent simulations at the preDN stage; orange line: connected points of the daily average of the HM simulations; blue line: connected points of the daily average of the MM simulations; black points: daily average of the inline observations; black-outlined triangles: daily laboratory samples.

However, for the testing period (Figure 5-14b) the HM has poor performance as indicated in Table 5-4 by the relative MAE (32%) and relative RMSE (39%) that corresponds to a RMSE of 9.2 mg/L. The model's bias is significantly negative (-32% mg/L), reflecting the HM's tendency to overestimate the effluent TSS concentrations, as clearly visible in the graph of figure b.

5.5 Conclusions

A parallel hybrid model was developed in a cooperative structure, in which a calibrated/validated mechanistic model is corrected by a data-driven model to improve the accuracy of water quality simulations at the outlet of a predenitrification biofilter of a large WWTP. Different ML techniques were tested to obtain an accurate estimate of the residual error of the mechanistic model; and this for NO_3-N, total COD and TSS concentrations in the preDN effluent. These data-driven models were given the simulated data of the effluent water qualities by the mechanistic model, as well as information on the influent flow and water quality and simulated treatment parameters (such as the controller set points).

The best results were obtained with a feedforward neural network using a ReLU activation function containing one hidden layer with 25 neurons. The simulations obtained over the training period and the testing period show that the precision of the hybrid model is much higher than that of the mechanistic model alone. This is in particular the case for NO_3-N, which is, in this case, the most important effluent quality variable that serves as an indicator of the process's denitrification performance. The hybrid model has a relative mean error of -11% during training compared to 47% for the mechanistic model, showing a 4 times lower error of the effluent nitrate concentrations. During testing, the DDM is still able to compensate for the mechanistic model's residual error, albeit to a lower degree, which is expected.

The Janus coefficient, which evaluates the capability of the model to perform on unseen conditions, has values of 1.0, 2.1 and 2.3 for NO_3-N, total COD and TSS model simulations respectively. Such coefficients are considered to be good and express the validity of the models. When taking into account the highly dynamic conditions observed in the secondary treatment biofiltration lane of the super-large wastewater treatment plant under study; the considerable measurement errors observed in the high-resolution monitoring of quality effluent variables; and the sensitivity of the model calculations to outliers and faults in the input data; it is fair to conclude that the modelling results could further benefit from more intense (and automated) data preparation and reconciliation processes, as well as more data from longer training and testing periods.

Overall, the results of this study with the developed hybrid model indicate that the data-driven component captured sufficient residual information to compensate for the inaccuracy of the mechanistic model. This has improved the confidence in the model's output.

6 Non-linear Hybrid Model Predictive Control of Methanol Dosing for Enhanced Nitrogen Removal in a Denitrification Biofilter – a case study as proof-of-concept

Keywords: *non-linear Model Predictive Control, Hybrid models, data-driven models, mechanistic models, methanol dosing, denitrification, biofiltration, wastewater treatment.*

Abstract

The wastewater sector is undergoing rapid changes thanks to the rising importance of sustainability and the principles of water conservation and circular economy, in which wastewater is considered a resource of water, energy, heat and chemicals. These changes are occurring as well because of the ongoing digitalization of the water industry with the application of inline monitoring equipment, advanced process models and data management systems. In order to meet these demands, and the evolving stricter regulations on effluent water quality, more reactive and complex treatment systems have been put in place, that often operate close to their design capacity, thereby reducing operational and economic margins. Process control has become an important tool to operate a system towards a defined goal (i.e. removal of nitrates) despite the many disturbances experienced in wastewater treatment. Advanced process control like model predictive controllers use a model to make predictions of future states and an optimizer to select the optimal sequence of actions to best reach a desired outcome.

As previously described in chapter 5, the hybridization of a data-based model using AI techniques with a mechanistic model reduces the model's prediction errors significantly. Here, an operational application of the hybrid model is tested in the context of a methanol dosing model predictive controller applied to a denitrification biofilter model. The thus created Hybrid Model Predictive Controller uses a moving simulation window (with 15-minute intervals) to predict future NO_3-N states based on multiple optimizer scenarios consisting of sequences of proposed control actions, in this case methanol dosing. It selects the best among the scenarios simulated on the basis of the overall error (RMSE) to the desired set point. The impact of this optimal dosing action on the real physical system is studied here by using a second mechanistic biofilter model that represents reality. Different trial configurations of (Hybrid) Model Predictive Controllers were tested, involving different levels of quality in updating the hybrid model to reflect reality. The results of this study indicate that a Hybrid Model Predictive Control is able to significantly improve the effluent water quality when compared to the current dosing control that was used as reference. However, not unexpected, this results in higher operational costs of methanol usage. Further analysis of these results is required.

6.1 Introduction

Improving the performance and efficiency of wastewater treatment plants involves strengthening the control of the processes that take place there. Optimization of energy performance, reduction of the use of reagents, of greenhouse gas emissions and of overall environmental impact of operations, the improvement of the recovery of resources and a stable effluent quality are nowadays the objectives for the wastewater treatment sector. The use of mathematical modelling in the wastewater sector has been established since the 1990's for research, system design, simulation of processes and operator training. More recently, the development of digital twins that 'mirror' the behaviour of a physical system or plant, and that receives data in (near) real time, allow for predictions of future states. An

important field of research aims to study the potential of data-driven models to achieve an even higher aspired performance of wastewater treatment plants (Schneider et al., 2022).

Model-based Predictive Control (MPC) can be used to optimize complex systems that have many multivariable constrained control issues. MPC is implemented to reduce the variation in process variables in industrial applications, which in turn leads to more efficiency and an increased throughput. The development of tools for online performance optimization using models is crucial for exploiting the full potential of digitalization (Stentoft et al., 2019). A predictive process model that has access to operational data (a digital twin) when combined with an optimizer and applied to process control, can be defined as a model-based predictive controller. MPC's have characteristics that enable them to control unstable and non-linear processes, to be flexible enough to be applicable under different conditions, to be easy to extend to multivariate control (i.e. Multi-input Multi-output, MIMO) and to be designed in a logical and transparent manner (Alex, 2022).

In this chapter, a hybrid model predictive controller (HMPC) that uses the hybrid model developed in chapter 5 is used to optimize the dosing of external carbon in the form of methanol in a denitrification biofilter. The hybrid model combines a mechanistic dynamic simulation model with a data-based model using high-frequent operational data. The hybrid model reduces the range of prediction errors and makes predictions of future states more reliable. The optimizer searches for the optimal action that brings the predicted process behaviour as close as possible to the desired behaviour using a minimal of control effort. To test the HMPC, it is applied to a virtual reality, represented by a second mechanistic model that is different to the model included in the HMPC.

Model-based Predictive Control (MPC) is a method to optimize process control. It applies the interaction between a model and an optimizer to calculate the optimal proposed sequence of control actions based on model predictions of future states and variables of the system subject to disturbances within a (short) prediction horizon, generally hours or minutes (Figure 6-1). Only the first proposed action is actually implemented on the process at time step t. The optimization iteration then moves to the next time step (t+1) to repeat the cycle (Vanrolleghem, 1995). For it to work in real-time, the models used in MPC should be fast and adaptable.

MPC requires a prediction of future process behaviour over a time interval as a function of potential control actions and then chooses the best control scenario based on some objective function that may combine a multitude of elements (desired performance, constraints, control efforts, …) (Stentoft et al., 2019). A model-based predictive controller thus consists of two essential components, namely a process model that predicts one or more state variables (e.g. NO_3-N effluent), and a controller with an optimizer, that allows to identify the best course of a manipulated variable (e.g. the dosing of methanol) over a prediction horizon using an objective function.

MPC is thus a multivariable strategy that encompasses multiple variables, such as constraints, handling of actuators, states and process outputs. It brings a structured approach to solutions where the main aim is to minimize a performance criterion in the future. The future behaviour is computed according to a model of the process being managed. It utilizes an internal dynamic model of the process, a history of past control moves, and an optimization cost function over the receding prediction horizon to calculate the optimum control needed. The first input in the optimal sequence is then transmitted to the physical WWTP for implementation, and the entire calculation is reiterated, and this for subsequent control intervals.

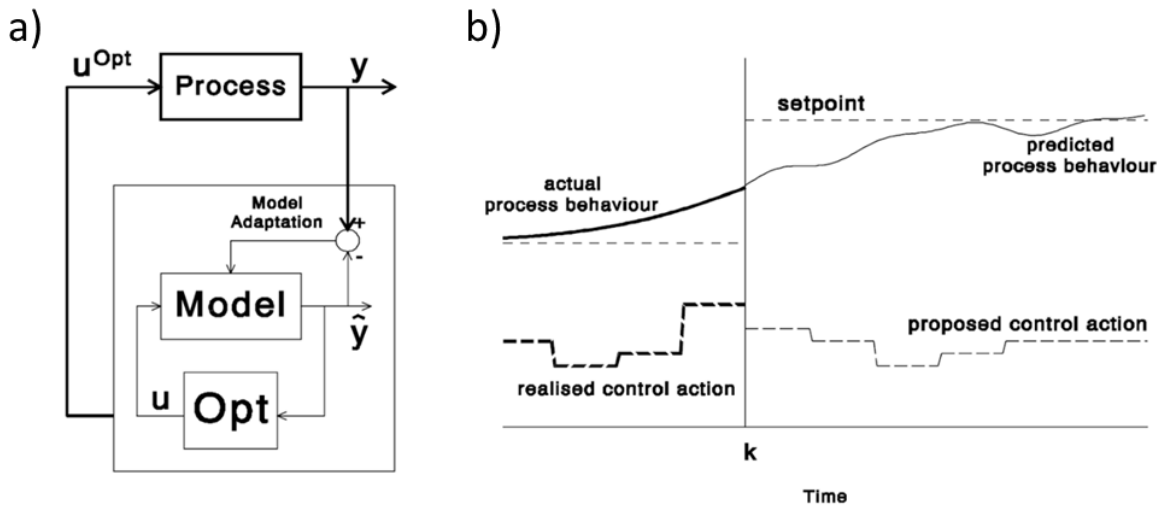


Figure 6-1: a) principle of model predictive control, b) Illustrative calculations performed by a MPC algorithm (Vanrolleghem, 1995).

As a process control technique, MPC is not (yet) widely applied in wastewater treatment due to complexity of biological processes that limit the predictive performance of mechanistic models. For MPC to function, it must address the non-linear system characteristics by applying non-linear models, which are computationally intensive to solve. Literature shows that the inclusion of data-driven models into the controller could be a good alterative (Newhart et al., 2019). In this approach, the mechanistic component is replaced by a machine learning model. However, HMPC's have not been widely applied in the wastewater industry so far (see also chapter 2). There are not many academic studies on the topic of HMPC, and most of them apply a different definition of hybrid models as defined in this study: using a mechanistic model supported by a data-driven model. The HMPC developed here is one of the first implementations of a hybrid MPC on an operational wastewater treatment facility. The only other known example is described by Sparks et al. (2022) using a similar approach as being developed here, but for the application of (i) chlorine for disinfection and (ii) aeration in an activated sludge system.

6.2 Materials and Methods

6.2.1 Case Study Site

The research project was carried out with operational data collected at the Seine aval (SAV) wastewater treatment plant situated in Achères (Yvelines), in the outskirts of Paris, which is managed by SIAAP, the public greater Paris sanitation authority. This plant treats an average volume of 1 500 000 m^3/d in dry weather (17.4 m^3/s; 6 000 000 P.E.).

As illustrated in Figure 6-2, the wastewater treatment consists of a primary physical-chemical treatment (screening, de-sanding, oil removal and enhanced primary settling with coagulant-flocculants addition), followed by a secondary biological treatment using biofiltration in a configuration of pre-denitrification (preDN), nitrification (NIT) and post-denitrification (postDN). External carbon is dosed in the predenitrification – if necessary due to low 'natural' carbon concentrations – and the postdenitrification stages, while the biofilters are aerated for nitrification. During the period studied, the average flow rate in dry weather conditions to the biofiltration lane was 1 090 000 m^3/d. More details can be found in chapter 2.

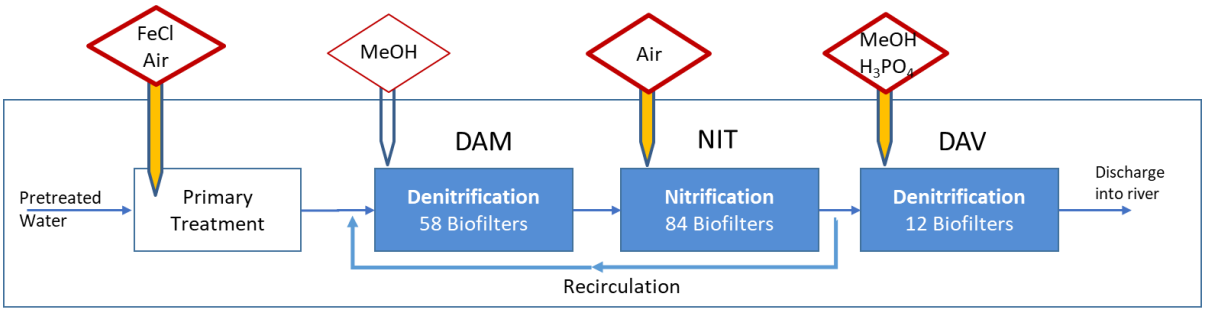


Figure 6-2: Schematic diagram of the wastewater treatment at the Seine aval wastewater treatment plant (SIAAP) with indication of chemical dosing and aeration.

The first stage of biofiltration (preDN) consists of 58 Biostyr® type biological filters (173 m^2 * 3.5 m height/biofilter - granular material: polystyrene 4.5 mm in diameter) in an anoxic environment for the elimination of nitrates (NO_3-N) and nitrites (NO_2-N) by heterotrophic bacteria using the organic matter (OM) present in the settled water. This configuration makes it necessary to recirculate nitrified water from the NIT outlet back to the preDN inlet. The second stage includes 84 Biostyr® type biological filters (173 m^2 * 3.5 m height/biofilter - granular material: polystyrene 4.0 mm in diameter). This stage is aerated by devices placed at the base of the filtering material that allow for the nitrification process (NIT) by autotrophic bacteria that transform ammonia (NH_4-N) into nitrates and nitrites. The final biofiltration stage (postDN) consists of 12 Biofor®-type biological filters (147 m^2 * 3 m height/biofilter – granular material: clay 4.5 mm in diameter) that treat nitrified effluent in a non-aerated environment. In this configuration, due to the low carbon content of the treated water, denitrification in the postDN stage requires the addition of an exogenous carbon source; here in the form of methanol CH_3OH (or MeOH). When the PO_4 concentration becomes too low and starts to limit bacterial growth, orthophosphoric acid (H_3PO_4) can be dosed at this point.

The average upwards flowrate through the Biostyr filters in the preDN is 10 m/h for an average NO_3-N reduction of 70%, while in NIT the average flow rate is 7.2 m/h that removes 90% of the NH_4-N. The average hydraulic flow rate of the Biofor filters in postDN is 19 m/h with an NO_3-N reduction of 85% over the period studied.

The biological filters are regularly backwashed according to a fixed time schedule and when the water pressure reaches a determined value by reversing the flow using treated water stored on site accompanied by the injection of large air bubbles. During washing, which lasts 30 minutes per day per filter, the trapped particles and part of the biomass growth are removed in order to restore the filtrating capacity.

6.2.2 Hybrid Model Architecture

The hybrid model concept presented here is identifiable as a cooperative parallel hybrid structure (Schneider et al., 2022) where the data-driven model is trained to predict the difference between the values of the mechanistic model simulations and observations (residual error); after which the two results are merged. This architecture is particularly suitable when working with a mechanistic model that is limited in the description of the process dynamics, for example due to incomplete knowledge or oversimplification of the processes (Schneider et al., 2022).

The key element for a well-developed hybrid model lies in the information content of these residuals; a well-trained machine learning model can pick-up 'hidden' dynamic information and use this to correct the prediction error of the mechanistic model (Lee et al., 2002).

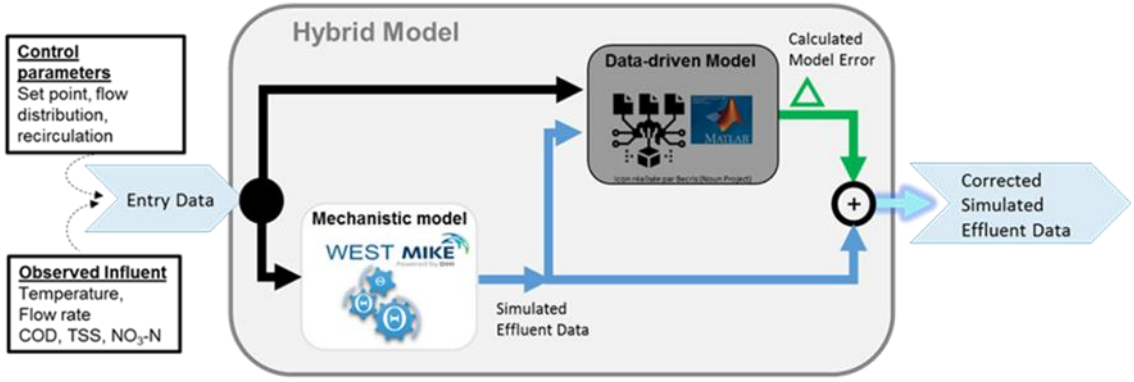


Figure 6-3: Schematic diagram of the parallel hybrid model structure (grey box) combining the predictions of a data-based model (black box) and a mechanistic model (white box) based on the process knowledge of the bio-physical-chemical system, in order to have corrected simulated effluent data.

The architecture of the developed hybrid model used in this study (Figure 6-3) is a parallel structure where the data-driven model is trained to predict the difference between the mechanistic model simulations and the observations (the residual error). The mechanistic model (MM) first makes predictions of the effluent quality variables, which are then passed to the data-driven model (DDM) together with data of the influent quality (black line), as well as information on the important treatment parameters. The DDM uses this information to estimate the residuals of the mechanistic model predictions. Finally, the two results are combined in order to obtain corrected values of the effluent variable. This is done for three effluent variables (NO_3-N, TSS, total COD) separately. For each effluent variable, a subset of process input variables is used. During the training phase of the data-driven model, measured effluent variables are used to calculate the difference between observed and simulated variables, which are used as labelled responses for the training algorithms of the DDM.

Mechanistic Model

The mechanistic model component of the hybrid model is designed as a 1-dimensional dynamic model of a submerged upwards flow biological filter where the biological activity of microorganisms inside a biofilm is calculated. The model was originally developed by Zhu (2020) and updated in this PhD study (Chapter 4). The model simulates the biological and physicochemical conversion of carbon, nitrogen and phosphorus by describing the evolution of substrates by a mass balance based on the ASM1 model (Henze et al., 2000). In total, 22 biological conversions are accounted for. The main modifications and additions focus on the mass transfer in biofilms, the kinetics of microorganism processes and the influence of backwashing cycles.

The dynamic model applied in the MPC controller is a simplified model in the sense that it simulates the behaviour of only one biofilter representing the typical behaviour assumed to occur. In the real system, the predenitrification stage consists of 58 biofilters divided over three batteries that are controlled independently from each other. Within each battery, the effluent of six or seven filters is combined, so that the effluent of a battery contains the mixed effluent of multiple groups of filters.

The simplified model used in the MPC controller thus simulates without the process control implemented in the system-wide biofiltration model of the three stages: predenitrification (preDN) – nitrification (NIT) – postdenitrification (postDN). Note, however, that the MPC model contains all same parameters, variables and equations as the system-wide model, however, it computes much faster, which is a crucial aspect required for MPC. Find a complete description of the mechanistic model in chapter 4.

The mechanistic biofilter model was calibrated on the first three and a half months of the dataset available from December 1, 2019. The model was then validated on the rest of the dataset from June to September 2020. The statistical scores describing the magnitude of the model prediction errors (residuals) are calculated to assess model performance and accuracy qualitatively and quantitatively. Each dynamic simulation was preceded by a 28-day steady-state simulation to initialize the model. The mechanistic modelling was performed in the WEST+ (2016) software (DHI) environment.

The methodology used for the calibration, including the global sensitivity analyses used for the estimation of influential model parameters, is well described in chapter 4.

Data-driven Model

The trained and tested data-driven model (DDM) developed in Chapter 5, consists of a three-layer feedforward ANN that uses a ReLU activation function. A five-fold cross validation was applied on the training set. The normalized variables that have been applied include the type of day (weekday or weekend), type of weather (dry or rainy), temperature, flow rates, observed influent concentrations of nitrates, TSS, total COD and soluble COD, as well as operational parameters such as the NO_3-N set point in the effluent and the methanol dosing rate. In addition, the mechanistic model effluent variables at the preDN outlet are included as features. During the training phase of the DDM, the error between the mechanistic model and the observations is also transmitted to allow for supervised learning. This was done for each output to cover the residuals of the NOx concentrations (NO_3-N + NO_2-N), total COD and total suspended solids. After training, the selected model was tested with a new dataset (1 300 points) covering the period from 1.3.2020 – 15.3.2020. The optimal structure was selected by comparing the performance of each model. The raw data processing, the development and training of the DDM were performed in the MATLAB R2022a (MathWorks) environment using the *Machine Learning* and *Deep Network Designer* toolkits.

Hybrid Model

The hybrid model, around which the HMPC is built, consists of a mechanistic model (MM) enhanced by a data-driven model (DDM). The MM simulates process behaviour and makes predictions of variables based on the process knowledge incorporated in the model's equations. The DDM is used in this hybrid modelling concept to correct the mechanistically simulated output variables in order to reduce the prediction error and thus increase the overall model performance inside the MPC structure. This then permits the optimizer component in the MPC to search for the optimal series of control actions based on modelled variables that have a higher level of precision. This is expected to result in a more efficient process control.

The methodology applied for the training and testing of the DDM and the configuration of the hybrid model are discussed in more detail in chapter 5.

6.2.3 Current Process Control of Methanol Dosing at Study Site

At the Seine aval WWTP, methanol is stored in pure form, but applied in diluted form (5%) – necessary for safety reasons.

According to the constructor's manual, the dosing of pure methanol was designed with an average daily value of 2660 L/d and a maximum of 18 460 L/d (18.5 m^3/d) for all filters active at the same time and is proportionally recalculated according to the number of active filters in each battery. Assuming that in each battery there is always one filter in backwashing session, and that an additional filter is offline due to maintenance, the maximum number of filters that can be considered active at a time is 55 filters for preDN stage.

The dosing takes place at the entry channel of the pumps that recirculate the nitrified water from the outlet of the NIT stage back to the entrance of the preDN stage. The maximum dosing ratio of methanol (g) to the nitrates (g) removed is 0.8 gMeOH/g NO_3-N denitrified, which corresponds to 1.25 gMeOH dosed per 1g NO_3-N removed. Operators can lower the dosing ratio if the nitrate concentration in the effluent is less than the set point, which is currently set at 1.3 mgNO_3-N/L.

The mechanistic biofilter model developed previously (Zhu, 2020) integrated methanol dosing control based on the current process control in the predenitrification stage at the Seine aval WWTP. The methanol dosing control in the model is a function of the following elements (see equation 6-1); (1) a measurement of the NO_3-N concentration at the inlet of the preDN stage that serves as a feedforward signal; (2) the desired set point of the NO_3-N at the outlet, (3) the ratio of methanol (g) dosed per gram of NO_3-N + NO_2-N removed, (4) the weather type indication, that can either be a heavy rain event (0) or a dry weather event (1).

IF Weather > 0:

$$Mass\_MeOH = \left(NO3_{Influent} - NO3_{setpoint}\right) * flow\ rate * (ratio\_MeOH/NOx\_removed)$$

ELSE: *Mass_MeOH* = 0

(Equation 6-1)

In equation 6-1, the Mass_MeOH is the flux of methanol dosed (g/d); the (NO_3_Influent) is the measured nitrate concentration (gN/m^3) in the inflow of the preDN, the flow rate of the wastewater (m^3/d) and the ratio of methanol per gram of nitrate to be removed (g/gN). The current settings for the NO_3-N set point and the ratio of methanol dosed per gram of nitrates removed are shown in Table 6-1.

The weather functions as an 'on' and 'off' function of methanol dosing control and simulates the operator's rules that during heavy rain events, the recirculation of nitrified water from the NIT stage to the inlet of the preDN is stopped, thus, removing the need for methanol dosing. A heavy rain event is classified by flow rates higher than 22 m^3/s (as an average value observed during a 15-minute interval) in combination with an observed conductivity measurement below 90 mS/m. The set point and the desired ratio are defined by the operator.

Table 6-1: Operational settings for the NOx set point and the applied ratio of methanol dosed per gram of NOx to be removed at the Seine aval WWTP.

	NOx SetPoint [mgN/L]	Ratio MeOH : NOx-N [g MeOH / g NOx-N]
MeOH_PreDN	1,3	0,75
MeOH_PostDN	2,0	3,0

The mass of methanol (MeOH) dosed is modelled in the biofilter model (see Figure 6-4) as soluble COD and added in the influent flow as the state variable S_S (readily biodegradable substrate) by applying a multiplication factor of 1.5 gCOD/gMeOH (Zhu, 2020). This factor is the quotient of the COD equivalent of methanol (48 g/mol) and the molecular weight of methanol (32 g/mol).

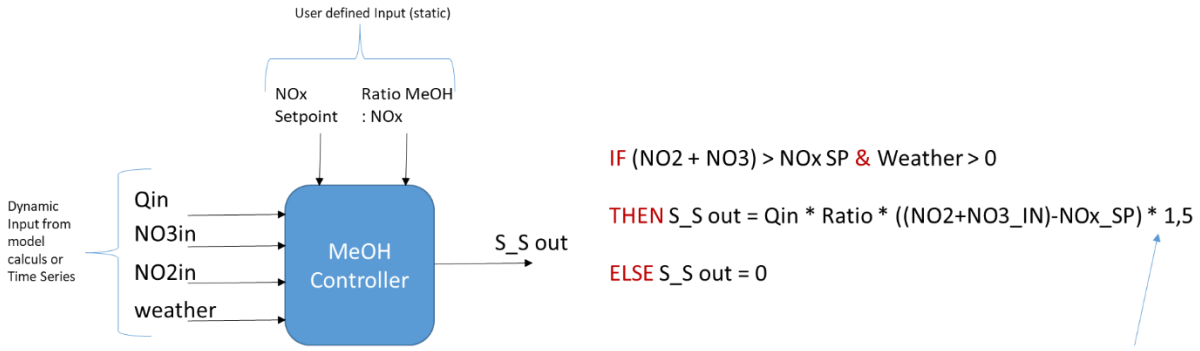


Figure 6-4: Control settings of methanol dosing in the original mechanistic biofilter model that are based on the current operational setting in the predenitrification stage of the Seine-aval WWTP.

6.3 Hybrid Model Predictive Controller

The concept of the hybrid model predictive control (HMPC) developed in this study for the dosing of methanol in a denitrification biofilter is illustrated in Figure 6-5. The HMPC functions by first searching for an optimal methanol dosing rate by running multiple scenarios through the hybrid biofilter model, which contains the calibrated mechanistic model (the MPC Model, blue box in the figure) simulating the behaviour of one filter as described earlier, and a correcting ANN-model (yellow box). It then compares the simulated nitrate effluent with the desired effluent NO_3-N concentration (set point). The MPC model receives the measured influent data and relevant information about the operational settings (e.g. number of active biofilters). This information could easily be extended to include cost for methanol dosing and environmental costs (e.g. GHG emissions) associated with the methanol dosing.

The dosing of methanol has a direct lowering effect on the nitrate concentration in the effluent. Higher doses of methanol will increase the metabolic activity of methanol-oxidizing microorganisms (methylotrophic denitrification), which will consume more nitrates and thus lower the nitrate concentration in the effluent. This assumes that sufficient methanol is added to enable fast grow-in of methylotrophs to establish a sufficiently high population. The optimizer searches for the smallest error between the simulated nitrate concentrations and the desired set point. The corresponding optimal dosing rate is then identified and transmitted to a second mechanistic biofilter model (the “Reality“ Process Model in the red box), which represents the reality to which this HMPC controller would eventually be connected. It is important to realize that this “Reality” process model is different from the MPC model in a way that approximates the difference that would occur between the MPC model

and the reality to which the HMPC is to be applied. That model represents the physical process taking place in the physical world and allows quantifying the impact of the proposed action.

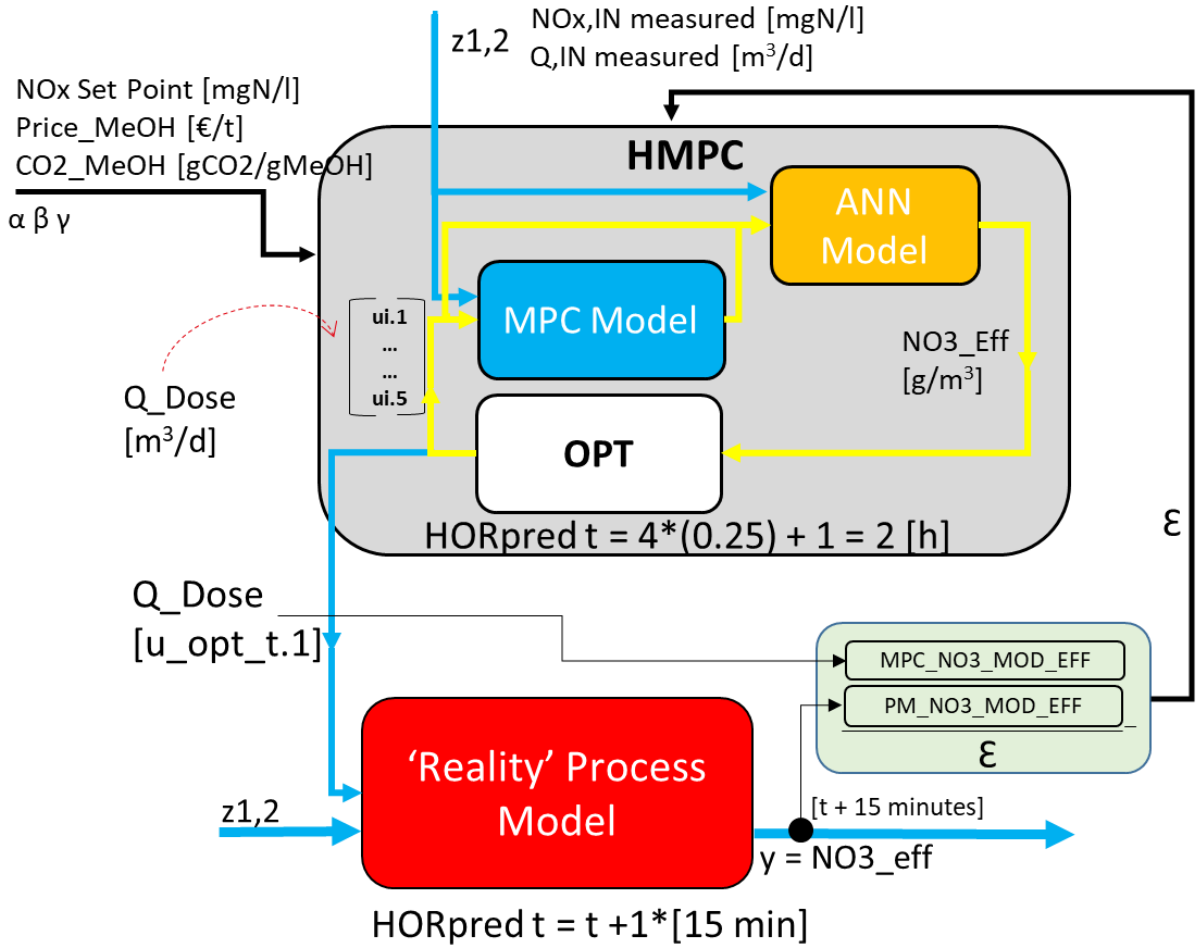


Figure 6-5: Conceptual scheme of the developed Hybrid Model Predictive Controller (grey box) that involves a mechanistic model (blue box) interacting with a correcting DDM (yellow box) to find the optimal control action value that is implemented in the "Reality" Process Model (red box). This model then makes a one-time-step ahead prediction of the effluent quality, which is used as the 'observed' effluent for the next iteration of the HMPC.

The influent files and the operational settings (e.g. information about the number of active filters in a treatment stage) are identical for both the MPC Model and the Reality Process Model. The initialization of both models is thus identical. However, due to certain model parameter differences (see section 6.3.3) the initial condition of each model in the dynamic simulation will be different, and so will their behaviour during the dynamic simulation.

When the predicted behaviour and that observed in reality (here represented by the 'Reality' Process Model) diverge from each other, it would be recommended to recalibrate the parameters of the mechanistic model used in the MPC configuration. However, in the case of wastewater treatment models, and especially for biofilm models, this would require an intensive and tedious procedure. A much faster method is to compensate for this model-mismatch by adjusting the set point of the MPC model. In fact, when the nitrate concentrations predicted by the MPC model differ from the real values (as simulated here by the "Reality" process model), the objective to reach a certain nitrate set point in the real world can be reached by aiming with the MPC for a set point that has been adjusted by this model mismatch. For instance, if the nitrate concentrations in the real world are 1.1 mgN/L and the

model in the MPC simulates a concentration of 1.5 mgN/L, the set point of 1.3 mgN/L would be increased to 1.7 mgN/L so that attaining that concentration in the MPC, would be expected to lead to attaining 1.3 mgN/L in reality.

This model mismatch correction through set point adjustment is implemented as follows: when the output of the Reality Process Model is compared to the predicted output by the MPC Model, the averaged difference between the two models of the last 15 minutes period (symbolized by Ɛ in Figure 6-5) is communicated back to MPC model, where it is used to modify the set point applied during the next iteration.

The objective function of the optimizer aims to minimize the error between the desired nitrate concentration, which is the set point, and the modelled nitrate effluent. It will seek to manipulate the dosing of external carbon into the biofilter simulated in the MPC model to meet the defined goal, despite disturbances occurring in the flow rate and influent concentrations. In the developed control structure, the process variable (PV) is the nitrate effluent concentration, and the manipulated variable (MV) is the flow rate of the methanol dosing (Q_Dose) expressed in m^3/d; the desired output is the set point nitrate concentration ($mgNO_3$-N/L) in the effluent (SP_NO3).

As can be seen in Figure 6-6, the prediction horizon of the MPC is set to 120 minutes, which gives ample time for the effect of a methanol dosing action on the biofilter to pass completely through the biofilter. The time step size (k) has been chosen as 15 minutes since the hydraulic retention time of the water in a biofilter is about 20 to 30 minutes. Any reagent dose injected at instant k will pass through the biofilter within 30 minutes, and its effect on the biological activity is assumed to be of the same order. The control horizon is set as the first four timesteps (thus from k to k+4), of which actually only the first proposed action is actually passed on to the real system (here the "Reality" process model). After the first iteration, the HMPC moves up one time step (k+1) in a receding prediction horizon, in order to start the optimization again. In this case study, this moving simulation window is repeated for a total period of 3 hours.

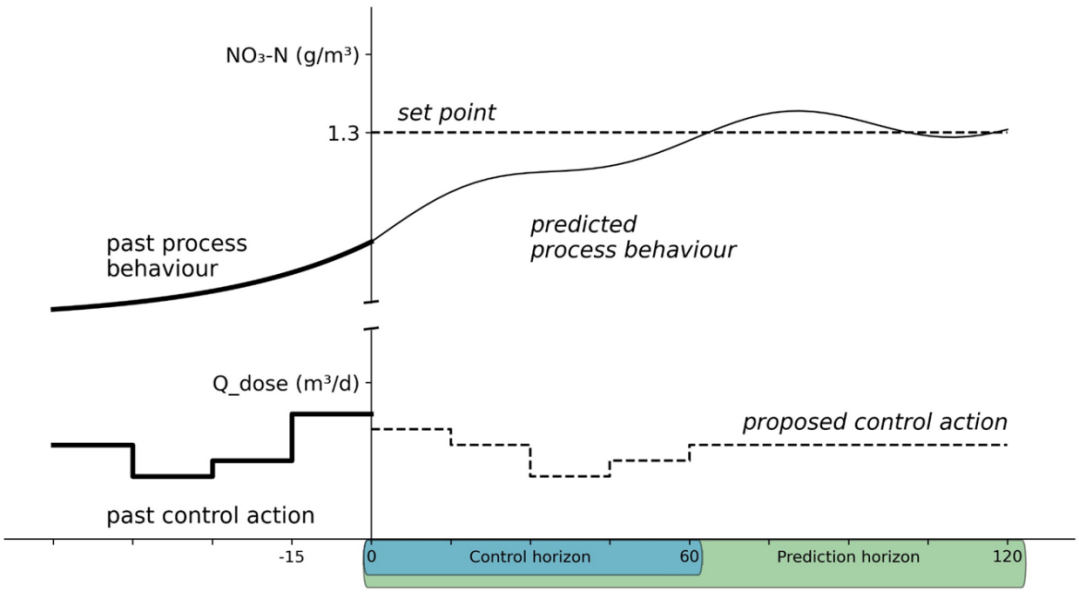


Figure 6-6: Applied (H)MPC control strategy showing the control horizon of 60 minutes and the prediction horizon of 120 minutes on the horizontal axis; as well as the desired NO_3-N set point value and the dosing flow rate (Q_Dose) on the vertical axis.

Besides the variable dosing flow rate (Q_Dose) and the desired set point, all other controller inputs are considered disturbances under 'ideal' conditions without taking in consideration measurement noise, time delays or other monitoring discrepancies. In a future expansion of this demonstration project, the Reality Process Model can easily be replaced by real-time measurements of nitrate effluent concentrations.

The HMPC controller is coded in the python (3.9) language in order to allow automatic execution and coupling of the DDM (developed in python) and the biofilter models, which were developed in WEST (DHI).

6.3.1 Reality Process Model Divergence from MPC model

In order to enforce a discrepancy between the process knowledge incorporated in the MPC model and the real physical system represented in Figure 6-5 by the red coloured Reality Process Model, it was decided to change some model parameters in the process model.

The selection of parameters to modify was based on the global sensitivity analyses performed earlier (see chapter 4) and the correlation of the parameter with the mass transfer of methanol into the biofilm, as well as the uptake of methanol by the active biomass. The global sensitivity analysis for the nitrate effluent variable identified the anoxic growth rate reduction factor of heterotrophic biomass consuming nitrates (n_h1) and the anoxic half-saturation constant of soluble COD of heterotrophic biomass consuming nitrates (K_S_NO3) as the third and seventh most influential parameters respectively (see Table 4-10). Next to two parameters affecting biokinetics, a parameter influencing the transport of the components in the biofilm was selected. The fraction of the biofilm that is removed during washing sessions (Washrate) was the most influential parameter on NO_3-N in the effluent of the mechanistic biofilter model.

When considering the sensitivity analysis of the parameters on the effluent soluble COD concentrations (see Table 4.11), the parameters n_h1 and K_S_NO3 have been identified as second and third most influential. The soluble COD is an important indicator that is directly correlated to the uptake of methanol by the microorganisms, since the dosed methanol in the model is completely 'assigned' to the state variable S_S, which is the readily biodegradable substrate that forms a large part of the soluble COD.

Regarding the influence of model parameters on the effluent TSS concentrations (see Table 4.12), the WashRate was pointed out as having the fourth strongest influence; recognizing the impact of the daily backwashing sessions on the filtration efficiency of a filter capturing suspended particulates.

The parameters finally selected to be different in the Reality Process Model are indicated in Table 6-2. The direction of the changes in parameter values was chosen such that effluent predictions of nitrate were 10% lower, in the sense of root mean square error, which measures the average difference between the values of the original MPC model ('the reference') and those predicted by the model with the modified parameters. The order of the modification of the parameters was: (1) change parameter n_h1 until a -10% difference in NO_3-N effluent is noticed; (2) change K_S_NO3 for a +5% difference in NO_3-N effluent, (3) change WashRate for a -5% difference in NO_3-N effluent. This made that the total change in NO_3-N effluent were -10% compared to the Reality Process Model predictions before the parameter modifications. This order of change will allow an effect of the different scenario to be more

easily visible, but not to change the rate of kinetics and transport of compounds in the model in extreme manners.

The evaluation pertained to time series of 12 hours, since this is the total duration studied here.

Table 6-2: Parameters of the Reality Process Model that were modified and their impact on the simulated effluent nitrate concentrations.

Order	Parameter	Description	Calibrated value	Modified value	Parameter change	% Impact on Δ NO_3-N effluent
1	**n_h1**	Anoxic growth rate reduction factor of heterotrophic biomass consuming nitrates	0.28 [-]	0.26 [-]	-7%	-10 %
2	**K_S_NO3**	Anoxic half-saturation constant of soluble COD of heterotrophic biomass consuming nitrates	41.37 [gCOD/m^3]	39.30 [gCOD/m^3]	-5 %	+5 %
3	**Washrate**	Biofilm fraction that is removed during washing	0.55 [-]	0.56 [-]	+2 %	-5 %

The effect on the modelled nitrate concentration of the final selected parameter values is shown in Figure 6-7.

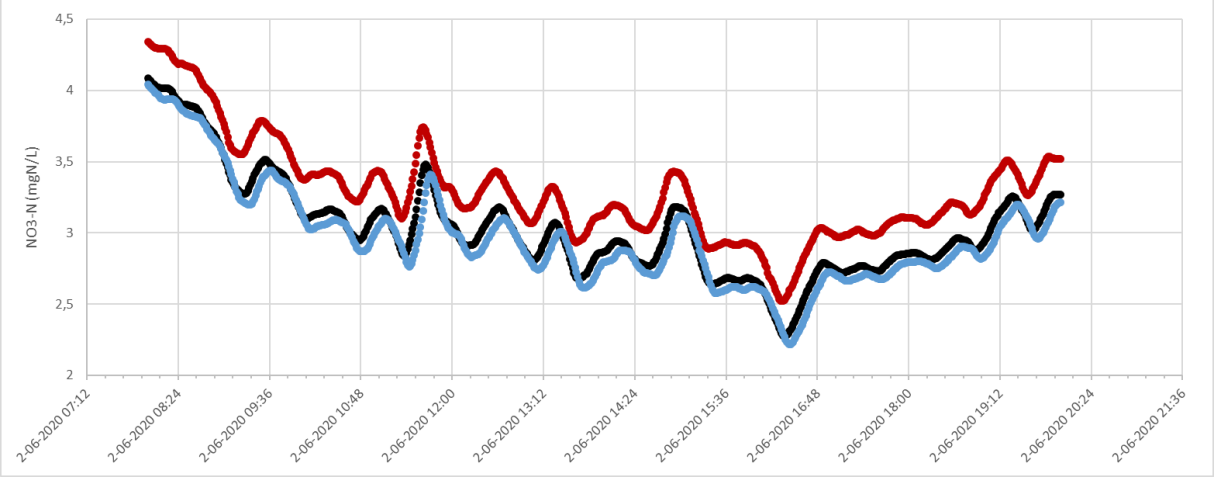


Figure 6-7: Evaluation of changes in selected model parameters (n_h1, K_S_NO3, WashRate) upon the modelled NO_3-N effluent under two scenarios: a positive 10% change (red line) and a negative 2% change (blue line) compared to the reference situation (black line).

6.3.2 HMPC Model Divergence

Another situation that will be studied in this chapter occurs when the mechanistic model that is part of the hybrid model is recalibrated so that it better corresponds to reality. In order to simulate such recalibration situation, it was decided to modify the same model parameters identified for the modifications in the Process Model. However, this time the amplitude and direction of the change was maximum +2%.

The order of the modification of the parameters indicated in Table 6-3 was the following: (1) modify n_h1 until a +2% difference in NO_3-N effluent is established; (2) change K_S_NO3 for a -1% difference

in NO_3-N effluent, (3) change WashRate for a +1% difference in NO_3-N effluent. This made that the total change in NO_3-N effluent was +2% compared to the MPC Model predictions before the parameter modifications.

Table 6-3: Parameters of the MPC Model that were modified and their impact on the simulated effluent nitrate concentrations.

Order	Parameter	Description	Calibrated value	Modified value	Parameter Change	% Impact on ΔNO_3-N effluent
1	**n_h1**	Anoxic growth rate reduction factor of heterotrophic biomass consuming nitrates	0.28 [-]	0.285 [-]	+2 %	+2 %
2	**K_S_NO3**	Anoxic half-saturation constant of soluble COD of heterotrophic biomass consuming nitrates	41.37 [gCOD/m^3]	42.41 [gCOD/m^3]	+2.5 %	-1 %
3	**Washrate**	Biofilm fraction that is removed during washing	0.55 [-]	0.5445 [-]	-1 %	+1 %

The result of the final selected parameter values is shown in Figure 6-7, where a slight shift of the time series is observed due to rounding of time stamps (all expressed in unit of days) by the model's integrator.

6.3.3 Description of Data Sets

The datasets used for the development of the HMPC controller were taken from the operational SCADA system for the June 1-8, 2020. It was the first week during which the methanol dosing came back online with certain regularity (see Figure 6-8). The dosing had been halted before that due to technical problems.

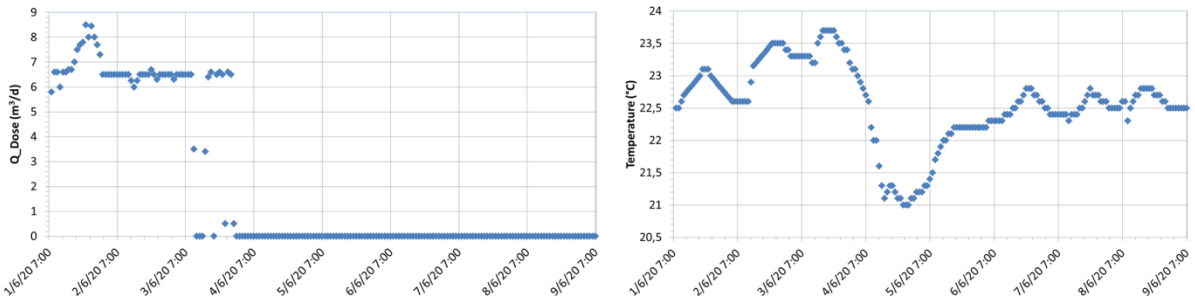


Figure 6-8: a) methanol dosing rate [m^3/d] and b) temperature profile of the settled wastewater for June 1-9, 2020.

Within these 8 days, a window of 12 hours was selected to simulate the MPC controller. The 7-day average of the influent conditions was extrapolated to cover 28 days to be used for the Steady-State simulations.

6.3.4 Hybrid Model Predictive Controller Performance Criteria

The performance of the hybrid model predictive controller was assessed by applying the Root Mean Square Error (equation 6.2). The objective function of the optimizer expresses that one tries to minimize the RMSE between the NO_3-N set point and the predicted NO_3-N values over the prediction horizon.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \qquad Equation\ 6.2$$

6.3.5 MPC performance evaluation procedure

In order to be able to evaluate the performance of the HMPC and compare it to the other types of process control, a scenario analysis has been performed consisting of the following high-level scenarios (see Figure 6-9):

Scenario 1: Reference of current control. The biofilter Process Model runs with the current dosing control that is based on feedforward rules of nitrate measurements at the entrance of the preDN stage (equation 6.1), thus without a model predictive controller involved. The simulated nitrate effluent and the dosed methanol (g) will serve as the reference situation.

Scenario 2: Reference situation for the MPC controller. The MPC controller applies a biofilter model (blue model) to find the optimal dosing rate. However, in this scenario, the MPC model has been modified to have exactly the same parameter values as compared the Process Model, so that the MPC model is completely aligned with 'reality'. This simulates an ideal situation.

Scenario 3: MPC control with calibrated MPC model. The MPC model is here the calibrated biofilter model without the data-based model that corrects for its deficiencies, which thus differs to the Reality Process Model. The model parameters are not in harmony.

Scenario 4: HMPC control with hybrid model. The predictive controller in this scenario uses the trained DDM to correct for the mechanistic model's error. The MPC's optimizer algorithm that aims to find the optimal dosing rate then uses these corrected simulated effluent data.

Scenario 5: HMPC control with modified biofilter model. In this scenario, the mechanistic biofilter model is modified in the sense that the model parameters are recalibrated to recover from a large divergence between the current calibrated model and the physical reality. The maximum allowable effect of the modifications on the NO_3-N effluent concentrations are + 2%.

These five high-level scenarios are illustrated in Figure 6-9, where the MPC model with the calibrated parameters is indicated by the blue coloured boxes, while the Reality model simulating the physical system is coloured red. The light-blue coloured box of the fifth scenario is to indicate that the model parameters have been modified compared to the dark blue coloured models. Also shown are the expected outputs on the nitrate effluent concentrations for each scenario and an assessment of the impact on the costs of methanol dosing.

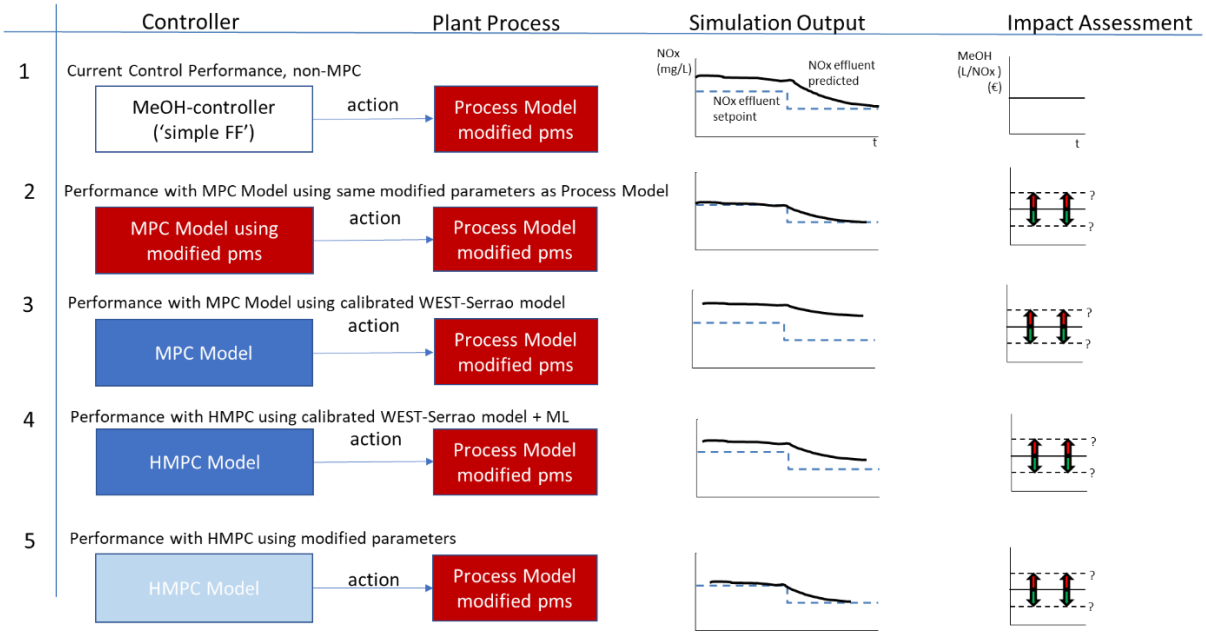


Figure 6-9: Definition of the high-level scenarios performed to compare the effect of using the MPC and HMPC under varying conditions.

6.3.6 Optimization Scenarios Workflow in Python

MPC optimization searches for the lowest error between the simulated effluent of the mechanistic model and the desired effluent concentration, which is the nitrate set point in the effluent of the preDN treatment stage. The MPC optimizer and the interactions with the MPC Model and the Reality Process Model have been written in python code. This permits an automated execution of the various optimization scenarios and executing the data-driven model and the two mechanistic models, as well as facilitating data exchange between the models. Flow paths were designed to run the controller in a MPC mode and an HMPC mode. Table 6-4 describes the sequence of steps of the two workflows that were written in modular python scripts. This approach allowed isolating each step of the workflow making it easier to debug and progressively put in additional features during development.

Table 6-4: Sequence of steps implemented for the MPC and HMPW workflows. The steps refer to the python scripts shown in the figures 6.10 and 6.11.

Step	Description	MPC	HMPC
1.py	Create executable xml files for mechanistic MPC model based on optimizer values and start & stop time of next iteration	V	V
2.py	Run mechanistic MPC model for 120 minutes and save simulation output in txt files for each optimization scenario	V	V
2b.py	Run data-driven model to correct the mechanistic MPC model's residual error of each optimization scenario; save simulation output in txt files	-	V
3.py	Evaluate the performance of each optimization scenario by comparing to the set point value; select best scenario with lowest RMSE	V	V
3b.py	Re-execute the identified best scenario in the (H)MPC model and reinitialize the mechanistic model	V	V
4.py	Create executable xml file for the Reality Process Model based on best optimization scenario and start & stop time of next iteration	V	V
5.py	Run the Reality Process Model for 15 minutes, save simulation output in txt file and reinitialize the model	V	V

6.py	Evaluate performance by comparing NO_3-N effluent to set point in terms of RMSE and update set point for MPC model in case of model mismatch	V	V
7.py	Count the number of performed iterations	V	V

MPC Workflow

The workflow created by the different python scripts for high-level scenarios 2 and 3 is shown in Figure 6-10. It consists of seven scripts that are executed in order. The optimization scenarios are a selection of predefined scenarios containing methanol dosing flow rate values (Q_Dose) in the range between 0 and 18 m^3/d, corresponding to options between no dosing, low, intermediate and high methanol dosing flow rates. The array contains the following values: 0, 4, 8, 12, 16. The MPC optimizer will run each of these scenarios and chooses the scenario that results in NO_3-N effluent values that are as close as possible to the desired set point, based on the lowest SSE error calculated over the prediction horizon.

The first script ('1.py') loads the different optimizing runs or low-level scenarios stored in a csv file and creates for each scenario, an executable file (.xml) for data exchange between systems. It reads the influent data for eight time steps with 15 minutes intervals (corresponding to 2 h). Before every iteration, it moves up one time step of 15 minutes in order to keep synchronized with the receding prediction horizon of the Reality Process Model.

Script '2.py' executes the xml files in the WEST mechanistic model that forms part of the MPC and creates output text files with the simulated NO_3-N concentrations.

The third script ('3.py') then makes a comparison between the simulated NO_3-N concentrations and its set points. It reads the set point values from an array stored in a txt file. This script allows calculating the Root Mean Squared Errors (RMSE) in order to select the best dosing flow rate (Q_dose) time series from the optimization scenarios that corresponds to the lowest SSE score. The identified best-fit scenario is then rerun for a second time through the mechanistic model ('3b.py'). In this way, the MPC model is properly reinitialized with the correct scenario values for the next iteration.

Script 4.py prepares the input xml file for executing the mechanistic Reality Process Model in WEST for the next 15 minutes. It takes the identified optimal Q_Dose of the first time step, as well as the new start-time and stop-time corresponding to the iteration number updated with 15 minutes time intervals.

The execution of the Reality Process Model takes place by script 5.py, which runs the model in dynamic mode for 15 minutes and creates an output file that contains the effluent NO_3-N values at the end of that simulation (t + 15 minutes). At the end of the script, it reinitializes the model by saving the end values of the state variables as initial values for the next iteration.

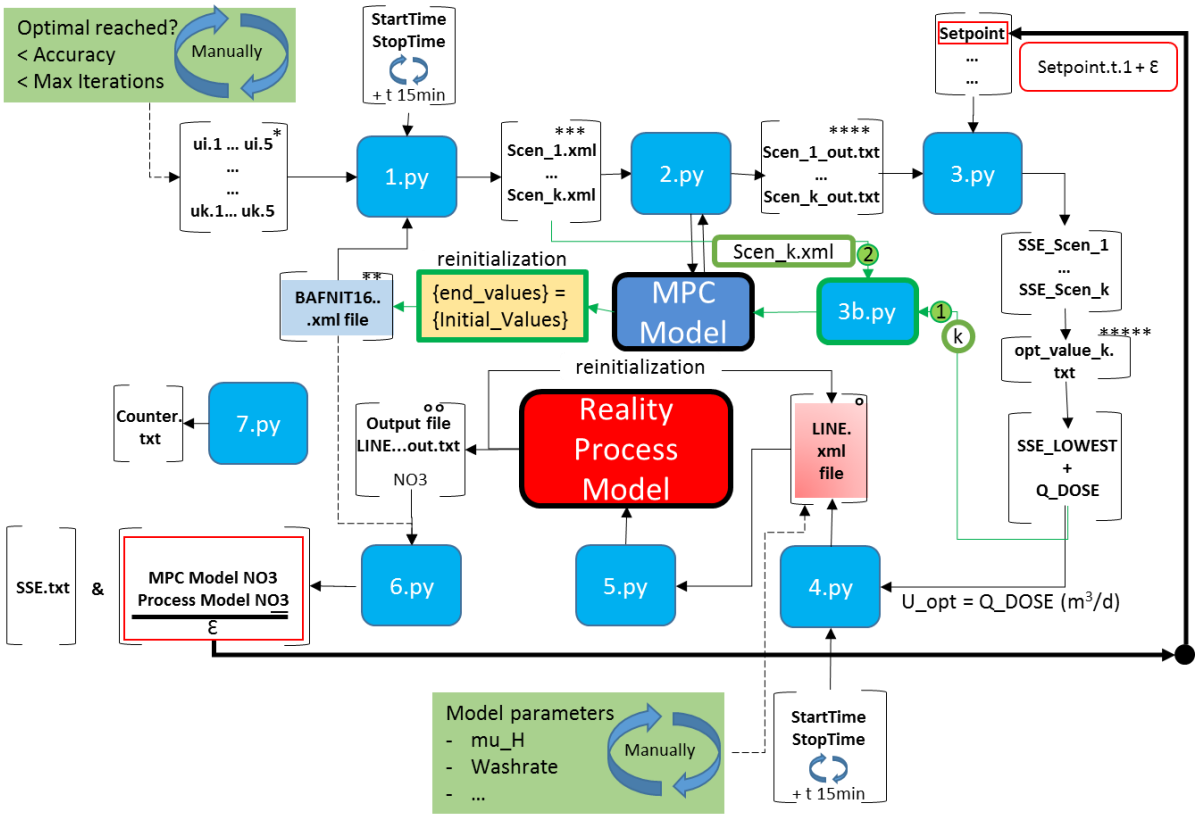


Figure 6-10: Workflow of python scripts as written for the execution of Scenario 2 and 3 with the Model Predictive Controller.

Feedback about the evolution of the nitrate effluent concentrations in the effluent of the Reality Process Model is now provided for in script 6.py by calculating the difference between the MPC Model prediction of NO_3-N effluent and what is happening in the physical system (the "Reality" Process Model) for the entire 15 minutes. This difference Ɛ ('epsilon') is added to the set point array, which is then stored in the updated txt file. During the next iteration, this txt file is read by script 3.py to collect the modified set point values. In addition, script 6.py also calculates the error between the NO_3-N set point and the predicted nitrate concentrations in the preDN effluent.

Finally, script '7.py' serves to indicate the current number of iterations using an external or global variable that allows keeping count of the iterations that have been completed. This global variable is used in the different scripts to account for the receding prediction horizon that moves forward with one time step interval of 15 minutes after each iteration.

Also indicated in the green boxes in the figure are the possible manual modifications to the optimizer scenario and the model parameters of the mechanistic models. For the moment, the optimizer scenarios are predefined based on process knowledge, but this could be automated in a future extension of the project. The same applies to the values of the parameters for both models used here that are currently manually modified, but this could be performed based on optimization results.

HMPC Workflow

The workflow describing the different python scripts for the utilization of the Hybrid Model in the context of the MPC, scenarios 4 and 5, is shown in Figure 6-11. On many accounts, it is similar to the flow of the scripts described for the MPC development in scenarios 2 and 3, except for the insertion of

the Data-Driven Model after script 2.py and before script 3.py. Here, script 2b.py executes the trained neural network to predict the MPC model's error and allow correcting for it. For this, the ML model uses the prediction variables, which are read from the influent and input files of the MPC model and includes operational parameters known to the MPC model. The updated corrected simulated variables are then used by script 3 to analyse performance of each scenario by calculating the SSE. From this point on, the scripts are identical and follow the path described earlier.

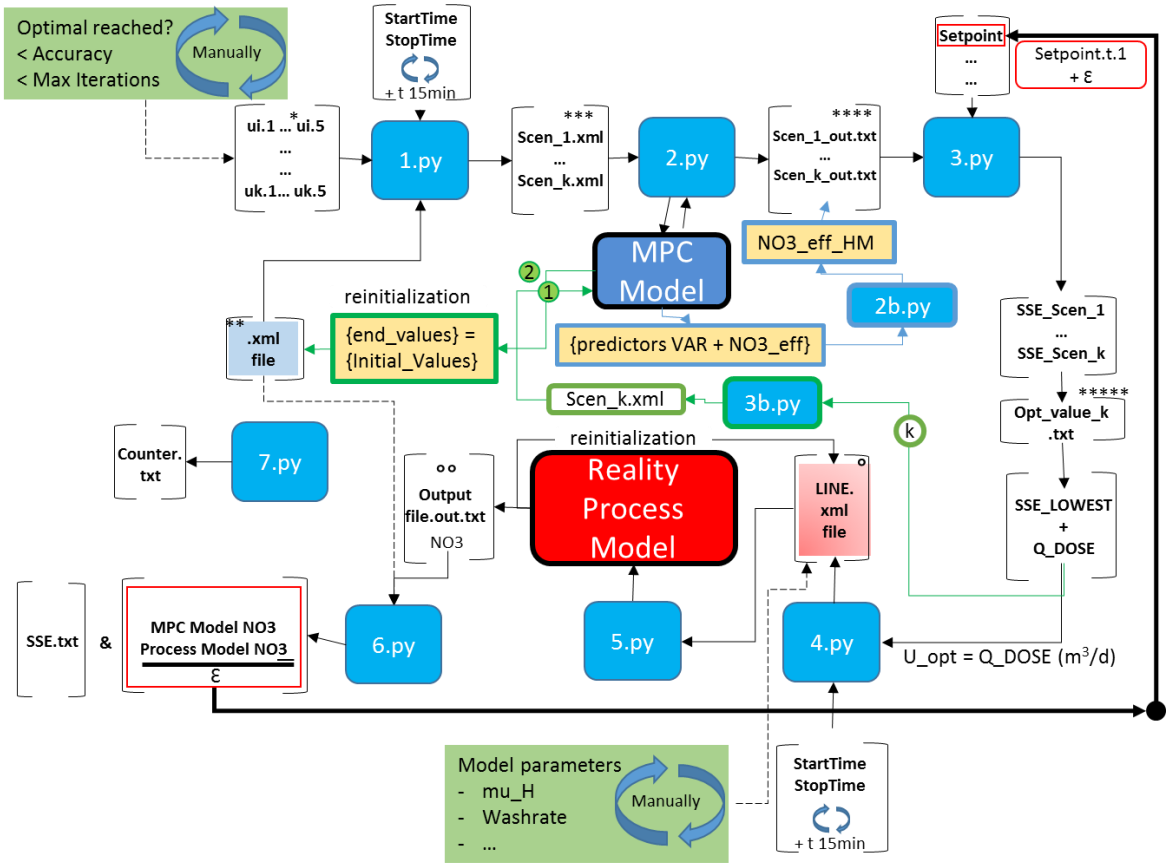


Figure 6-11: Workflow of python scripts as written for the execution of Scenario 4 and 5 with the Hybrid Model Predictive Controller.

For easy reference, the Python codes for the scripts are made available in Annex IV.

6.4 Results & Discussion

The simulations performed with the High-Level scenarios as presented in Figure 6-9 test different structures of a model predictive controller (i.e.: with and without support of a correcting machine learning model) under different conditions (i.e.: the modifications of the selected parameters n_h1, K_S_NO3 and WashRate). All simulations used the same input data for the period of 1 to 8 June, a period characterized by high NO_3-N influent concentrations and actual methanol dosing in the physical system.

Figure 6-12 shows the performance of each scenario by comparing the simulated NO_3-N concentration in the Reality Process Model with the nitrate set point. The performance of the current dosing control settings is the reference with which this work is to be compared with. Clearly visible is the significant

improvement of the MPC and HMPC controllers when compared to the current applied control (Table 6-5). The impact of the MPC and HMPC strategies is a reduction of the nitrate concentration, resulting in an improved effluent quality.

Table 6-5: Overview of the RMSE errors in the Reality Process Model for the total of 11 iterations performed in each scenario. The RMSE is defined as the difference between the NO_3-N set point and the simulated values for a period of nearly 3 hours

No.	Scenario	RMSE mgNO_3-N/L
1	Reality_Current_Control	1,77
2	Reality_MPC_A	0,89
3	Reality_MPC_B	0,94
4	Reality_MPC_C	1,01
5	Reality_HMPC_B	1,20
6	Reality_HMPC_C	1,24

In more detail, it is noticeable that the MPC_A scenario has the lowest error to the desired effluent concentration. This is expected since in this scenario the model parameters of both the MPC model and the Reality Process Model are identical. Here, the MPC model is able to closely predict the behaviour occurring in the 'real world'. When there is a mismatch between the assumed behaviour and the 'real' behaviour, however, here simulated by the other scenarios ending with B and C where selected model parameters in the MPC model are diverging from the "Reality" ones, the error gets larger. The overall direction of the parameter modifications seems to make a difference on the error as well. Scenario MPC_B, where modifications led to a 10% increase in nitrate effluent concentrations (see section 6.3.1), performed better than scenario MPC_C with a -2% change (section 6.3.2). This could be related to the efficiency of methanol, which is higher at higher NO_3-N values in the wastewater.

Between scenarios B and C, it is observed that the hybrid variants HMPC_B and HMPC_C, which apply the data-driven model to correct for the mechanistic MPC model's error, results in the highest RMSE values. This could indicate that the lack of the ML model's training with operational data that include methanol dosing is detrimental to their performance in a control context. In fact, as described in chapter 5, the DDM was trained on observed data from a period without external carbon dosing. It could benefit from a retraining on prolonged period of extensive carbon addition in the denitrification treatment stage.

Figure 6-12 indicates the overall performance of the "Reality" Process Model under the different scenarios in terms of RMSE between the desired NO_3-N set point and the simulated effluent NO_3-N concentrations. The figure concatenates the 15-minute simulations of a total of 11 iterations, each simulating a receding 15-minute prediction horizon in the "Reality" Process Model.

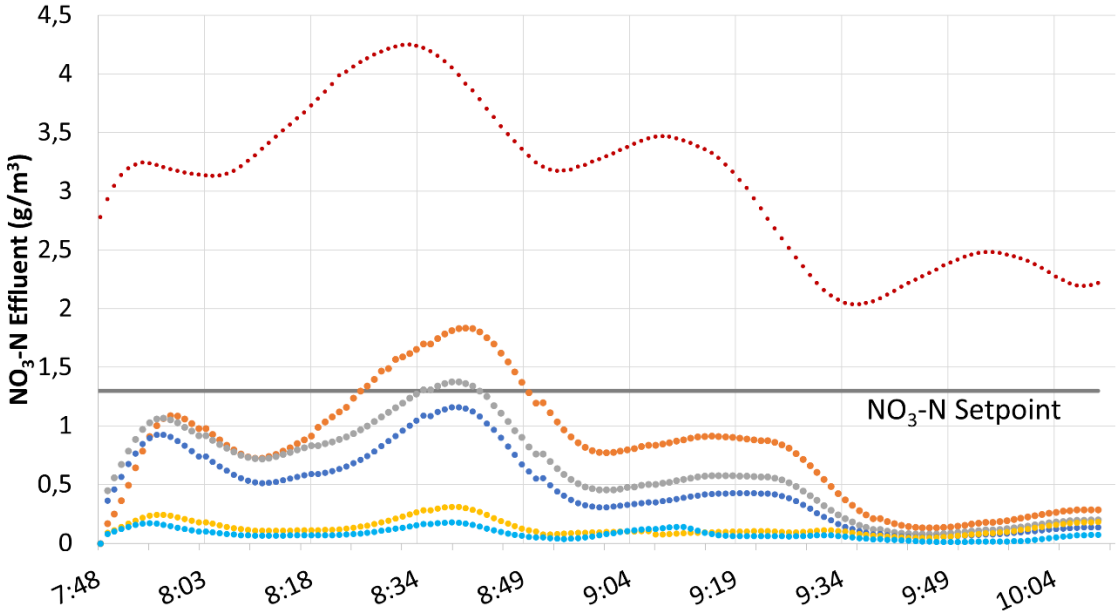


Figure 6-12: Simulated NO_3-N effluent concentrations in the Reality Process Model of the various High-Level scenarios using MPC and HMPC variants, as well as the effect of the current dosing control settings. Indicated are the current controller result (red colour), MPC_A (orange), MPC_B (grey), MPC_C (dark blue), HMPC_B (yellow) and HMPC_C (light blue).

Clearly noticeable in the figure is that the current feedforward control strategy (red line) results in a too high NO_3-N concentration with an average of around 3.1 mgN/L for the period studied, corresponding to an RMSE of near 1.8 mgN/L when compared to the desired set point of 1.3 mgN/L. The MPC control strategies result in much lower effluent concentrations that are converging to the desired set point during the first hour from 07:45 to 08:45. The average values of the MPC and HMPC variants are three times smaller than those of the current controller.

Figure 6-12 also shows that the MPC and HMPC control strategies result in nitrate values that are actually too far below the desired set point; thereby raising the RMSE. To cause such low nitrate values in the effluent, the MPC and in particular, the HMPC strategies dose large volumes of methanol. This results in higher operational costs of methanol usage.

As mentioned earlier in this paragraph, a retraining of the DDM with more data stemming from a period with active methanol dosing could certainly benefit to reduce the difference in outcome of scenarios B (MPC) and C (HMPC). However, further review of the results shown in Figure 6-12 suggests that the DDM may have been trained on the 'wrong' residual error. Instead of defining the residual error as the difference between measured and simulated nitrate effluent measurements in the preDN treatment system, for this demonstration the DDM should in fact have been trained on the difference between the Reality Process Model simulation results, which represent the physical system, and the simulation results of the MPC Model, which is the simplified process model included in the MPC. Due to the different characteristics of both models, the modelled effluent nitrate concentrations of both models are not identical. By training a DDM on this difference, it is expected that the resulting HMPC will be able to correct for the residuals, thus giving the lowest RMSE and thus the best performance with the HMPC that takes advantage of this best model. This demonstration project will be finalized by properly training the hybrid model and results will be presented at the IWA Watermatex2023

conference (Québec, QC, Canada, September 2023) and afterwards submitted for publication in the form of a scientific paper.

Note that, in the current optimizer structure, the optimizer only seeks to minimize the RMSE error between the set point and the modelled effluent nitrate concentrations and there is thus no additional incentive to limit methanol dosing. In a continuation of this study, reducing unnecessary dosing would be an aspect that definitely requires more attention by designing a multi-objective optimization problem.

6.5 Conclusions & Perspectives

A previously developed cooperative parallel Hybrid Model was applied in a Model Predictive Control strategy that searches for the optimal proposed control action based on predictions made by the hybrid model. The process control manipulates methanol dosing as an external carbon source to enhance the performance of denitrification in a submerged upward flow biofilter. The HMPC developed here is one of the first implementations of a hybrid MPC to a real, large water resource recovery facility treating 1.5 million m^3/day. The only other known example is described by Sparks et al. (2022) using a similar approach as being developed here, but for the application of (i) chlorine for disinfection and (ii) aeration in an activated sludge system.

This study shows that the realistically simulated Hybrid Model Predictive Controller significantly reduces the effluent NO_3-N concentrations and thereby improves the effluent quality. The computational demand of the hybrid model approach using a python script calling WEST is much smaller when compared to running the mechanistic model simulations within WEST's graphical user interface. This has been achieved by developing Python scripts that execute the mechanistic model and the machine learning models automatically and without the overhead of a graphical user interface.

Further studies of the control strategies using the Hybrid Model are required to understand the low simulated nitrate values. The HMPC optimizer can certainly benefit from regular retraining of the neural network. This would further increase the confidence in the model's performance that allows applying it in an active process control setting.

7 Conclusions and Perspectives

The research work of this thesis centres on wastewater treatment process modelling; on ways how to improve the model performance and make them applicable for advanced (near) real-time control of treatment systems. This chapter discusses the main conclusions of the work on mechanistic modelling of nitrogen removal by biofiltration, of the development of a hybrid model to compensate for the mechanistic model's residual error, and of the use of this hybrid model in a hybrid model predictive control application developed for the dosing of methanol in a denitrification filter. All this was developed at the super-scale case study of Seine aval in Paris, France.

The overall aim of the PhD study was to contribute to the integration of hybrid models that combine mechanistic and data-driven models as a tool for operators and managers in the wastewater industry to aid in decision-making, design and planning, as well as a process control. Wastewater treatment nowadays has become a complex chain of systems and processes that require effective control and pro-active decision-making based on data and information derived from the current state of a process, and from data-based analyses and models predicting future states. The mining of data will lead to more information and the mechanistic modelling of processes gains a deeper understanding of the various biological, physical and chemical processes at hand in a bioreactor. It is expected that this increased knowledge will result in more intelligent actions taken (e.g. Corominas et al., 2018; Therrien et al., 2020) that allow the wastewater industry to meet the wide range of demands and regulations, such as compliance with stringent effluent qualities, the recovery of resources, the reduction of energy demand and greenhouse gas emissions.

In this regard, the comprehensive objective of this study is to demonstrate that hybrid modelling brings added value to model performance of effluent quality simulations when used for complex process control of wastewater treatment applying high-resolution operational data. This is achieved by reaching the following principal objectives:

- Objective 1: A mechanistic model of a biofilm reactor for predicting nitrogen removal, power consumption, filtration and greenhouse gas (GHG) emission.
- Objective 2: A machine learning (ML) model to predict the residual error for the main effluent quality variables providing a better accuracy of the hybrid model.
- Objective 3: A hybrid model predictive controller (HMPC) that demonstrates the potential application of hybrid models in an operational setting.

The first objective was pursued in chapter 4, the second objective in chapter 5 and the third objective in chapter 6. The general conclusions related to these works will now be discussed in more depth, starting with those of the literature review.

7.1 Conclusions

7.1.1 Literature Review

The study of articles and textbooks on biofilter modelling and the application of data-driven models in the wastewater sector showed that the mechanistic models of conventional activated sludge systems are widely accepted in academia and industry; more than 35 years since the development of the ASM1 model has led to a large number of scientific papers written and modelling studies performed; as well as protocols that aid the modelling community. Such well-developed mechanistic models can be

defined as digital twins in certain contexts as they allow analysis of data and monitoring of systems to avoid problems from occurring, plan for future upgrades and (re-)designs, and optimization of processes. However, this is not the case (yet) for biofiltration, which is a biological wastewater treatment technique that contains large surfaces for bacteria to grow on so they treat pollutants and nutrients in the ecological conditions that are created for them inside a reactor. These microorganisms are prominently active inside a thin layer with a typical thickness of approximately 100 µm and obtain their substrate from the water through diffusion. Compared to the ASMs, the mechanistic models describing biofiltration require adaptation in order to model the behaviour of biomass inside the biofilm and the transport of compounds by diffusion. Biofilm models are therefore considered to be more complex and tedious to calibrate and validate; and moreover, lack widely accepted protocols for model development. Additionally, these biofilm models require more computing power, which makes them less suitable for real-time process control applications with a short time action horizon (Morgenroth, 2008b).

The rapidly increasing interest of the wastewater sector for data-based and data-driven methods that can smartly use increasing volumes of big-data being collected in modern wastewater treatment plants, has resulted in a proliferation of scientific papers on the topics of artificial intelligence, data mining and machine learning within the water and wastewater sector since the year 2000 (e.g. Hadjimichael et al., 2016). Data-driven models that apply machine learning models, such as ANN's and SVM's, have emerged as alternative approaches for process modelling. More recently, hybrid models that combine data-driven models with the mechanistic models are reported to have the potential of achieving more accurate model performance (e.g., Schneider et al., 2022).

Nevertheless, so far, the development of hybrid models for biofiltration systems has only found limited application in the industry. The main reasons are the reluctance to accept data-driven models since these lack a knowledge-based understanding of the underlying processes; the lack of data-management experience and training of operators and managers alike; as well as the long lifetime and high investment costs of infrastructure that needs to withstand harsh and very dynamic conditions.

The research work performed here has shown that it is indeed possible to develop a hybrid model that simulates a complex biofiltration treatment process with more accuracy than a mechanistic model alone can achieve using high-resolution (15-minute time steps) operational data from a very large wastewater treatment plant (dry weather average flow rate 1.5 million m^3/d; 6 million P.E). This study then demonstrated the application of a hybrid model in a comparative case study in which a conventional feed-forward controller in use today at the plant under study is compared to a Hybrid Model Predictive Controller of the methanol dosing to a biofilm reactor that takes advantage of the developed hybrid model. This, with the definition of hybrid model used here, has not been reported in literature before.

7.1.2 Calibration-Validation of a Mechanistic Biofiltration Model

The first objective of the PhD thesis focused on the calibration and validation of a mechanistic model describing an entire secondary biological treatment facility for nitrogen and organic matter removal by biofiltration consisting of predenitrification, nitrification and postdentrification stages. For the calibration, high-frequent data at 15 minutes intervals collected over 3.5 months were used, while for validation a different data set was used covering another 3 months of data.

It was found that the quality of the observations is of profound importance and highly affects the modelling results. Monitoring data gathered by inline sensors can greatly suffer from inconsistencies, faults and drifts that require an intensive data cleaning operation before it can be applied in a modelling study. Based on the data cleaning results acquired in this study, it is fair to state that the preprocessing of high-resolution operational data from municipal wastewater treatment plants can result in significant (12 to 16% on average) losses of data, in particular with data originating from fault-prone monitoring techniques (such as is the case for DO and TSS sensors), due to the high percentage of outliers and faulty data detected. Daily laboratory analyses data typically show better reliability and can be used to reconcile the high-frequent data by removing drifts and biases.

Another important aspect of data quality is related to the information being collected in relationship to the location of sampling points. With voluminous and complex treatment systems, it is often the case that sampling points are located at both the entry and exit of treatment stages (e.g. NIT), thereby monitoring the 'average' performance of a group of batteries or of a battery that groups multiple units that produce a 'mixed' effluent. However, the biofilter models developed so far describe the behaviour of an individual filter, which is impacted by the accumulation of biomass and suspended solids in the filter bed, necessitating regular backwashing cycles that temporarily halt treatment. Since negative flow directions can cause numerical problems for the model integrator, the washing sessions are modelled in the same (continuous) flow direction as that during treatment. This can result in an overestimation of biomass and suspended solids in the effluent of a modelled biofilter when compared to the averaged mixed quality monitored in the real system. Caution has been paid to avoid such errors from occurring.

A previously developed biofilter model (Zhu, 2020) has been successfully recalibrated using operational data from a large wastewater treatment plant. The model simulates a 3-stage nitrogen removal in a sequence of predenitrification, nitrification and postdenitrification processes.

The recalibrated mechanistic model is capable of predicting the effluent quality variables at the predenitrification stage with acceptable errors compared to the observations. This is particularly the case for the NO_3-N, NH_4-N, total and soluble COD, and TSS. Measured values that fall in the higher range tend to be more underestimated by the model, while lower values are generally overestimated. This could indicate that one or more kinetic rate parameter(s) are too high or that one or more diffusion constant(s) are too low. However, this could also be related to uncertainty in measurements, when probes and sensors have difficulty with extreme high or low concentrations.

The Janus-coefficient, which evaluates the capability of the model to perform outside trained conditions, i.e. validation, obtained good values for important effluent variables and the model is thus generally quite satisfactory in simulating effluent variables in terms of trends and amplitudes, except for the summer period, when it was shown to suffer from overestimation of COD. The validation of the nitrogen components (NO_3-N and NO_2-N) simulations resulted in a Janus coefficient of around 1.5, which is an excellent result and an indicator of well-modelled biological behaviour. TSS simulations under validation yield a similar coefficient reflecting the good description of filtration by the model. This builds confidence in the model's validity outside trained conditions.

The modelled output of the Nitrification stage fit the observations very well for NO_3-N, soluble COD, TSS and NH_4-N with an absolute mean error range between 0.1 to 3.7 mg/L. There is, however, a consistent overestimation of the nitrite concentrations by the model that could be explained by the

instability of NO_2-N in samples and the difficulty to measure it in the lower range. In the case of total COD, the model error is stronger than average with a mean error of -10.9 mgO_2/L. This is partly explained by the contribution to total COD of particulate organic matter, which is overestimated by the model as indicated by the TSS mean average of -3.73mg/L during the same period.

Concerning the postdenitrification stage, the daily averaged simulations of the modelled nitrates and ammonia concentrations in the effluent fit the observations over the whole range quite well. The effluent values for total COD and soluble COD are however overestimated, as well as the effluent nitrite concentrations. Overall, the modelled TSS concentrations follow the trends observed, although it starts too high. The negative mean average scores for all variables but NO_3-N confirm that the model in general overestimates the effluent concentrations in the postdenitrification stage.

7.1.3 Hybrid Model Development

The second part of this thesis involved the development of a hybrid model that can rapidly correct for the mechanistic model's error. It was developed as a parallel cooperative structure, where both models receive the same input information of the influent water quality and flow rate and operational parameters, such as the dosing of methanol and the number of active biofilters. However, the data-driven component received in addition the mechanistically modelled effluent quality. It was trained on the basis of these variables to estimate the mechanistic model's error. The data-driven model component used data for the training (01.12 – 31.12.2019) and testing (01.03 – 15.03.2020) period from the predenitrification stage at the wastewater plant being studied.

The developed Hybrid Model was capable of significantly increasing the performance of the mechanistic biofilter model during training, which is to be expected, but, importantly, also during testing. A cross-validation technique was applied that partitions the data into 5 folds and trains a model for each subset using the corresponding data; and then calculates the average validation error over all subsets. This method makes efficient use of all the data and still gives a good estimate of the predictive accuracy of the trained model, although it can take a longer time to execute because the model is trained repeatedly on the subsets.

The results show how the hybrid model outperforms and significantly reduces the size of the mechanistic model's prediction errors of the effluent nitrate concentration from a relative mean error of 12% for the mechanistic model to 2% for the hybrid model during training. While the error on nitrate simulations increased to 8% during hybrid model testing, this is significantly lower than the error of 110% for the mechanistic model under the same testing conditions. The overall good performance of the hybrid model allows for future applications of hybrid biofilm models in an operational context, such as for process control with a Hybrid Model Predictive Controller.

The Janus coefficient for NO_3-N, total COD and TSS can be considered to be good, confirming the validity of the model. When taking into account the highly dynamic conditions observed in the secondary treatment biofiltration lane of the super-large wastewater treatment plant under study, the considerable measurement errors observed in the high-resolution monitoring of quality effluent variables, and the sensitivity of the model calculations to outliers and faults in the input data; it is fair to conclude that the modelling results could further benefit from more intense (and automated) data preparation and reconciliation processes, as well as more data from longer training and testing periods.

7.1.4 Case-study: development of a Hybrid Model Predictive Controller

The hybrid model developed in this thesis was then tested in an operational application in the context of a methanol dosing controller applied to a full-scale predenitrification biofilter model. A Hybrid Model Predictive Controller was developed that uses a moving simulation window (15 minutes intervals) to predict future effluent nitrate concentrations based on multiple optimization scenarios consisting of a sequence of proposed control actions (flow of methanol dosing). The impact of the optimal dosing action on the real physical system was studied by sending the action to a second mechanistic biofilter model.

High-level scenarios were defined to evaluate different trial configurations of (Hybrid) Model Predictive Controllers over 3 hours using operational data from a period that experienced external carbon dosing. The results of this case study indicate that a Hybrid Model Predictive Control is able to significantly improve the effluent water quality when compared to the current dosing control. However, this results in higher methanol usage and thus operating costs. More extensive simulations over longer prediction horizons will permit evaluating the stability of the (H)MPC controllers.

As far as known, the HMPC developed here is one of the first implementations of a hybrid MPC on an operational municipal wastewater treatment facility. The only other known example is described by Sparks et al. (2022) using a similar approach as developed here, but for the application of the control of chlorine dosing for disinfection and for aeration control in an activated sludge system.

7.2 Perspectives

7.2.1 Data Quality

The sensitivity of models, mechanistic and data-driven, to the quality of data is an important aspect of the modelling exercise. As mentioned, the quality of measured data can be improved by applying (automated) data cleaning methods, such as those based on faults and outliers detection. The final accepted cleaned data series will contain data gaps, imposing an additional data gap filling technique. The cleaning of raw data by machine learning techniques could be integrated with an influent generator model that allows gap filling in order to guarantee the quality of the data that feeds the model (e.g. Li & Vanrolleghem, 2022).

However, model parameters and parameters for fractionation also have an important effect. Influent characterization is often based on fractionation models developed for ASM models and uses parameter values that come as default in ASM models, but do not necessarily apply for biofiltration treatment and for the conditions prevalent in the Paris region. Since no direct measurements of the fractionation parameters have been made in the framework of this research, values found in other studies focusing on similar treatment processes have been used as a best second approach. Nonetheless, these data are wastewater-specific and differences in wastewater composition can occur due to climatic differentiation, topography and consumer habits (e.g. Van Nieuwenhuijzen et al., 2004). Several studies indicate that a more detailed characterisation of municipal wastewater will yield a better determination of the fractionation parameters; in particular for the particulate matter (e.g. Levine et al., 1991; Roeleveld & Van Loosdrecht, 2002). In order to be able to interpret the wastewater composition of the treatment plant being studied, it is recommended to perform laboratory analyses that allow determination of the values of these fractions. Additionally, data-driven models could be applied that allow estimating model parameters (e.g. Mesta et al., 2023).

7.2.2 Mechanistic Model Calibration and Validation

The results of the recalibrated model for the predenitrification (preDN) stage show a generally good performance for the effluent variables. However, for the nitrification (NIT) and postdenitrification (postDN) stages this is not always the case. The errors for all variables but the NO_3-N confirm that the model in general overestimates the effluent concentrations at the NIT and postDN stages. The parameters for these modelled stages have not been recalibrated and its values were taken from previous modelling studies that were based on data collected at other wastewater treatment plants using similar biofiltration systems. However, since the wastewater plant under study here consists of exceptionally large and complex treatment systems that have important dynamics, it is strongly recommended to recalibrate the NIT and postDN stage models with data collected at the same wastewater plant under study for the preDN stage.

7.2.3 Hybrid Model Training

As shown in this study, the Hybrid Model outperforms the mechanistic model both during calibration/training and during validation/testing. However, the performance of the hybrid model reduces fast when it is exposed to 'new' untrained conditions. More research is needed to evaluate the stability of the model and to quantify the recommended time period for recalibration. A training period that contains data from longer periods, including different seasons, and/or regular re-training of the neural network would definitely increase the robustness of the model. A connection with a database storing pre-treated and cleaned inline measurement and laboratory observational data and/or the Supervisory Control and Data Acquisition (SCADA) system will allow a data-driven model (DDM) to update its training dataset on a regular base, such as being described by Sparks et al. (2022). The frequency of training the model and the size of the dataset to be used remain to be determined.

In addition to the parallel structure of the hybrid model developed in this study, where the data-driven model component consists of a trained neural network, there is increased interest in the academic world for the application of multiple data-driven models that support each other, commonly referred to as 'ensemble modelling', as testified by the large number of scientific papers appearing on this topic (e.g.: Wai et al., 2022). Ensemble models are developed by combining multiple algorithms, different Machine Learning and/or statistical models, such as simple and weighted linear averaging, and neural networks. Each model possesses unique individual characteristics that allow certain inherent features to be extracted. A common application is where ensemble models combine complementary advantages to achieve better forecasting performance (Wai et al., 2022). In recently published works, the ensemble models studied outperform their standalone counterparts and add versatility and interpretability (e.g.: Nourani et al., 2018; Sharafati et al., 2020).

It might be interesting to apply an ensemble modelling approach to the data-driven component of the hybrid model developed here. In a first step, different machine learning models could be combined to evaluate the performance of estimating the mechanistic model's error. The performance of the stand-alone ANN model developed in this study (chapter 5) is dependent on the model structure and configuration of hyper-parameters. It is possible that its performance could benefit from a second regression model that has (as much as possible) unique mechanisms in approaching the same regression problem, such that the resulting errors of the two models should not be highly correlated.

Another approach worth investigating, is training a DDM to actually estimate the effluent variable itself. In this regard, both the DDM and the mechanistic model make the same prediction of the

effluent concentration of the variable of interest, which can then be combined using a coefficient or factor (e.g.: Ghosh et al., 2019). It is a hybrid modelling technique that is considered by some as an ensemble approach, in which different components seek to solve the same problem independently to improve the ensemble model performance (Schneider et al., 2022). This way, the performance quality of each model type can be assessed and expressed in quantification terms. This approach is especially interesting in the case where certain (independent) process behaviours can be uncoupled so that each effect can be represented in a separate model (von Stosch et al., 2014). The main difficulty here remains however the definition of the weighting factors that allow combining the output of the two models, which are affected by the large variabilities observed in the availability and quality of measured data and the dynamic conditions prevailing in wastewater treatment systems.

7.2.4 HMPC Application for Methanol Dosing

The development of a Hybrid Model Predictive Controller for the addition of external carbon in the form of methanol in a denitrification biofilter aimed to showcase the advantages of a hybrid model in process control and optimization. The goal was to have a proof-of-concept; and not a plug-and-play kind of software. This means that many improvements can be made to the programming code and the modular Python scripts; this will allow removing redundancies and speed up the computational time.

In this study, it has been shown that implementing a (H)MPC automatic controller can significantly improve the effluent quality in terms of nitrate concentrations. However, this comes at an increased cost for the consumption of external carbon and associated greenhouse gas emissions. Other studies highlight the existence of trade-offs and the need for a balancing act of several objectives (e.g. Flores-Alsina et al., 2011). Sweetapple and Butler (2014) demonstrate that significant improvements can be realised with the application of a multi-objective optimisation of wastewater treatment plant control. The authors apply a multi-objective evolutionary algorithm (NSGA-II) to derive sets of Pareto optimal solutions for operational and control parameter values of a conventional activated sludge (CAS) system, with objectives including minimisation of greenhouse gas (GHG) emissions, operational costs and effluent pollutant concentrations. The set of Pareto-optimal solutions are non-dominated based upon a given objective set so that it is not possible to further improve in terms of any one objective without worsening another. It is very interesting to enhance the current optimization problem applied here for biofiltration by a multi-objective optimization that investigates the relationships and trade-offs between the operational costs of methanol dosing, the resultant effluent water quality and GHG emissions. The mechanistic model developed in Chapter 4 allows performing such optimization studies.

Furthermore, in this study the COD content of the influent of the predenitrification (preDN) stage is considered to be the limiting constraint of the biokinetics; hence the need for external carbon dosing to enable fast growth of bacteria that transform nitrates and nitrites. However, typically, denitrification is both carbon and NOx (sum of nitrate and nitrite) limited; since without carbon there are insufficient active bacteria, but without a substantial amount of NOx there is nothing to grow on. It is therefore interesting to study a control strategy that engages the recirculation flow rate of nitrified water from the effluent of the Nitrification (NIT) stage back to the anoxic zone at the entry of the predenitrification (preDN) stage. In this proposed strategy, the optimizer also includes control over the internal recycle pumps that allows for variable speeds between the lower limit (pump is turned off) and the higher limit (pump is at maximum speed).

Additionally, it could be interesting to look at the downstream effect of process optimization of methanol dosing. It is known that conventional methanol dosing controllers do not pay attention to the conversions by methylotrophs and add methanol even when insufficient active biomass is present in the reactor. The over-dosing of carbon raises the chemical costs and results in addition of unused methanol to the treatment stages that are next (Rieger et al., 2015). In the case of the Seine aval WWTP, this could lead to a loading of the Nitrification stage with readily biodegradable COD, which can reduce the performance of the autotrophs and increases the need for additional aeration. The objective of the proposed control solution would be to include the trade-off of methanol dosing to minimize soluble COD loading of the NIT aerated zone while ensuring that sufficient methanol is added to allow a sufficiently high population of methylotrophs to exist, and that takes into consideration additional aeration costs in the NIT stage.

7.3 Take Home Message

Overall, this study demonstrates the potential of a hybrid modelling approach that combines a mechanistic biofilm model describing the behaviour of denitrifying and nitrifying bacteria under anoxic and aerobic conditions, with a data-driven model that compensates for the mechanistic model's simulation error of the effluent quality. The hybrid model developed here can better describe the denitrification behaviour in a secondary biofiltration system of a very large-scale wastewater treatment plant (6 million P.E.) than a mechanistic model alone, thus significantly improving the simulation performance of nitrogen removal in a biofilter model. The hybrid model is then tested in an operational dosing control, where it outperforms the current feed-forward control in significantly improving the effluent nitrate concentration. This study increases the confidence in hybrid process modelling using (near) real-time operational data and contributes to the application of smart water technologies in a dynamic and evolving wastewater sector.

References

Abrahart, R. J., See, L. M., & Solomatine, D. P. (Eds.). (2008). Practical hydroinformatics: computational intelligence and technological developments in water applications (Vol. 68). Springer Science & Business Media.

Acevedo, V., Kaiser, T., Babist, R., Fundneider, T. & Lackner, S. (2021). A multi-component model for granular activated carbon filters combining biofilm and adsorption kinetics. Water Res., 197, 117079.

Agarwal, M., Singh, K., Dohare, R. K., & Upadhyaya, S. (2016). Process control and optimization of wastewater treatment plants using simulation softwares: a review. International Journal of Advanced Technology and Engineering Exploration, 3(22), 145.

Aguado, D., & Rosen, C. (2008). Multivariate statistical monitoring of continuous wastewater treatment plants. Engineering Applications of Artificial Intelligence, 21(7), 1080-1091.

Ahnert, M., Marx, C., Krebs, P. & Kuehn, V. (2016). A black-box model for generation of site-specific WWTP influent quality data based on plant routine data. Water Science and Technology 74 (12), 2978–86.

Ajala, O. J., Tijani, J. O., Salau, R. B., Abdulkareem, A. S., & Aremu, O. S. (2022). A review of emerging micro-pollutants in hospital wastewater: Environmental fate and remediation options. Results in Engineering, 100671.

Al Aani, S., Bonny, T., Hasan, S. W., & Hilal, N. (2019). Can machine language and artificial intelligence revolutionize process automation for water treatment and desalination?. Desalination, 458, 84-96.

Alferes, J. & Vanrolleghem, P. A. (2014). Automated data quality assessment: Dealing with faulty on-line water quality sensors. In Proceedings 7th International Congress on Environmental Modelling and Software (iEMSs2014), San Diego, USA.

Alferes J. & Vanrolleghem P.A. (2016). Efficient automated quality assessment: Dealing with faulty on-line water quality sensors. AI Communications, 2016, vol. 29, No. 6, p. 701-709.

Åmand L., Olsson G. & Carlsson B. 2013 Aeration control – a review. Water Science and Technology 67 (11), 2374–2398.

Anderson, J. S., McAvoy, T. J. & Hao, O. J. (2000). Use of hybrid models in wastewater systems. Industrial & Engineering Chemistry Research 39(6), 1694–1704.

Andrews, J. F. (1974). Dynamic models and control strategies for wastewater treatment processes. Water Research, 8(5), 261-289.

Andrews, N., Willis, J., & Muller, C. (2016). Assessment of technology advancements for future energy reduction. IWA Publishing, London, UK.

Antwi, P., Zhang, D., Luo, W., wen Xiao, L., Meng, J., Kabutey, F. T., Ayivi, F. & Li, J. (2019). Performance, microbial community evolution and neural network modeling of single-stage nitrogen removal by partial-nitritation/anammox process. Bioresource technology, 284, 359-372.

Arnaldos M., Amerlinck Y., Rehman U., Maere T., Van Hoey S., Naessens W. & Nopens I. 2015 From the affinity constant to the half-saturation index: understanding conventional modeling concepts in novel wastewater treatment processes. Water Research 70, 458–470.

Azimi S. and Rocher V. (2017). Energy consumption reduction in a wastewater treatment plant. Water Practice and Technology, 12(1), 104–116.

Baeten J. E., van Loosdrecht M. C. M. & Volcke E. I. P. 2018 Modelling aerobic granular sludge reactors through apparent half-saturation coefficients. Water Research 146, 134–145.

References

Baeza, J. A., Gabriel, D., & Lafuente, J. (2002). In-line fast OUR (oxygen uptake rate) measurements for monitoring and control of WWTP. Water Science and Technology, 45(4-5), 19-28.

Bahramian, M., Dereli, R. K., Zhao, W., Giberti, M., & Casey, E. (2022). Data to Intelligence: The Role of Data-Driven Models in Wastewater Treatment. Expert Systems with Applications, 119453.

Bartman, A.R., Christofides, P.D., Cohen, Y., (2009). Nonlinear model-based control of an experimental reverse-osmosis water desalination system. Ind. Eng. Chem. Res. 48, 6126-6136

Batstone D. J., Amerlinck Y., Ekama G., Goel R., Grau P., Johnson B., Kaya I., Steyer J. P., Tait S., Takács I., Vanrolleghem P. A., Brouckaert C. J. & Volcke E. (2012) Towards a generalized physicochemical framework. Water Science and Technology 66 (6), 1147–1161.

Beider, P. (2002). Future investment in drinking water and wastewater infrastructure. Congressional Budget Office.

Belanche, L. A., Valdés, J. J., Comas, J., Roda, I. R., & Poch, M. (1999). Towards a model of input–output behaviour of wastewater treatment plants using soft computing techniques. Environmental modelling & software, 14(5), 409-419.

Belloir, C., Stanford, C., & Soares, A. (2015). Energy benchmarking in wastewater treatment plants: the importance of site operation and layout. Environmental technology, 36(2), 260-269.

Bemporad, A., Borrelli, F., & Morari, M. (2000). Optimal controllers for hybrid systems: Stability and piecewise linear explicit form. In Proceedings of the 39th IEEE Conference on Decision and Control (Cat. No. 00CH37187) (Vol. 2, pp. 1810-1815). IEEE.

Bernardelli, A., Marsili-Libelli, S., Manzini, A., Stancari, S., Tardini, G., Montanari, D., G. Anceschi, P. Gelli & Venier, S. (2020). Real-time model predictive control of a wastewater treatment plant based on machine learning. Water Science and Technology, 81(11), 2391-2400.

Bernier, J. (2014). Modélisation Simultanée de l'Enlèvement des Nutriments et de l'Evolution de la Perte de Charge en Biofiltration des Eaux Usées. (Thèse de Doctorat). Université Laval, Québec, QC, Canada.

Bernier, J., Rocher, V., & Lessard, P. (2014a). Modelling headloss and two-step denitrification in a full-scale wastewater post-denitrifying biofiltration plant. Journal of Environmental Engineering and Science, 9(3), 171-180.

Bernier J., Rocher V., Guerin S. & Lessard P. (2014b). Modeling the nitrification in a full-scale tertiary biological aerated filter unit. Bioprocess and Biosystems Engineering, 37(2), 289-300.

Bernier J., Guérin S., Azimi S., Rocher V. & Lessard P. (2016). Modélisation du fonctionnement des décanteurs physico-chimiques lamellaires. Calibration et validation à l'échelle industrielle d'un modèle simple à une dimension. L'Eau, l'Industrie, les Nuisances, 391, 75-82.

Blumensaat, F., Leitão, J. P., Ort, C., Rieckermann, J., Scheidegger, A., Vanrolleghem, P. A., & Villez, K. (2019). How urban storm-and wastewater management prepares for emerging opportunities and threats: Digital transformation, ubiquitous sensing, new data sources, and beyond-a horizon scan. Environmental science & technology, 53(15), 8488-8498.

Bollon, J., Filali, A., Fayolle, Y., Guerin, S., Rocher, V., & Gillot, S. (2016). Full-scale post denitrifying biofilters: sinks of dissolved N2O?. Science of the Total Environment, 563, 320-328.

Boltz, J. P., & Daigger, G. T. (2010). Uncertainty in bulk-liquid hydrodynamics and biofilm dynamics creates uncertainties in biofilm reactor design. Water Science and Technology, 61(2), 307-316.

Boltz, J.P., Morgenroth, E., Brockmann, D., Daigger, G.T., Henze, M., Rittmann, B., Sorensen, K.H., Takács, I., Vanrolleghem, P.A. and van Loosdrecht, M., (2013). A biofilm reactor model calibration framework. In 9th. International Conference on Biofilm Reactors.

Borzooei, S., Miranda, G. H., Abolfathi, S., Scibilia, G., Meucci, L., & Zanetti, M. C. (2020). Application of unsupervised learning and process simulation for energy optimization of a WWTP under various weather conditions. Water Science and Technology, 81(8), 1541-1551.

Bourgeois, W., & Stuetz, R. M. (2000). Measuring wastewater quality using a sensor array: prospects for real-time monitoring. Water Science and Technology, 41(12), 107-112.

Bourgeois, W., Hogben, P., Pike, A., & Stuetz, R. M. (2003). Development of a sensor array based measurement system for continuous monitoring of water and wastewater. Sensors and Actuators B: Chemical, 88(3), 312-319.

Briggs, R. (1967). Monitoring and Automatic Control of Dissolved Oxygen Levels in Activated Sludge, Effluent and Water Treatment Convention, London.

Brok, N. B., Munk-Nielsen, T., Madsen, H., & Stentoft, P. A. (2020). Unlocking energy flexibility of municipal wastewater aeration using predictive control to exploit price differences in power markets. Applied Energy, 280, 115965.

Brun, R., Kühni, M., Siegrist, H., Gujer, W., & Reichert, P. (2002). Practical identifiability of ASM2d parameters—systematic selection and tuning of parameter subsets. Water research, 36(16), 4113-4127.

Busby, J.B. and Andrews, J.F. (1975). Dynamic Modelling and Control Strategies for the Activated Sludge Process, Journal WPCF, 47(5).

Camacho, E. F., Ramírez, D. R., Limón, D., De La Peña, D. M., & Alamo, T. (2010). Model predictive control techniques for hybrid systems. Annual reviews in control, 34(1), 21-31.

Carlson, S., & Walburger, A. (2007). Energy index development for benchmarking water and wastewater utilities. American Water Works Association.

Carpenter, E. J., Anderson, S. J., Harvey, G. R., Miklas, H. P., & Peck, B. B. (1972). Polystyrene spherules in coastal waters. Science, 178(4062), 749-750.

Cheng, G., Ramirez-Amaro, K., Beetz, M. & Kuniyoshi, Y. (2019). Purposive learning: robot reasoning about the meanings of human activities. Sci. Rob. 4.

Cheng, X., Guo, Z., Shen, Y., Yu, K., & Gao, X. (2021). Knowledge and data-driven hybrid system for modeling fuzzy wastewater treatment process. Neural Computing And Applications, 1-22.

Choubert, J. M., Marquot, A., Stricker, A. E., Racault, Y., Gillot, S. & Héduit, A. (2009). Anoxic and aerobic values for the yield coefficient of the heterotrophic biomass: Determination at full-scale plants and consequences on simulations. Water Sa, 35(1).

Chu K. H., van Veldhuizen H. M. & van Loosdrecht M. C. M. (2003). Respirometric measurement of kinetic parameters: effect of activated sludge floc size. Water Science and Technology 48 (8), 61–68.

Cohen-Shapira, N., & Rokach, L. (2022). Learning dataset representation for automatic machine learning algorithm selection. Knowledge and Information Systems, 64(10), 2599-2635.

References

Copp, J.B. (Ed.), (2002). The COST simulation benchmark—description and simulator manual. COST action 624, Office for Official Publications of the European Communities, Luxembourg.

Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U., & Poch, M. (2018). Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. Environmental modelling & software, 106, 89-103.

Cosenza A., Mannina G., Vanrolleghem P.A. & Neumann M.B. (2013). Global sensitivity analysis in wastewater applications: A comprehensive comparison of different methods. Environmental Modeling & Software, 49, 40-52.

Costa, C. (2022). A Comprehensive View of the ASM1 Dynamic Model: Study on a Practical Case. Water, 14(7), 1046.

Cote, M., Grandjean, B. P., Lessard, P., & Thibault, J. (1995). Dynamic modelling of the activated sludge process: improving prediction using neural networks. Water research, 29(4), 995-1004.

Daigger G. T. (2014). Oxygen and carbon requirements for biological nitrogen removal processes accomplishing nitrification, nitritation, and anammox. Water Environment Research 86 (3), 204–209.

Daigger, G. T. (2017). Flexibility and adaptability: Essential elements of the WRRF of the future. Water Practice and Technology, 12(1), 156-165.

Daigger, G. T., Voutchkov, N., Lall, U. & Sarni, W. (2019). The Future of Water: A collection of essays on “disruptive” technologies that may transform the water sector in the next 10 years. Inter-American Development Bank, Water and Sanitation Division (IDB-DP-657).

Dalmau, M., Atanasova, N., Gabarrón, S., Rodriguez-Roda, I., & Comas, J. (2015). Comparison of a deterministic and a data driven model to describe MBR fouling. Chemical Engineering Journal, 260, 300-308.

Dellana, S. A., & West, D. (2009). Predictive modeling for wastewater applications: Linear and nonlinear approaches. Environmental modelling & software, 24(1), 96-106.

Descoins, N., Deleris, S., Lestienne, R., Trouvé, E. & Maréchal, F. (2012). Energy efficiency in waste water treatments plants: Optimization of activated sludge process coupled with anaerobic digestion. Energy, 41(1), 153-164.

Dris, R., Gasperi, J., Rocher, V., Saad, M., Renault, N., & Tassin, B. (2015). Microplastic contamination in an urban area: a case study in Greater Paris. Environmental Chemistry, 12(5), 592-599.

Dochain, D., & Vanrolleghem, P. A. (2001). Dynamical modelling & estimation in wastewater treatment processes. IWA publishing.

Dolejš P., Varga Z., Luza B., Pícha A., Jeníček P., Jeison D. & Bartáček J. (2019). Maximizing energy recovery from wastewater via bioflocculation-enhanced primary treatment: a pilot scale study. Environ Technol 1–11

Dürrenmatt, D. J., & Gujer, W. (2012). Data-driven modeling approaches to support wastewater treatment plant operation. Environmental Modelling & Software, 30, 47-56.

Eberl H., Morgenroth E., Noguera D.R., Picioreanu C., Rittmann B., Van Loosdrecht M. & Wanner O. (2006). Mathematical Modeling of Biofilms (IWA Task Group on Biofilm Modelling, ed.) Scientific and Technical Report No. 18, IWA Publishing, London, UK.

Eggimann, S., Mutzner, L., Wani, O., Schneider, M. Y., Spuhler, D., Moy de Vitry, M., Beutler, P. & Maurer, M. (2017). The potential of knowing more: A review of data-driven urban water management. Environmental science & technology, 51(5), 2538-2553.

Elías-Maxil, J. A., Van Der Hoek, J. P., Hofman, J., & Rietveld, L. (2014). Energy in the urban water cycle: Actions to reduce the total expenditure of fossil fuels with emphasis on heat reclamation from urban water. Renewable and Sustainable Energy Reviews, 30, 808-820.

Estrada, J. M., Kraakman, N. B., Muñoz, R., & Lebrero, R. (2011). A comparative analysis of odour treatment technologies in wastewater treatment plants. Environmental science & technology, 45(3), 1100-1106.

European Union: European Commission (2015). Closing the loop - An EU action plan for the Circular Economy. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. 614/2.

European Union: European Commission (2022). Proposal for a directive of the european parliament and of the council concerning urban wastewater treatment (recast). 26 October 2022. COM(2022) 541 final.

Fan, M., Hu, J., Cao, R., Ruan, W., & Wei, X. (2018). A review on experimental design for pollutants removal in water treatment with the aid of artificial intelligence. Chemosphere, 200, 330-343.

Fenu, A.; Guglielmi, G.; Jimenez, J.; Spérandio, M.; Saroj, D.; Lesjean, B.; Brepols, C.; Thoeye, C. & Nopens, I. (2010). Activated sludge model (ASM) based modelling of membrane bioreactor (MBR) processes: A critical review with special regard to MBR specificities. Water Res., 44, 4272–4294.

Fernández, C., Vega, J. A., Fonturbel, T. & Jiménez, E. (2009). Streamflow drought time series forecasting: a case study in a small watershed in North West Spain. Stochastic Environ. Res. Risk Assess. 23, 1063–1070.

Fernández-Arévalo T., Lizarralde I., Fdz-Polanco F., Pérez-Elvira S., Garrido J., Puig S., Poch M., Grau P. & Ayesa E. (2017). Quantitative assessment of energy and resource recovery in wastewater treatment plants based on plant-wide simulations. Water Research 118, 272–288.

Fito, J., & Van Hulle, S. W. (2021). Wastewater reclamation and reuse potentials in agriculture: towards environmental sustainability. Environment, Development and Sustainability, 23, 2949-2972.

Flores-Alsina, X., Gallego, A., Feijoo, G., & Rodriguez-Roda, I. (2010). Multiple-objective evaluation of wastewater treatment plant control alternatives. Journal of Environmental Management, 91(5), 1193-1201.

Flores-Alsina, X., Corominas, L., Snip, L., & Vanrolleghem, P. A. (2011). Including greenhouse gas emissions during benchmarking of wastewater treatment plant control strategies. Water Research, 45(16), 4700-4710.

Foladori, P., Vaccari, M., & Vitali, F. (2015). Energy audit in small wastewater treatment plants: methodology, energy consumption indicators, and lessons learned. Water Science and Technology, 72(6), 1007-1015.

Garrido-Baserba, M., Corominas, L., Ces, U., Rosso, D. & Poch, M. (2020). The fourth-revolution in the water sector encounters the digital revolution. Environmental Science & Technology, 54(8), 4698-4705.

Garrido-Baserba, M., Barnosell, I., Molinos-Senante, M., Sedlak, D. L., Rabaey, K., Schraa, O., Verdaguer, M., Rosso, D. & Poch, M. (2022). The third route: A techno-economic evaluation of extreme water and wastewater decentralization. Water Research, 218, 118408.

References

Gebremariam S. Y., Beutel M. W., Christian D. & Hess T. F. 2011 Research advances and challenges in the microbiology of enhanced biological phosphorus removal–a critical review. Water Environment Research 83 (3), 195–219.

Gernaey, K. V., Van Loosdrecht, M. C., Henze, M., Lind, M., & Jørgensen, S. B. (2004). Activated sludge wastewater treatment plant modelling and simulation: state of the art. Environmental modelling & software, 19(9), 763-783.

Ghosh, D., Hermonat, E., Mhaskar, P., Snowling, S. & Goel, R. (2019). Hybrid modeling approach integrating first-principles models with subspace identification. Ind. Eng. Chem. Res. 58, 13533–13543.

Gikas, P. (2017). Towards energy positive wastewater treatment plants. Journal of environmental management, 203, 621-629.

Giorgetti, N., Bemporad, A., Tseng, H. E., & Hrovat, D. (2006). Hybrid model predictive control application towards optimal semi-active suspension. International Journal of Control, 79(05), 521-533.

Glassey, J., & Von Stosch, M. (Eds.). (2018). Hybrid modeling in process industries. CRC Press.

Goldstein, R., & Smith, W. E. P. R. I. (2002). Water & sustainability (volume 4): US electricity consumption for water supply & treatment-the next half century. Electric Power Research Institute.

Gonzalez-Rivas, N., Reyes-Pérez, H., & Barrera-Díaz, C. E. (2019). Recent advances in water and wastewater electrodisinfection. ChemElectroChem, 6(7), 1978-1983.

Grieves, M. (2014). Digital twin: manufacturing excellence through virtual factory replication. White paper, 1(2014), 1-7.

Guerra-Rodríguez, S., Oulego, P., Rodríguez, E., Singh, D. N., & Rodríguez-Chueca, J. (2020). Towards the implementation of circular economy in the wastewater sector: Challenges and opportunities. Water, 12(5), 1431.

Gujer, W. (2011). Is modeling of biological wastewater treatment a mature technology?. Water Science and Technology, 63(8), 1739-1743.

Hadjimichael, A., Comas, J., & Corominas, L. (2016). Do machine learning methods used in data mining enhance the potential of decision support systems? A review for the urban water sector. Ai Communications, 29(6), 747-756.

Haimi, H., Mulas, M., Corona, F., Marsili-Libelli, S., Lindell, P., Heinonen, M. & Vahala, R. (2016). Adaptive data-derived anomaly detection in the activated sludge process of a large-scale wastewater treatment plant. Eng. Appl. Artif. Intell. 52, 65–80.

Hamed, M. M., Khalafallah, M. G., & Hassanien, E. A. (2004). Prediction of wastewater treatment plant performance using artificial neural networks. Environmental Modelling & Software, 19(10), 919-928.

Hauduc H., Neumann M.B., Muschalla D. , Gamerith V., Gillot S. & Vanrolleghem P.A. (2015). Efficiency criteria for environmental model quality assessment: A review and its application to wastewater treatment. Environmental Modeling & Software, 68, 196-204.

Henze M., Van Loosdrecht Mc, Ekama G.A. & Brdjanovic D. (Eds.). (2008). Biological Wastewater Treatment. IWA Publishing, London, UK.

Henze M., Gujer W., Mino T. & Van Loosdrecht M.C. (2000). Activated Sludge Models ASM1, ASM2, ASM2d and ASM3. IWA Publishing, London, UK.

Hermansson, A.W. & Syafiie, S., (2015). Model predictive control of pH neutralization processes: a review. Contr. Eng. Pract. 45, 98-109.

Hiatt, W. C., & Grady, C. L. (2008). An updated process model for carbon oxidation, nitrification, and denitrification. Water Environment Research, 80(11), 2145-2156.

Hu, H.; Liao, K.; Xie, W.; Wang, J.; Wu, B. & Ren, H. (2020). Modeling the formation of microorganism-derived dissolved organic nitrogen (mDON) in the activated sludge system. Water Res., 174, 115604.

Hulsbeek, J. J. W., Kruit, J., Roeleveld, P. J., & Van Loosdrecht, M. C. M. (2002). A practical protocol for dynamic modelling of activated sludge systems. Water Science and Technology, 45(6), 127-136.

Ingildsen P. & Olsson G. (2016). Smart Water Utilities. Complexity Made Simple. IWA Publishing, London, UK.

IWA and Xylem Inc. (2019). Digital Water Report. Technical Report. London, UK.

Jansson, Å., Röttorp, J., & Rahmberg, M. (2002). Development of a software sensor for phosphorus in municipal wastewater. Journal of Chemometrics: A Journal of the Chemometrics Society, 16(8-10), 542-547.

Jenicek, P., Kutil, J., Benes, O., Todt, V., Zabranska, J., & Dohanyos, M. (2013). Energy self-sufficient sewage wastewater treatment plants: is optimized anaerobic sludge digestion the key?. Water Science and Technology, 68(8), 1739-1744.

Jeppsson, U., Rosén, C., Alex, J., Copp, J., Gernaey, K. V., Pons, M. N., & Vanrolleghem, P. A. (2006). Towards a benchmark simulation model for plant-wide control strategy performance evaluation of WWTPs. Water Science and Technology, 53(1), 287-295.

Jones, D., Snider, C., Nassehi, A., Yon, J., & Hicks, B. (2020). Characterising the Digital Twin: A systematic literature review. CIRP Journal of Manufacturing Science and Technology, 29, 36-52.

Joveniaux, A., Legrand, M., Esculier, F., & De Gouvello, B. (2022). Towards the development of source separation and valorization of human excreta? Emerging dynamics and prospects in France. Frontiers in Environmental Science, 10, 2388.

Kehrein, P., Van Loosdrecht, M., Osseweijer, P., Garfí, M., Dewulf, J., & Posada, J. (2020). A critical review of resource recovery from municipal wastewater treatment plants–market supply potentials, technologies and bottlenecks. Environmental Science: Water Research & Technology, 6(4), 877-910.

Khiewwijit R., Temmink H., Rijnaarts H. & Keesman K. J. (2015). Energy and nutrient recovery for municipal wastewater treatment: how to design a feasible plant layout? Environmental Modelling & Software 68, 156–165.

Kim, M. H., Kim, Y. S., Prabu, A. A., & Yoo, C. K. (2009). A systematic approach to data-driven modeling and soft sensing in a full-scale plant. Water Science and Technology, 60(2), 363-370.

Kolditz O. Rink K. Nixdorf E. Fischer T. Bilke L. Naumov D. Liao Z. & Yue T. (2019). Environmental information systems: paving the path for digitally facilitated water management (Water 4.0). Engineering 5 (5), 828–832.

Langergraber, G., Rieger, L., Winkler, S., Alex, J., Wiese, J., Owerdieck, C., Ahnert, M., Simon, J. & Maurer, M. (2004). A guideline for simulation studies of wastewater treatment plants. Water Science and Technology, 50(7), 131-138.

Lee, D. S., Park, J. M., & Vanrolleghem, P. A. (2005a). Adaptive multiscale principal component analysis for on-line monitoring of a sequencing batch reactor. Journal of Biotechnology, 116(2), 195-210.

Lee, D. S., Vanrolleghem, P. A., & Park, J. M. (2005b). Parallel hybrid modeling methods for a full-scale cokes wastewater treatment plant. Journal of Biotechnology, 115(3), 317-328.

Lee, D. S., Lee, M. W., Woo, S. H., Kim, Y. J., & Park, J. M. (2006). Nonlinear dynamic partial least squares modeling of a full-scale biological wastewater treatment plant. Process Biochemistry, 41(9), 2050-2057.

Lee, M. W., Hong, S. H., Choi, H., Kim, J. H., Lee, D. S., & Park, J. M. (2008). Real-time remote monitoring of small-scaled biological wastewater treatment plants by a multivariate statistical process control and neural network-based software sensors. Process Biochemistry, 43(10), 1107-1113.

Levine, A.D. & Asano, T. (2004). Recovering sustainable water from wastewater. Environ. Sci. Technol., 38, 201A–208A.

Levine, A.D., Tchobanolous, G. and Asano, T. (1991). Size distributions of particulate contaminants in wastewater and their impact on treatability. Wat. Res., 25(8), 911–922.

Li, F. & Vanrolleghem, P.A. (2022). An influential generator for WRRF design and operation based on a recurrent neural network with multi-objective optimization using a genetic algorithm. Water Science and Technology, 85(5), 1444-1453.

Li, Z., Zheng, L., Koottatep, T., & Vinnerås, B. (2023). Decentralized wastewater treatment technologies. Frontiers in Environmental Science, 11, 1199552.

Liu, H., Ramnarayanan, R. & Logan, B.E. (2004). Production of electricity during waste-water treatment using a single chamber microbial fuel cell. Environ Sci Technol; 38:2281–5.

Londong, J. (1992). Strategies for Optimised Nitrate Reduction With Primary Denitrification, Wat. Sci. Tech., 26(5), 1087–1096.

Longo, S., d'Antoni, B. M., Bongards, M., Chaparro, A., Cronrath, A., Fatone, F., Lema, J.M., Mauricio-Iglesias, M., Soares, A. & Hospido, A. (2016). Monitoring and diagnosis of energy consumption in wastewater treatment plants. A state of the art and proposals for improvement. Applied energy, 179, 1251-1268.

Lu, H., Chandran, K., & Stensel, D. (2014). Microbial ecology of denitrification in biological wastewater treatment. Water research, 64, 237-254.

Luccarini, L., Bragadin, G. L., Colombini, G., Mancini, M., Mello, P., Montali, M., & Sottara, D. (2010). Formal verification of wastewater treatment processes using events detected from continuous signals by means of artificial neural networks. Case study: SBR plant. Environmental Modelling & Software, 25(5), 648-660.

Lynggaard-Jensen, A. (1999). Trends in monitoring of waste water systems. Talanta, 50(4), 707-716.

Maere, T., Neethling, J. B., Clark, D., Pramanik, A., & Vanrolleghem, P. A. (2016). Wastewater treatment nutrient regulations: an international perspective with focus on innovation. WEF/IWA Nutrient Removal and Recovery, Colorado, USA, 10-13.

Maktabifard, M., Zaborowska, E., & Makinia, J. (2019). Evaluating the effect of different operational strategies on the carbon footprint of wastewater treatment plants–case studies from northern Poland. Water Science and Technology, 79(11), 2211-2220.

Mamais, D., Noutsopoulos, C., Dimopoulou, A., Stasinakis, A., & Lekkas, T. D. (2015). Wastewater treatment process impact on energy savings and greenhouse gas emissions. Water Science and Technology, 71(2), 303-308.

Mannina G., Cosenza A., Vanrolleghem P.A. & Viviani, G. (2011). A practical protocol for calibration of nutrient removal wastewater treatment models. Journal of Hydroinformatics, 13(4), 575-595.

Mannina G., Ekama G., Caniani D., Cosenza A., Esposito G., Gori R., Garrido-Baserba M., Rosso D. & Olsson G. (2016) Greenhouse gases from wastewater treatment – A review of modelling tools. Science of The Total Environment 551–552, 254–270.

Masson, M. H., Canu, S., Grandvalet, Y., & Lynggaard-Jensen, A. (1999). Software sensor design based on empirical data. Ecological Modelling, 120(2-3), 131-139.

Mateos, F., & Rodríguez, S. (2012). Estudio de Prospectiva Consumo Energético en el sector del agua. Ministerio de Industria, Energía Y Turismo, Fundación Escuela de Organización Industrial.

Melcer, H., Dold, P. L., Jones, R. M., Bye, C. M., Takacs, I., Stensel, H. D., Wilson, A.W., Sun, P. & Bury, S. (2003). Methods for Wastewater Characterization in Activated Sludge Modelling. Narrative, 1, 12.

Mendoza-Espinosa, L., & Stephenson, T. O. M. (1999). A review of biological aerated filters (BAFs) for wastewater treatment. Environmental engineering science, 16(3), 201-216.

Metcalf & Eddy, Abu-Orf, M., Bowden, G., Burton, F.L., Pfrang, W., Stensel, H.D., Tchobanoglous, G., Tsuchihashi, R. and AECOM (Firm), (2014). Wastewater engineering: treatment and resource recovery. McGraw Hill Education.

Morgenroth, E., & Wilderer, P. A. (2000). Influence of detachment mechanisms on competition in biofilms. Water research, 34(2), 417-426.

Morgenroth, E., Kommedal, R., & Harremoës, P. (2002). Processes and modeling of hydrolysis of particulate organic matter in aerobic wastewater treatment–a review. Water Science and Technology, 45(6), 25-40.

Morgenroth, E. (2008a). Modelling biofilms. In: Henze, M.C.M. van Loosdrecht, G.A. Ekama & D. Brdjanovic (Eds.), Biological Wastewater Treatment: Principles, Modelling and Design, pp. 457-492. IWA Publishing, London, UK.

Morgenroth, E. (2008b). Biofilm reactors. In M. Henze, M. C. M. van Loosdrecht, G. A. Ekama, & D. Brdjanovic (Eds.), Biological Wastewater Treatment: Principles, Modelling and Design, pp. 493-511. IWA Publishing, London, UK.

Nadeem, K., Plana, Q., Albasi, C., Bernier, J., Azimi, S., Rocher, V. & Alliet, M. (2021). Integrated model for simulation of super-large scale membrane bioreactor process and energy consumption. In: Proceedings Euromembrane 2021, Copenhagen, Denmark, November 28 - December 2, 2021.

Newhart, K.B., Holloway, R.W., Hering, A.S. & Cath, T.Y. (2019). Data-driven performance analyzes of wastewater treatment plants: A review. Water Research, 157, 498-513.

Ngo, H. H., Vo, H. N. P., Guo, W., Chen, Z., Liu, Y., & Varjani, S. (2020). Sustainable management and treatment technologies for micro-pollutants in wastewater. In Current Developments in Biotechnology and Bioengineering (pp. 1-22). Elsevier.

Niero, M., & Hauschild, M. Z. (2017). Closing the loop for packaging: finding a framework to operationalize Circular Economy strategies. Procedia Cirp, 61, 685-690.

Ninassi, M. V., Peladan, J. G., & Pujol, R. (1998, October). Predenitrification of municipal wastewater: the interest of up-flow biofiltration. In 71st Annual Conference and Exposition, WEFTEC (Vol. 98, pp. 445-466).

Niu, G., Yi, X., Chen, C., Li, X., Han, D., Yan, B., Huang, M. & Ying, G. (2020). A novel effluent quality predicting model based on genetic-deep belief network algorithm for cleaner production in a full-scale paper-making wastewater treatment. J. Cleaner Prod. 265, 121787.

Nourani, V., Elkiran, G., & Abba, S. I. (2018). Wastewater treatment plant performance analysis using artificial intelligence–an ensemble approach. Water Science and Technology, 78(10), 2064-2076.

References

Novak, P. J., Arnold, W. A., Henningsgaard, B., Hozalski, R. M., Kessler, K., LaPara, T. L., Parrella, A., Rogacki, L., Thompson, C., Thorson, R., Zimmerman, R.A., Bott, C.B., Daigger, G.T., & Neethling, J. B. (2015). Innovation promoted by regulatory flexibility.

Nowak, O., Keil, S., & Fimml, C. (2011). Examples of energy self-sufficient municipal nutrient removal plants. Water Science and Technology, 64(1), 1-6.

Oberdieck, R., & Pistikopoulos, E. N. (2015). Explicit hybrid model-predictive control: The exact solution. Automatica, 58, 152-159.

O'Brien, M., Mack, J., Lennox, B., Lovett, D., & Wall, A. (2011). Model predictive control of an activated sludge process: A case study. Control Engineering Practice, 19(1), 54-61.

Ocampo-Martinez, C., Bemporad, A., Ingimundarson, A., & Cayuela, V. P. (2007). On hybrid model predictive control of sewer networks. Identification and Control: The Gap between Theory and Practice, 87-114.

Ocampo-Martinez, C., & Puig, V. (2009). Fault-tolerant model predictive control within the hybrid systems framework: Application to sewer networks. international journal of adaptive control and signal processing, 23(8), 757-787.

Ocampo-Martinez, C. (2010). Suboptimal MPC approach based on piecewise linear functions. In: Grimble, M.J., Johnson, M.A. (Eds.), Model Predictive Control of Wastewater Systems, Advances in Industrial Control. Springer.

Ohasi, H., Sugawara, T., Kikuchi, K. and Konno, H. (1981). Correlation of liquid-side mass-transfer coefficient for single particles and fixed-beds. Journal of Chemical Engineering of Japan 14(6), p. 433-438.

Olsson, G. (2008). Process control. Biological Wastewater Treatment. Principles, Modelling and Design. Edited by Henze, M., van Loosdrecht, M.C.M, Ekama, G.A & Brdjanovic, D. IWA Publishing, London, 393-414.

Olsson, G., & Newell, B. (1999). Wastewater treatment systems. IWA publishing.

Olsson, G., Aspegren, H., & Nielsen, M. K. (1998). Operation and control of wastewater treatment—a Scandinavian perspective over 20 years. Water Science and Technology, 37(12), 1-13.

Olsson, G. (2012). ICA and me – a subjective review. Water Res., 46 (2012), pp. 1585-1624

Orhon, D.; Yucel, A.B.; Insel, G.; Solmaz, B.; Mermutlu, R. & Sözen, S. (2021). Appraisal of Super-Fast Membrane Bioreactors by MASM—A New Activated Sludge Model for Membrane Filtration. Water, 13, 1963.

Ortigara, A. R. C., Kay, M., & Uhlenbrook, S. (2018). A review of the SDG 6 synthesis report 2018 from an education, training, and research perspective. Water, 10(10), 1353.

Panepinto, D., Fiore, S., Zappone, M., Genon, G., & Meucci, L. (2016). Evaluation of the energy efficiency of a large wastewater treatment plant in Italy. Applied Energy, 161, 404-411.

Papias, S., Masson, M., Pelletant, S., Prost-Boucle, S., & Boutin, C. (2018). In situ continuous monitoring of nitrogen with ion-selective electrodes in a constructed wetland receiving treated wastewater: an operating protocol to obtain reliable data. Water Science and Technology, 77(6), 1706-1713.

Petersen, B., Vanrolleghem, P. A., Gernaey, K., & Henze, M. (2002). Evaluation of an ASM1 model calibration procedure on a municipal–industrial wastewater treatment plant. Journal of Hydroinformatics, 4(1), 15-38.

Philippe, R. (2019). Outils automatiques d'évaluation de la qualité des données pour le suivi en continu de la qualité des eaux usées (Doctoral dissertation, Université Laval).

Pistocchi, A., Dorati, C., Grizzetti, B., Udias, A., Vigiak, O., & Zanni, M. (2019). Water quality in Europe: effects of the Urban Wastewater Treatment Directive. European Commission Joint Research Centre, JRC115607.

Plana, Q., Alferes, J., Fuks, K., Kraft, T., Maruéjouls, T., Torfs, E., & Vanrolleghem, P. A. (2019). Towards a water quality database for raw and validated data with emphasis on structured metadata. Water Quality Research Journal, 54(1), 1-9.

Quaghebeur, W., Nopens, I., & De Baets, B. (2021). Incorporating unmodeled dynamics into first-principles models through machine learning. IEEE Access, 9, 22014-22022.

Quaghebeur, W., Torfs, E., De Baets, B., & Nopens, I. (2022). Hybrid differential equations: integrating mechanistic and data-driven techniques for modelling of water systems. Water Research, 213, 118166.

Ráduly, B., Gernaey, K. V., Capodaglio, A. G., Mikkelsen, P. S., & Henze, M. (2007). Artificial neural networks for rapid WWTP performance evaluation: Methodology and case study. Environmental modelling & software, 22(8), 1208-1216.

Rahman S. M., Eckelman M. J., Onnis-Hayden A. & Gu A. Z. (2016) Life-cycle assessment of advanced nutrient removal technologies for wastewater treatment. Environmental Science & Technology 50 (6), 3020–3030.

Razavi, S. (2021). Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling. Environmental Modelling & Software, 144, 105159.

Regmi, P., Stewart, H., Amerlinck, Y., Arnell, M., García, P. J., Johnson, B., Maere, T., Miletić, I., Miller, M., Rieger, L. and Samstag, R., Santoro D., Schraa O., Snowling S., Takács I., Torfs E., van Loosdrecht, M. C. M., Vanrolleghem, P. A., Villez, K., Volcke, E. I. P., Weijers, S., Grau, P., Jimenez J., & Rosso, D. (2019). The future of WRRF modelling–outlook and challenges. Water Science and Technology, 79(1), 3-14.

Reijken, C.; Giorgi, S.; Hurkmans, C.; Pérez, J. & van Loosdrecht, M. (2018). Incorporating the influent cellulose fraction in activated sludge modelling. Water Res., 144, 104–111.

Reinders, M., Greditgk-Hoffamnn, S., Risse, H., & Lange, M. (2012). Solution approaches for energy optimization in the water sector. In IWA world congress on water, climate and energy, Dublin, Ireland, May 2012 (pp. 13-18).

Richalet, J., Rault, A., Testud, J.L., Papon, J. (1978). Model predictive heuristic control: applications to industrial processes. Automatica 14, 413e428.

Rieger, L., Alex, J., Schraa, O., & Bott, C. B. (2015). A Highly Efficient Methanol Dosage Controller for Multi-stage WRRFs. In Proceedings of the 88th Annual Water Environment Federation Technical Exhibition and Conference, September 2015 (pp. 26-30).

Rieger L., Gillot S., Langergraber G., Ohtsuki T., Shaw A., Takacs I. & Winkler S. (2012). Guidelines for Using Activated Sludge Models. Scientific and Technical Report No. 22. IWA Publishing, London, UK.

Rieger L., Takács I. & Siegrist H. (2012). Improving nutrient removal while reducing energy use at three Swiss WWTPs using advanced control. Water Environment Research 84 (2), 170–188.

Rittmann B.E., Boltz Jp, Brockmann D., Daigger Gt, Morgenroth E., Sørensen Kh, Takacs I., Van Loosdrecht M. & Vanrolleghem, P.A. (2018). A framework for good biofilm reactor modeling practice (GBRMP). Water Science and Technology, 77(5), 1149-1164.

Rocher, V., Paffoni, C., Gonçalves, A., Azimi, S., & Gousailles, M. (2008). La biofiltration des eaux résiduaires urbaines: retour d'expérience du SIAAP. Revue des sciences de l'eau, 21(4), 475-485.

Rocher V., C. Paffoni, A. Goncalves, S. Azimi Et C. Briand (2011a). Conditions d'apparition des nitrites lors de la dénitrification des eaux usées par biofiltration. 1. Importance de la charge appliquée en azote. L'Eau, l'Industrie, les Nuisances, 344, 384-388.

Rocher V., C. Paffoni, A. Goncalves, S. Azimi, S. Guerin-Rechdaoui & A. Pauss (2011b). Conditions d'apparition des nitrites lors de la dénitrification des eaux usées par biofiltration. 2. Importance des apports en substrat carboné. L'Eau, l'Industrie, les Nuisances, 344, 389-397.

Rocher, V., Paffoni, C., Gonçalves, A., Guérin, S., Azimi, S., Gasperi, J., ... & Pauss, A. (2012a). Municipal wastewater treatment by biofiltration: comparisons of various treatment layouts. Part 1: assessment of carbon and nitrogen removal. Water Science and Technology, 65(9), 1705-1712.

Rocher, V., Paffoni, C., Gonçalves, A., Azimi, S., & Pauss, A. (2012b). Municipal wastewater treatment by biofiltration: comparisons of various treatment layouts. Part 2: assessment of the operating costs in optimal conditions. Water Science and Technology, 65(9), 1713-1719.

Roeleveld, P. J., & Van Loosdrecht, M. C. M. (2002). Experience with guidelines for wastewater characterisation in The Netherlands. Water Science and Technology, 45(6), 77-87.

Rosén, C., Röttorp, J., & Jeppsson, U. (2003). Multivariate on-line monitoring: challenges and solutions for modern wastewater treatment operation. Water science and technology, 47(2), 171-179.

Rosen, C., Jeppsson, U. and Vanrolleghem, P.A. (2004). Towards a common benchmark for long-term process control and monitoring performance evaluation. Water Science and Technology, 50 (11) (2004), pp. 41-49

Rother, E., & Cornel, P. (2004). Optimising design, operation and energy consumption of biological aerated filters (BAF) for nitrogen removal of municipal wastewater. Water Science and Technology, 50(6), 131-139.

Rother, E., & Cornel, P. (2007). Potentials and limits of a pre-denitrification/nitrification biofilter configuration for advanced municipal wastewater treatment. Water science and technology, 55(8-9), 115-123.

Saltelli A., Ratto M., Andres T., Campolongo F., Cariboni J., Gatelli D., Saisana M. & Tarantola S. (2008). Global Sensitivity Analysis: The Primer. John Wiley & Sons, Chichester, UK.

Selvakumar, A., & Tafuri, A. N. (2012). Rehabilitation of aging water infrastructure systems: key challenges and issues. Journal of Infrastructure Systems, 18(3), 202-209.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. & Tarantola, S. (2008). Global sensitivity analysis: the primer. John Wiley & Sons.

Samstag R. W., Ducoste J. J., Griborio A., Nopens I., Batstone D. J., Wicks J. D., Saunders S., Wicklein E. A., Kenny G. & Laurent J. 2016 CFD for wastewater treatment: an overview. Water Science and Technology 74 (3), 549–563.

Santos, J.M.M.; Rieger, L.; Lanham, A.B.; Carvalheira, M.; Reis, M.A.M. & Oehmen, A. (2020). A novel metabolic-ASM model for full-scale biological nutrient removal systems. Water Res., 171, 115373.

Sarpong, G., Gude, V.G. Near Future Energy Self-sufficient Wastewater Treatment Schemes. Int J Environ Res 14, 479–488 (2020).

Schneider M.Y., Quaghbeur W., Borzoei S., Froemelt A., Li F., Saagi R. & Torfs E. (2022). Hybrid modeling of water resource recovery facilities: Status and opportunities. Water Science and Technology, 85(9), 2503-2524.

Serralta J., Ribes J., Seco A. & Ferrer J. 2002 A supervisory control system for optimising nitrogen removal and aeration energy consumption in wastewater treatment plants. Water Science and Technology 45 (4–5), 309–316.

Serrao M., Jauzein V., Azimi S., Rocher V., Tassin B., Juran I. & Vanrolleghem P.A. (2022a). Decision support tools for the treatment of urban wastewater: development of a hybrid model for the optimization of the biofiltration sector. Journées Information Eaux, 11 – 13 October 2022, Poitiers, France.

Serrao M., Jauzein V., Azimi S., Rocher V., Tassin B., Juran I. & Vanrolleghem P.A. (2022b). Hybrid model of a super-large scale biofiltration wastewater treatment system to optimize process modelling. IWA Specialist Conference on Instrumentation, Control & Automation (ICA 2022), October 17-21, 2022, Beijing, China.

Serrao M., Sparks J., Mesta K., Therrien J.D., Covre F., Li N., Nicolaï N. & Vanrolleghem P.A. (2023a). Modelling of Urban Wastewater Treatment Processes covering the entire Data-Pipeline: from Raw Data to a Digital Twin to Optimize Performance and Increase Efficiency. 8th IWA Water Resource Recovery Modelling Seminar (WRRmod2022+), 18-22 January 2023, Stellenbosch, South Africa.

Serrao M. & White S. (2023b). Scientific Collaboration for Water Research. MAWAC-ENAR Water Research & Innovation Workshop; 20-21 March 2023, New York City, USA.

Sharafati, A., Asadollah, S. B. H. S., & Hosseinzadeh, M. (2020). The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. Process Safety and Environmental Protection, 140, 68-78.

Shi, S. & Xu, G. (2018). Novel performance prediction model of a biofilm system treating domestic wastewater based on stacked denoising autoencoders deep learning network. Chem. Eng. J. 347, 280–290.

Shirkoohi, M.G., Tyagi, R.D, Vanrolleghem, P. A., & Drogui, P. (2021). Artificial neural networks and genetic algorithms: An efficient modelling and optimization methodology for active chlorine production using the electrolysis process. The Canadian Journal of Chemical Engineering, 99, S389-S403.

Shirkoohi M.G. (2022). Techniques d'intelligence artificielle dans la modélisation du processus électrochimique pour le traitement des eaux résiduaires (Thesis). Université du Québec, Institut national de la recherche scientifique, Canada.

Shirkoohi, M. G., Tyagi, R. D., Vanrolleghem, P. A., & Drogui, P. (2022). Modelling and optimization of psychoactive pharmaceutical caffeine removal by electrochemical oxidation process: A comparative study between response surface methodology (RSM) and adaptive neuro fuzzy inference system (ANFIS). Separation and Purification Technology, 290, 120902.

SIAAP (2018). Innovate in the monitoring and operating practices of wastewater treatment plants. Scientific and technical lessons learned from phase I (2014-2017) of the Mocopée program. Report, SIAAP, Paris, France.

SIAAP (2019). Rapport d'activité et de développement durable 2019. Les indicateurs techniques, financiers et de développement durable. SIAAP, Paris, France.

SIAAP (2020a). Rapport d'activité et de développement durable 2020. Les indicateurs techniques, financiers et de développement durable. SIAAP, Paris, France.

SIAAP (2020b). Usine Seine aval. Saint-Germain-en-Laye, Achères, Maisons-laffitte (78). SIAAP, Paris, France.

SIAAP (2021). Rapport d'activité et de développement durable 2021. Les indicateurs techniques, financiers et de développement durable. SIAAP, Paris, France.

Sin, G., Van Hulle, S. W., De Pauw, D. J., Van Griensven, A., & Vanrolleghem, P. A. (2005). A critical comparison of systematic calibration protocols for activated sludge models: A SWOT analysis. Water Research, 39(12), 2459-2474.

Sin, G., Gernaey, K. V., Neumann, M. B., van Loosdrecht, M. C., & Gujer, W. (2011). Global sensitivity analysis in wastewater treatment plant model applications: prioritizing sources of uncertainty. Water research, 45(2), 639-651.

Sin, G. & Al, R. (2021). Activated sludge models at the crossroad of artificial intelligence – a perspective on advancing process modeling. Npj Clean Water 4, 1–7.

Solon, K., Jia, M., & Volcke, E. I. (2019). Process schemes for future energy-positive water resource recovery facilities. Water Science and Technology, 79(9), 1808-1820.

Sparks, J., Vanrolleghem, P., Bott, C., & Wadhawan, T. (2022). It's OK to be a Control Freak: Deploying machine learning algorithms and model-based controllers for WRRF optimization. In Proceedings of the WEF Innovations in Process Engineering Conference (WEF/IPE), Miami, US, June 20 - 23, 2022.

Spérandio, M. & Espinosa, M.C. (2008). Modelling an aerobic submerged membrane bioreactor with ASM models on a large range of sludge retention time. Desalination, 231, 82–90.

Stare, A., Vrečko, D., Hvala, N., & Strmčnik, S. (2007). Comparison of control strategies for nitrogen removal in an activated sludge process in terms of operating costs: a simulation study. Water research, 41(9), 2004-2014.

Stentoft, P. A., Guericke, D., Munk-Nielsen, T., Mikkelsen, P. S., Madsen, H., Vezzaro, L., & Møller, J. K. (2019). Model predictive control of stochastic wastewater treatment process for smart power, cost-effective aeration. Ifac-papersonline, 52(1), 622-627.

Stentoft, P.A., (2020). Stochastic Modelling and Predictive Control of Wastewater Treatment Processes. PhD Thesis, Technical University of Denmark

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

Sweeney M. & Kabouris J. (2013). Modeling, instrumentation, automation, and optimization of wastewater treatment facilities. WATER Environ. Res. 85, 1322–1338.

Sweetapple, C., Fu, G., & Butler, D. (2014). Multi-objective optimisation of wastewater treatment plant control to reduce greenhouse gas emissions. Water Research, 55, 52-62.

te Braake H.A.B., Babuska R. & van Can H.J.L. (1994). Fuzzy and neural models in predictive control. Journal A (Special Issue on Model Predictive Control), 35(3), 44-51.

Talvitie, J., Mikola, A., Koistinen, A., & Setälä, O. (2017). Solutions to microplastic pollution–Removal of microplastics from wastewater effluent with advanced wastewater treatment technologies. Water research, 123, 401-407.

Therrien Jd, Nicolai N. & Vanrolleghem P.A. (2020). A critical review of the data pipeline: How wastewater system operation flows from data to intelligence. Water Science and Technology, 82 (12), 2613-2634.

Tinelli, S., & Juran, I. (2019). Artificial intelligence-based monitoring system of water quality parameters for early detection of non-specific bio-contamination in water distribution systems. Water Supply, 19(6), 1785-1792.

Treilles, R., Gasperi, J., Gallard, A., Saad, M., Dris, R., Partibane, C., Breton J. & Tassin, B. (2021). Microplastics and microfibers in urban runoff from a suburban catchment of Greater Paris. Environmental Pollution, 287, 117352.

Torfs E., Nicolai N., Daneshgar S., Copp J.B., Haimi H., Ikumi D., Johnson B., Plosz B.B., Snowling S., Townley Lr, Valverde-Perez B., Vanrolleghem P.A., Vezzaro L. & Nopens I. (2022). The

transition from WRRF models to digital twin applications. Water Science and Technology, 85(10), 2840-2853.
Venkatesh, G., & Brattebø, H. (2011). Energy consumption, costs and environmental impacts for urban water cycle services: Case study of Oslo (Norway). Energy, 36(2), 792-800.
Van Ekeren, H., Negenborn, R. R., van Overloop, P. J., & De Schutter, B. (2011). Hybrid model predictive control using time-instant optimization for the Rhine-Meuse delta. In 2011 International Conference on Networking, Sensing and Control (pp. 215-220). IEEE.
Van Nieuwenhuijzen, A. F., van der Graaf, J. H. J. M., Kampschreur, M. J., & Mels, A. R. (2004). Particle related fractionation and characterisation of municipal wastewater. Water Science and Technology, 50(12), 125-132.
Vanrolleghem, P.A. (1995, June). Model-based control of wastewater treatment plants. In Proceedings ESF Workshop Integrated Environmental Bioprocess Design. Obernai, France.
Vanrolleghem, P. A., Insel, G., Petersen, B., Sin, G., De Pauw, D., Nopens, I., Dovermann, H., Weijers, S. & Gernaey, K. (2003). A comprehensive model calibration procedure for activated sludge models. In WEFTEC 2003 (pp. 210-237). Water Environment Federation.
Vanrolleghem P.A., Benedetti L. & Meirlaen J. (2005). Modeling and real-time control of the integrated urban wastewater system. Environmental Modeling & Software, 20(4), 427-442.
Valverde-Pérez, B., Johnson, B., Wärff, C., Lumley, D., Torfs, E., Nopens, I., & Townley, L. (2021). Digital Water-Operational digital twins in the urban water sector: case studies. International Water Association, London, UK, White paper.
Vigne E. (2007). Etude et modélisation dynamique d'un procédé par biofiltration en nitrification tertiaire. Thèse de doctorat, Université Laval, Québec.
Vigne, E., Choubert, J. M., Canler, J. P., Héduit, A., Sorensen, K., & Lessard, P. (2010). A biofiltration model for tertiary nitrification of municipal wastewaters. Water research, 44(15), 4399-4410.
Villez, K., Billeter, J. & Bonvin, D. (2019). Incremental parameter estimation under Rank-Deficient measurement conditions. Processes.
Villez, K., Vanrolleghem, P. A., & Corominas, L. (2020). A general-purpose method for pareto optimal placement of flow rate and concentration sensors in networked systems–with application to wastewater treatment plants. Computers & Chemical Engineering, 139, 106880.
Viotti, P., Eramo, B., Boni, M. R., Carucci, A., Leccese, M., & Sbaffoni, S. (2002). Development and calibration of a mathematical model for the simulation of the biofiltration process. Advances in Environmental Research, 7(1), 11-33.
Von Stosch M., Oliveira R., Peres J. & De Azevedo S.F. (2014). Hybrid semi-parametric modeling in process systems engineering: Past, present and future. Computers & Chemical Engineering, 60, 86-101.
Wade, M. J., Sánchez, A. & Katebi, M. R. (2005). On real-time control and process monitoring of wastewater treatment plants: real-time process monitoring. Trans. Inst. Meas. Control 27, 173–193. London, UK.
Wágner, D.S.; Valverde-Pérez, B.; Sæbø, M.; Bregua de la Sotilla, M.; van Wagenen, J.; Smets, B.F. & Plósz, B.G. (2016). Towards a consensus-based biokinetic model for green microalgae-The ASM-A. Water Res., 103, 485–499.
Wai, K. P., Chia, M. Y., Koo, C. H., Huang, Y. F., & Chong, W. C. (2022). Applications of deep learning in water quality management: A state-of-the-art review. Journal of Hydrology, 128332.

Wang, X., Kvaal, K. & Ratnaweera, H. (2019). Explicit and interpretable nonlinear soft sensor models for influent surveillance at a full-scale wastewater treatment plant. J. Process Control 77, 1–6.

Wanner, O., Eberl, H. J., Morgenroth, E., Noguera, D. R., Picioreanu, C., Rittmann, B. E., & van Loosdrecht, M. C. M. (2006). Mathematical Modeling of Biofilms. IWA Publishing, London, UK.

Weijers, S. R. & Vanrolleghem, P. A. (1997). A procedure for selecting best identifiable parameters in calibrating activated sludge model No.1 to full-scale plant data. Wat. Sci. Technol. 36(5), 69–79.

Wentzel, M. C., & Ekama, G. A. (1997). Principles in the modelling of biological wastewater treatment plants. In Microbial Community Analysis: The Key to the Design of Biological Wastewater Treatment Systems (Vol. 5, pp. 73-82). IWA Scientific and technical Report nr. 5.

Wett B., Buchauer K. & Fimml C. (2007). Energy self-sufficiency as a feasible concept for wastewater treatment systems. In: Paper Presented at the IWA Leading Edge Technology Conference, 21 September 2007, Singapore, pp. 21–24.

Wilderer, P.A. & Characklis, W.G. (1989). Structure and function of biofilms. In Structure and function of biofilms (eds. W.G. Characklis & P.A. Wilderer), 5-17, John Wiley & Sons, New York, USA.

Xu, P., Janex, M. L., Savoye, P., Cockx, A., & Lazarova, V. (2002). Wastewater disinfection by ozone: main parameters for process design. Water research, 36(4), 1043-1055.

Yang, D., Karimi, H. R., Kaynak, O., & Yin, S. (2021). Developments of digital twin technologies in industrial, smart city and healthcare sectors: a survey. Complex Engineering Systems, 1(1), 3.

Yoo, C. K., Lee, J. M., Lee, I. B., & Vanrolleghem, P. A. (2004). Dynamic monitoring system for full-scale wastewater treatment plants. Water Science and Technology, 50(11), 163-171.

Yoo, C. K., Villez, K., Van Hulle, S. W., & Vanrolleghem, P. A. (2008). Enhanced process monitoring for wastewater treatment systems. Environmetrics: The official journal of the International Environmetrics Society, 19(6), 602-617.

Zhang, B. & Zhang, W., (2006). Adaptive predictive functional control of a class of nonlinear systems. ISA Trans. 45, 175-183.

Zhang, D., Dong, X., Zeng, S., Wang, X., Gong, D., & Mo, L. (2023). Wastewater reuse and energy saving require a more decentralized urban wastewater system? Evidence from multi-objective optimal design at the city scale. Water Research, 235, 119923.

Zhao, L., Dai, T., Qiao, Z., Sun, P., Hao, J., & Yang, Y. (2020). Application of artificial intelligence to wastewater treatment: A bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse. Process Safety and Environmental Protection, 133, 169-182.

Zhu, J.-J., Kang, L. & Anderson, P. R. (2018a). Predicting influent biochemical oxygen demand: balancing energy demand and risk management. Water Res. 128, 304–313.

Zhu, J.-J. & Anderson, P. R. (2019). Performance evaluation of the ISMLR package for predicting the next day's influent wastewater flowrate at Kirie WRP. Water Sci. Technol. 80, 695–706.

Zhu J., Bernier J., Pauss A., Vanrolleghem P.A. & Rocher V. (2018b). Modeling of the Seine Aval station – Towards real-time optimization of operating and environmental costs. In: Proceedings Water Information Days, Poitiers, France, 11 October 2018.

Zhu, J. (2020). Modélisation détaillée du fonctionnement de la filière complète de biofiltration de la station de traitement des eaux usées Seine Aval (PhD Thesis). Université de Technologie de Compiègne, Sorbonne Université, France.

List of Annexes

Annex I: Municipal Wastewater Treatment

1. Development of Wastewater Treatment

Human activities, especially in concentrated urbanized areas in well-developed regions, create large quantities of wastewater that have to be treated prior to their discharge in the natural environment. If not treated properly, the pollutant and nutrients load of urban wastewater, such as organic, nitrogen and phosphorus compounds, will cause serious environmental deterioration within the recipient environment, with a direct impact on aquatic biotopes, fisheries and human health.

The decomposition of organic matter will lead to nuisance conditions, including gas production, and the pathogenic microorganisms found in wastewater can be harmful to other life forms. The nutrients in the wastewater will stimulate the growth of aquatic plants that will swiftly reduce the dissolved oxygen content of the natural water, thereby limiting the conditions for fish and other aquatic life to exist. Wastewater treatment thus constitutes a major component of pollution control in preventing contamination of surface water and groundwater and is necessary to protect public health and the environment (Metcalf & Eddy, 2014).

The nature of the nutrients and pollutants, as well as the large volumes of sewage in which they are carried make treatment of municipal wastewater a challenging technical problem. Modern wastewater treatment started in the early 1900s when it became evident that disposal of sewage in a natural river was no longer sustainable. In earlier periods, the 'solution to pollution was often found in dilution', meaning that disposal of small quantities of waste into a river would not negatively impact that river's capacity to treat it, mainly due to the fact of dilution. In addition, the water from rivers was often used as a resource for consumption and industrial manufacturing processes. Due to the increased urbanization in the 19th and 20th century, extended sections of the rivers flowing through many urban centres became highly polluted.

The first treatment techniques constituted of collecting the sewage and transporting it to an area outside the city where it was applied directly onto the soil by intermittent irrigation and filtration. This method produces strong odours and required large sections of land that became less available with increasing city development. Other methods of treatment had to be developed that were accelerating the natural processes of decay and decomposition under controlled conditions in engineered facilities (Metcalf & Eddy, 2014).

The evolution of regulations in Europe and the US plays a key role in the innovation of wastewater treatment processes by defining water quality objectives. The first objectives were concerned with the removal of suspended and floatable materials and the treatment of biodegradable organic matter. Then, since the 1970s and 80s, the removal of constituents that may cause long-term health issues and environmental impact, such as pathogenic organisms, were included. More recently, the emphasis has shifted towards a recovery of resources and a neutrality in energy consumption as part of a circular economy with wastewater treatment plants (WWTP) transitioning to a water resources recovery facility (WRRF). Sustainability has become an important business motto, with a focus on the overall energy balance of a facility, its greenhouse gas emissions, the total chemical usage and the carbon footprint of 'production' of clean and safe water. All these elements have become of high importance in the design and operation of WWRFs.

These developments have resulted in complex and reactive municipal wastewater treatment processes that require a high degree of control in near real-time to guarantee effluent that meets the

water objectives and allows for recovery of resources, with energy efficient technologies that limit the environmental impact.

1.1 Wastewater Treatment Processes

Organized sewage transfer and collective treatment of wastewater started in the beginning of the 20th century. Since then, a wide range of technologies, processes and techniques have been developed and implemented for the treatment of domestic and industrial wastewaters.

The typical processes involved in wastewater treatment are preliminary treatment, primary treatment, secondary treatment and tertiary treatment. At the start, large floatable items, sand and greases are removed during preliminary treatment to prevent damages to the equipment of the following stages. Primary treatment focuses on the removal of the suspended solids and organic matter from the wastewater. This process can be further enhanced by chemical addition, innovative design, such as the insertion of inclined flaps in the upper section of a primary settling tank ('lamallated settler') or by filtration techniques. Secondary treatment is concerned with the removal of biodegradable organic matter and the remaining suspended solids. Often nutrient removal of nitrogen and phosphorus takes places at this stage. Finally, during tertiary treatment any residuals of suspended solids and nutrients are to be removed in order to achieve even lower levels of oxygen demand when discharged into a river. Disinfection often takes place at this final stage. During each of these processes settled biosolids, or sludge, is removed and further treated, such as by a digester, incineration, composting or directly applied to agricultural lands. When wastewater is to be reused, additional levels of treatment are required to remove any remaining dissolved and suspended materials, such as heavy metals and inorganic solids.

The basic layout of a wastewater treatment plant (WWTP) using an activated sludge system, which is one of the most commonly applied methods for the treatment of urban wastewater, is shown in **Erreur ! Source du renvoi introuvable.**. The treatment starts with a preliminary screening to stop f loating objects and a grit removal to take out gravel and sands, followed by primary treatment, which separates the suspended solids from the water through a physical gravitational settling process. The relatively long detention times within these low velocity tanks allow the heavier solids to settle to the bottom, while the lighter solids float to the surface. Mechanically rotating arms collect the sludge and solids on the bottom, as well as the grease and oils floating at the top, allowing for distribution to the sludge digester for further treatment.

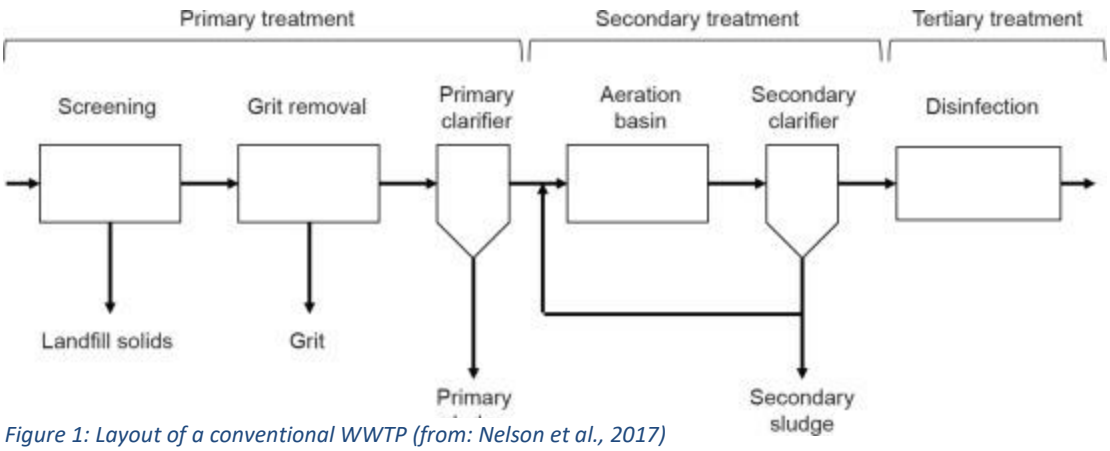


Figure 1: Layout of a conventional WWTP (from: Nelson et al., 2017)

Secondary treatment consists of treatment by microorganisms (mainly bacteria) that eliminate the organic and mineral compounds dissolved in the wastewater by using it for their growth. These processes are based on the natural functioning of bacteria to consume substrate in order to close the

elemental cycles of carbon (C), nitrogen (N) and phosphates (P). In a wastewater treatment plant, naturally occurring bacteria are used. By engineering the system, natural limitations for bioconversion such as limited aeration and limited amount of biomass can be overcome (Henze et al., 2008).

The different conditions or environments for bacteria to be active and consume nutrients are: a) aerobic (an aerated environment, with the presence of oxygen, most often added by forced aeration), b) anaerobic (an environment that has low concentrations of dissolved oxygen and no forced aeration), or c) anoxic, where there is a near total depletion of dissolved oxygen. Since molecular oxygen is a highly oxidizing agent, the aerobic processes are more efficient for bacterial growth than anaerobic processes where microorganisms are forced to use other less-oxidizing substances, such as available nitrates in the water.

Biological treatment of carbon is achieved by heterotrophic bacteria that are omnipresent in the water and are dependent on external energy sources (substrate), mainly the organic matter content of the water. They can be active under aerobic, anaerobic and anoxic conditions. The organic matter is transformed into carbon dioxide CO_2, water H_2O and more biomass that end up as biosolids.

Autotrophic bacteria that use carbon dioxide and the oxidation of inorganic compounds (such as ammonia) as their energy source achieve biological treatment of nitrogen in subsequent oxidation and reduction phases. The autotrophs create nitrates (nitrification) under aerobic conditions, followed by heterotroph bacteria under anoxic conditions that use the nitrates as their oxygen source to create nitrogen gas (denitrification).

Biological removal of phosphorus is achieved by promoting the growth of organisms collectively known as the polyphosphate accumulating organisms (PAOs) that take up and store excessive amounts of phosphorus. Since the bacterial removal rate for phosphates is not efficient within short time frames, additional treatment by chemical precipitation is often applied.

Effluent water quality standards are generally measured by the level of biodegradable organics, expressed in terms of five-day biochemical oxygen demand (BOD_5), total suspended solids (TSS), and the nitrogen (TKN, NH_4, NO_3, NO_2) and phosphorus (TP, PO_4) compounds. Since the biochemical oxygen demand (BOD_5) indicator requires a five-day measurement, it is often supplemented by the rapid chemical oxygen demand (COD) that is measurable in the matter of a few hours. Other indicators include the hydrogen ion concentration expressed as pH, the temperature and turbidity.

1.1.1 Biological Treatment

Biodegradabale components of wastewater can be treated biologically – that is by microorganisms; in order to oxidize dissolved and particulate carbonaceous BOD; to capture suspended and non-settleable colloidal solids into a biological floc or biofilm; and to transform nutrients, such as nitrogen and phosphorus (P), into end products that are less harmful to a large degree. A variety of microorganisms are used, principally bacteria, that not only convert organic matter by oxidation and nitrogen (by oxidation and reduction), but also create new cells that become additional biomass. For phosphorus removal, a specific group of bacteria are stimulated to take up and store large amount of inorganic P into the cells. The resulting bioflocs are heavier than water and can be removed by gravity settling (Metcalf & Eddy, 2014).

There are two main categories of biological wastewater treatment distinguished on the nature of the bacteria; either floating in suspension (‘suspended growth’) or attached to a support media (‘attached growth’). In suspended growth, the bacteria are held in liquid suspension requiring constant mixing of the bulk liquid. The activated sludge process for biodegradation of organic substances is the most commonly applied method for municipal wastewater treatment. It has been developed since the early

20th century, and requires large installations (footprint) and relative long hydraulic retention times. In attached growth processes however, the bacteria responsible for the conversion of organic matter and nutrients are attached to an inert packing material, and grow inside a thin layer called the biofilm. This process is therefore also known as a 'fixed film' process. The packing materials that constitute the filter bed can be made from rocks, gravel, sand, plastics and other synthetic materials. The filter bed can either be completely or partially submerged into the wastewater, with air or gas space above the biofilm liquid layer. Commonly applied examples of attached growth process is the trickling filter, where the wastewater is distributed on the top layer of the filter bed and allowed to infiltrate the packing material, and biological activated filtration ('biofiltration') (Metcalf & Eddy, 2014).

1.1.2 Biofilms

A biofilm is the thin layer within a typical range of 10 to 1000 µm (average thickness of 500 µm) attached to the surface of a support media within which the bacteria are active and transform different organic and inorganic substrates (see Figure 2a). It can be defined as a layer of prokaryotic and eukaryotic cells anchored to a substratum surface and embedded in an organic matrix of biological origin (Wilderer & Characklis, 1989). A biofilms is a gel-like substance composed of bacteria embedded in a matrix of extracellular polymeric substances (EPS) containing polysaccharide, proteins, free nucleic acids, and water (Sutherland, 2001). The solids give the biofilm its reactive and structural properties (Eberl & Morgenroth, 2006), while the EPS is basically the glue that holds the biofilm in place (Henze et al., 2008).

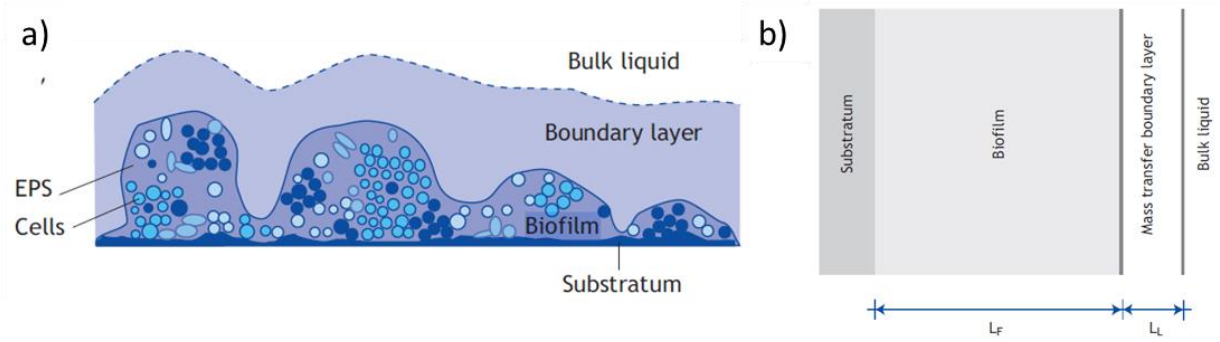


Figure 2: a) schematic representation of the different components of a biofilm system: bulk liquid, boundary layer, biofilm and substratum (from Henze et al.,2008 adapted from Wanner et al.,2006); b) The modelling approach with flat surfaces, a homogenous biofilm with thickness L_F and a boundary layer of homogenous thickness L_L. (Henze et al., 2008).

Bacteria growing in biofilms have an important advantage over bacteria in suspended growth, since the biofilm provides a higher level of protection against washing. Also, the geometry of a biofilm allows for distinct biological niches to occur – as long as there is food available in the form of substrate - that protect and stimulate the growth of 'weaker' bacteria that normally wouldn't survive in a homogenous system. Active biomass concentrations inside the biofilm are much larger compared to activated sludge systems (Henze et al., 2008).

The microbial activity inside the biofilm can furthermore modify the internal environmental conditions making it a more hospitable place for different bacteria to grow that affect microbial composition. For example, in a biofilm the slow-growing bacteria can procreate in the deeper layers where they are protected from detachment, while faster-growing microorganisms can establish themselves at the layer bordering with the bulk-liquid. Biofilm growth happens as a result of the production of microbial mass in the biofilm. Generally, at the start of a treatment, the fast growing heterotrophic organisms dominate. As the biofilm gets thicker, the slow-growing autotrophic bacteria become more abundant in the biofilm depth, while the heterotrophs still dominate at the top section near the surface. In addition, the inert mass of products formed by microbial decay starts to accumulate in the biofilm.

Whenever a detachment event occurs (such as a sloughing event or a through washing session applying large air bubbles and high pressured flow); a large part of the biofilm is removed. In the remainder biofilm, the autotrophs now dominate (Eberl & Morgenroth, 2006).

Essential features of a biofilm include the compartments, which define the sections of a biofilm (e.g. bulk liquid, biofilm and support media), and the particulate and dissolved components (e.g. biomass, substrates, products formed) contained within each compartment. Transformation, transport and transfer are the possible processes acting upon the components. Transformation processes are the biochemical reactions that produce or consume one or more components (e.g. utilization of substrate, production of biomass) and the abiotic chemical reactions that result in the adsorption or precipitation onto the biomass. Transport processes can happen inside the biofilm mainly by molecular diffusion of soluble components, as well as outside of the biofilm by advection and turbulent dispersion. The transport of particulate components can also happen through the expansion or contraction of the biofilm's solid matrix as a function of bacterial growth or decay and of the extracellular polymeric substances (EPS). Transfer processes exchange mass of dissolved or particulate components between two compartments (e.g. exchange of dissolved and particulates from the bulk liquid into the biofilm).

The principal compartments are the biofilm, which contains the biofilm itself; the bulk liquid that contains the wastewater rich in nutrients and pollutants, and the mass-transfer boundary layer (MTBL) that symbolizes a very thin zone of stagnant water on the external side of the biofilm that represent the mass transfer resistance of dissolved components into the biofilm.

The transfer of mass from the bulk liquid into the biofilm is an important process that allows for the substrate to move from the wastewater ('bulk liquid') to the biofilm so that it can be used by the microorganisms. The transfer is driven by a concentration gradient across the hypothetical mass-transfer boundary layer. For biofilms that have relative flat surfaces, the MTBL is assumed to be of uniform thickness (Eberl & Morgenroth, 2006).

1.2 Treatment by Biofiltration

Biofiltration, also called biological activated filtration (BAF), is an intensive treatment using microorganisms in compact bioreactors. This technology combines both physical and biological purification processes in a reactor using immersed filter media onto which the bacteria sustain themselves and break down organic materials and nutrients. It employs a dual function of biological oxidation of BOD and/or ammonia soluble components by diffusion and biodegradation kinetics in the biofilm, and physical removal of particulate and colloidal solids by adsorption and filtration (Metcalf & Eddy, 2014).

The term biological activated filtration includes both the aerated attached growth process ('Biological Aerated Filter') for removal of organic matter and ammonia, and the filtration under anoxic conditions for denitrification. The earliest development of a BAF started in the late 1970's with experiments of upward and downward flows of the wastewater through a tank filled with granular material. The design by Leglise et al. (1980) incorporated a down flow reactor with granular media that used activated carbon. The first commercial installation of this type ('Biocarbone©') took place in 1980 at a WWTP near Paris. Since then, several hundred BAF's have been installed worldwide, although principally in Western Europe (Metcalf & Eddy, 2014).

The main advantages of BAF's when compared to the more commonly applied suspended growth processes are the relative small footprint of the installation that allows for it to be totally covered inside a structure (less odours and visual pollution), the ability to treat diluted wastewaters and the absence of a requirement to settle the biomass. However, the process is more complex from an

operational and maintenance point of view, has a high capital cost for installation and operation, and it is vulnerable for blocking with solids and accumulated biomass, requiring frequent backwashing to maintain high performance (Metcalf & Eddy, 2014).

Biofiltration is a practical solution to regions with strong land pressures (such as in urbanized areas) due to the development of a concentrated biomass of heterotrophic and autotrophic bacteria on the filtering media. Biofilter reactors are generally constructed of columns of 2 to 4 meters height filled with granulated materials through which the wastewater flows, as shown in Figure 3. Commonly used materials for the filter bed or biolite are beads of clay, polystyrene, sand, pozzolan rocks or active carbon. The beads have a diameter size of between 3,5 and 4,5 mm. The small diameter results in a large total surface area for the bacteria to colonize, while still keeping a small footprint of the reactor (Metcalf & Eddy, 2014).

The inert filter material used in the BAF can be mineral, plastic or synthetic. Water flow can be ascending or descending; the first case is commonly preferred in municipal plants due to the ease of controlling the regular backwashing needed to remove excess biomass, as well as to prevent foul odours from escaping (Rocher, 2012a).

The removal of pollution is achieved by physical filtration, due to the small open spaces between the beads, larger particles are filtered; and by biological activity. The microorganisms form a thin layer at the surface of the beads, the biofilm, in which the biological reactions are taking place. The rate of biodegradation is affected by the specific kinetics of the type of active bacteria. In addition, soluble substrate removal is a diffusion-limited process with removal rates affected by the biofilm area, bulk liquid substrate and dissolved oxygen concentrations, fluid velocity in upwards or downwards direction and the water temperature. The particulate removal in a filter is affected by the media size, fluid velocity, operating headloss and backwash frequency. Under aerobic conditions, the bacteria remove carbon and produce nitrates (nitrification), while in an anaerobic environment the microorganisms transform nitrates to produce nitrogen gas (denitrification).

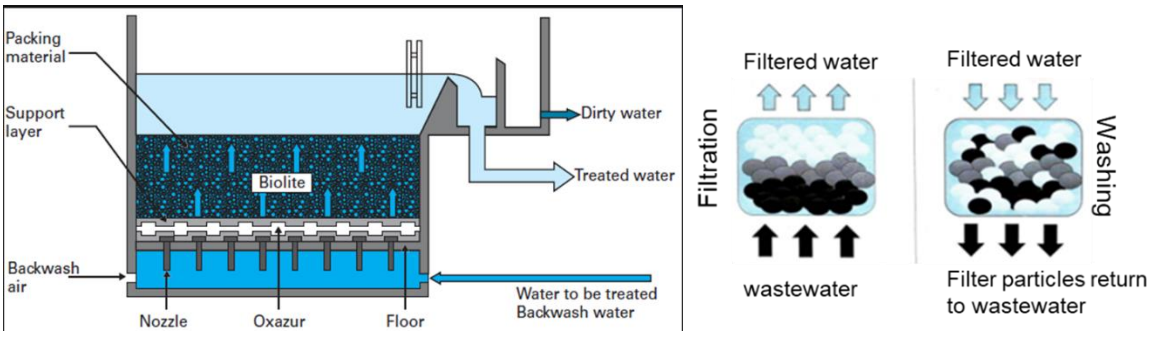


Figure 3: a) Schematic description of a submerged biofilter with a upwards flow direction; b) schematic representation of the accumulation of biomass at the entrance of the filter (left) and the effect of the backwashing with a redistribution of the beads in the biolite (right).

Some limitations are the risks of blockage of the filters due to the accumulation of suspended matter and swollen biofilms, requiring frequent reversing of the water flow to wash of the filters. Biofiltration is reported to endure higher costs for investment, maintenance and operation when compared to more conventional technologies, particularly due to significant energy needs for aeration and water pumping, as well as the utilization of reagents (Rocher et al., 2012a).

An important feature of treatment by biofiltration is the requirement for regular backwashing, which not only prevents suspended solids from clogging the filter, but also directly affects the thickness of the accumulated biomass and the microbial composition of the biofilm. In addition to scheduled

washing events, at certain moments a sloughing event can occur that detaches a larger part of the biofilm from the surface (Eberl & Morgenroth, 2006).

1.3 BAF Process Design and Operational Performance

Important factors that affect design and performance of BAF processes are the media characteristics (e.g. diameter of the beads), process loadings, aeration design and sludge production. The configuration of biofilters placed in serial that each have different conditions for typical bacterial growth can promote performance further.

1.3.1 Media size

The inert support media upon which the bacteria will accumulate affects both the biological and physical functions of a biofilter. For most common applications, spherical beads are used. The smaller the diameter of the beads, the larger will be the surface area available per unit of reactor volume for biofilm growth and substrate removal rates. Besides, the diameter also directly affects the pore size and filtration capacity of the filter.

The selection of best media size in the design of BAF is based on the loading of suspended solids and organic matter that are entering the filter, as well as the desired filtration and organic removal efficiencies. Requirements for regular backwashing is also a consideration, since during backwashing the treatment is basically halted resulting in operational 'down-time'.

In general, a smaller media size allows to treat 'heavier' wastewaters that contain high organic matter concentrations due to the higher specific surface area available. At the same time, a filter bed with small diameter beads will experience a reduction of the filtration capacity sooner than large diameter materials. If not timely removed by backwashing; a small diameter size filterbed can be more easily blocked; it thus significantly reduces the time between two sequential washing sessions.

In literature, media sizes of 4 to 5 mm are recommended for secondary treatment and a size of 6 mm for partial organic matter removal. For tertiary treatment, a smaller diameter size of 3 mm is more effective (Mendoza-Espinosa & Stephenson, 1999).

1.3.2 Process Loadings

The design of BAF's are concentrated on the volumetric process loading rates, which is expressed as the mass of BOD or nitrogen removed per day in function of the media bed volume applied, expressed as (kg BOD or N/(m^3*d); and the hydraulic application rate (m/h), defined as the wastewater flow rate (m^3/h) applied to the surface area of the filterbed (m^2). Table 1**Erreur ! Source du renvoi introuvable.** i ndicates a typical range of loading units for biological aerated filters with different applications together with their expected removal efficiencies, as presented by Metcalf and Eddy (2014).

Table 01: Process volumetric loadings for biological aerated filters (Metcalf & Eddy, 2014)

Process application	Loading units	Range	Removal efficiency (%)
BOD removal	kg BOD / (m^3 * d)	3.5 – 5.5	≥ 85
BOD removal and nitrification	kg BOD / (m^3 * d)	1.8 – 2.5	≥ 85
Tertiary nitrification	kg NH_4-N / (m^3 * d)	1.0 – 1.5	≥ 90

A nitrate-N loading of 1.0 to 1.5 kg/(m^3*d) and a hydraulic application rate of 20 to 30 m/h has been reported for a preanoxic upflow BAF (Ninassi et al., 1998).

Higher volumetric loads applied to a filterbed will result in more biological activity and thus an improved effluent water quality. However, it also entails a faster growth and accumulation of biomass that can lead to lowering of the filtration efficiency and eventually a blocking of the filter. For BAFs in secondary treatment, the BOD and TSS loading rates are important process parameters that affect performances. In particular, the applied TSS load should be strictly monitored to remain low in order to avoid filter blockage (Rocher et al., 2008).

With tertiary treatment, where the organic matter and TSS loads are already very low, the volumetric nitrogen load is the key process control parameter. However, if this becomes too low for a prolonged period, the bacteria in the biofilm experience a lack of substrate and start to decay faster. Values of 4 to 5 $kgNO_3$-N/(m3*d) are reported to result in a 90% removal efficiency in denitrification filters, but this becomes impossible to attain when the loading rates drop to 3 $kgNO_3$-N/(m3*d) (Rocher et al., 2011).

1.3.3 Aeration

The oxygen demand in an aerated BAF is a function of the wastewater characteristics, the BOD loading and backwashing frequency. Aerated biofilters have a volumetric oxygen uptake rate (OUR) that is much larger compared to the activated sludge systems due to the short hydraulic retention time of the wastewater in the filter. Values of up to 250 mg/(L*h) have been reported (Metcalf and Eddy, 2014).

Due to the smaller media size applied in BAFs, air bubbles are hold up in the filterbed this increasing the oxygen transfer efficiency. In BAF's the aeration application can be effective with larger bubble sizes (Metcalf and Eddy, 2014).

1.3.4 Biofilter Configuration

The installation of aeration nozzles inside a biofilter tank permit to create the different types of growth conditions for heterotrophs and autotrophs that allow to transform NH_4 into NO_3 (a process called nitrification) and then the transformation of NO_3 into N_2 (denitrification). As such, biofilters can be applied for carbon and nitrogen removal and their configuration has a significant impact upon the operational costs.

As thoroughly explained by Rocher et al (2012a), nitrogen removal with submerged upwards-flow biofilters can be achieved using one of three different layouts named after the location of the denitrifying reactor: 1) upstream denitrification (UD), 2) downstream denitrification (DD) and 3) combined upstream-downstream denitrification (U-DD) as shown in Figure 4 (Rocher et al., 2012b).

In the DD layout the effluent flows through the successive treatment stages without a need for recirculation since the denitrification is placed following the nitrification stage. In the first stage, carbon is consumed. At the second stage, the limitation of organic carbon allows the growth of a nitrifying biomass that converts ammonia into nitrates. At the third stage, both anoxic conditions and the addition of organic carbon (not indicated in the figure) promote the growth of denitrifying bacteria (Henze et al. 2000). In the UD layout, the biological treatment of the effluent is carried out over two stages; first denitrification and carbon removal combined, followed by the nitrification process. In this configuration, recirculation by pumping of (a part of) the nitrified water towards the inlet of the first stage is required. The advantage here is however that the organic content of the settled effluent is a sufficient source of energy and no external carbon source is required. In the U-DD configuration, both upstream and downstream denitrifications are combined. The first two stages are similar to the UD mode, but a third stage is for a post-denitrification with methanol addition. This allows stabilizing the recirculation ratio and create a more robust system, even during high load demands.

Research showed that economic savings of 10% can be achieved combining upstream and downstream denitrification when compared with downstream denitrification. Despite the extra energy consumption due to the pumping of recirculated water, savings induced by the restricted injections of methanol and air remained substantial (Rocher et al, 2012b).

In each of the configurations used, the comparison studies of Rocher (2012a and b) showed that the carbon pollution was almost fully removed and that the concentrations of organic matter in the effluent were residual. Likewise, ammonia was oxidized to over 97% upon biofiltration, the ammonium concentrations in the effluent not exceeding 1 mgN/L. The oxidized nitrogen, resulting from nitrification, was reduced to atmospheric nitrogen (N_2) during denitrification. This reduction was efficient enough in the UD and the U-DD layouts to comply with the regulatory thresholds.

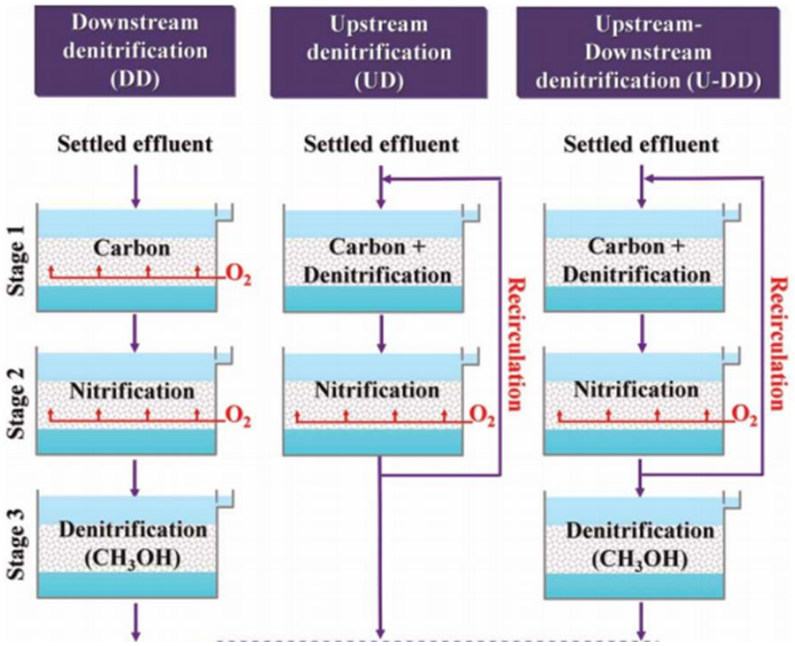


Figure 4: Three possible configurations to achieve nitrogen and carbon removal by the biofiltration with downstream denitrification (DD), upstream denitrification (UD) and combined upstream-downstream denitrification (U-DD) (Rocher et al., 2012a).

Rother and Cornel (2007) report that the combination of a pre-DN, N and post-DN filters is much more advisable for most municipal wastewaters, because the recycle rate can be reduced and external carbon can be saved. For minimum use of external carbon, 100–150% recycle rate should not be exceeded. Then, approximately 50–60% of the total NO_3-N can be depleted in the pre-DN stage. On average, 10 g total (t) COD/g NO_3-N were required in the pre-DN stage for denitrification in the pilot and full-scale plant and 0.4–0.5 kg NO_3-N/(m^3*d) can be reached without external carbon. As only 40–70% of the COD load is eliminated in the pre-DN, the remaining COD load is removed in the nitrification stage. 1 kg COD/(m^3*d) suppresses nitrification rates by approximately 0.1 kg NH_4-N/(m^3*d). For nitrification rates, >0.5 kg NH_4-N/(m^{3*} d) at 12°C not more than 2 kg COD/(m^3 d) may be eliminated in the nitrification (Rother and Cornel, 2007).

Annex II: Smart Monitoring and Process Modelling of Wastewater Treatment in Paris

M. Serrao[1,2,3,4], J. Bernier[1], V. Rocher[1], I. Juran[2], B. Tassin[3], P. Vanrolleghem[4]
[1]SIAAP – Syndicat interdépartemental pour l'assainissement de l'agglomération parisienne, Département d'innovation, Colombes, France;
[2]W-SMART International Association of Water & Wastewater Utilities for Sustainable Water Security, Paris, France;
[3]Leesu, Ecole des Ponts, Université Paris Est Créteil, Marne-la-Vallée, France;
[4]model*EAU* - Département de génie civil et de génie des eaux, Université Laval, Québec, QC, Canada;

2.1 Introduction

SIAAP is the public utility responsible for the management of wastewater as well as storm water in the Greater Paris region with 9 million inhabitants covering an area of 1 800 km^2, as illustrated in figure 1. With the transport and treatment of nearly 2,5 million m3 of water per day of municipal and industrial wastewater (respectively 94% and 6% of dry weather flow), SIAAP plays an important role in the protection of public health and ecological status of the Seine and Marne rivers. It has been implementing a strategy that promotes the integration of smart systems for the control and management of operational processes. One of the principal objectives of this organizational-wide mobilization towards becoming a 'smart wastewater utility' is to improve the stability and reliability of the physical, biological and chemical processes required for treatment of water and sludge as well as the recovery of resources. The application of such advanced process control tools permits to advance optimization and efficiency, to increase the reuse of materials and to move towards energy neutrality as part of a circular economy. As a support tool, smart systems allow visualizing the effects of recommended control strategies and provide indications of the expected economic and ecological benefits. This allows building confidence in decision making by operators and managers alike with the determination to reconcile the evolving stricter regulatory effluent quality requirements with energy conservation, the recovery of resources and the reduction of environmental impact.

Figure 13: SIAAP's service area map with indication of the wastewater treatment plants and an overview of the design of the Seine-Center WWTP integrated into the urban landscape. (© SIAAP)

This Smart Water vision forms part of the operator's sustainability program 'SIAAP 2030'. The plan formulates concrete actions and timeframes to mitigate the effects of population growth and urbanization (which increase soil impermeability), the emergence of new types of pollution linked to altering consumption styles and the effects of climate change (such as increased water scarcity and changing rainfall patterns). However, overall, it formulates the desired transition towards environmental and economic sustainability (e.g. lowering GHG emissions, increasing nutrient recovery and energy revalorization) while increasing operational efficiency by 'going smart'. Furthermore, the plan calls for communication among stakeholders and hopes to contribute to a smart city concept by inspiring other organizations active in the urban and natural water cycles (SIAAP, 2017b).

In the last few decades, there have been experiences gained and lessons learned with the application of smart infrastructure in the sanitation sector (e.g. Ingildsen and Olsson, 2016). This is observed in particular in the fields of in-line monitoring of water flow and of qualitative process performance indicators used by automatic controllers (e.g. Sharma et al, 2011), as well as the development of computer aided mathematical models for simulation and scenario analysis (e.g. Benedetti et al., 2013;. This has resulted in more stable and reliable process control, increased regulatory compliance and overall increase in our understanding of the hydraulic, chemical and biological processes taking place during the transport and treatment of municipal wastewater.

Like many other industries, the municipal wastewater treatment sector has been experiencing a sharp increase in the application of instrumentation, control and automation (ICA) in the last decades (e.g. Ingildsen and Olsson, 2016; Garrido-Baserba et al., 2020). The ICA technologies were introduced in the 1970's in order to control operations and assure good quality of the treated water in order not to harm the quality of the natural surface water bodies into which the effluent is discharged. Wastewater treatment has some specific characteristics when compared to other process industries, such as highly variable flow rates and pollutants loads that need to be removed to very low values under all circumstances. Wastewater processing consists of several successive physical, biological and chemical processes for both water and the settled sludge that exhibit non-linear and non-stationary behavior. This requires accurate process control to assure overall performance and compliance with regulatory norms and quality standards.

The operator's strategy supporting the digital transition is based on a multidisciplinary approach of smart monitoring and data-based information management systems, on science-based modelling of processes, and the integration of Artificial Intelligence techniques. It involves the participation of various organizational layers, on operational, administrative and management levels (SIAAP, 2017b. Implementation of smart water quality measurement systems allows capturing the process dynamics and has increased knowledge and understanding of treatment processes. The integration of theory-based and data-based modelling has the potential for improved process control, for prevention of down-time and preemptive risk mitigation, for detection of anomalies in process variables (e.g. pollutants loads and flow rates) and the building of urban resilience in a city-wide cooperative approach. The digital revolution in the water sector, part of the so-called fourth industrial revolution, utilizes the strengths of digital data management together with process control and optimization tools designed for use in the physical plant. It promises to carry the sanitation sector to the level of sustainable management of an Integrated Urban Water System (IUWS) (Garrido-Baserba et al., 2020). Nevertheless, the acceptance and incorporation of Smart Water Systems in the daily operations and management of large size centralized municipal sewage plants still faces many challenges.

The implementation of a digital transformation is leading to increasing amounts of generated data that has to be managed for it to become useful information. This has required SIAAP to further invest in ICT and regularly updated management systems, such as Laboratory Information Management Systems (LIMS), Maintenance Management Systems (MMS) and next generation Supervisory Control and Data Acquisition (SCADA) systems (see figure 2). In a recent international survey by Blumensaat et al. (2019) among wastewater utilities, many organizations report that they discover opportunities in the digitalization of treatment processes, as well as challenges in understanding exactly in what manner digitalization allows solving technological issues. Additionally, the rapid increase in digital data collection in the wastewater sector is reported to cause complex technological issues, for example with interconnectivity of various information systems and cybersecurity. It also imposes organizational and cultural changes to allow for cross-disciplinary training, and interdisciplinary research in areas of integrated urban water management and beyond (Blumensaat *et al*., 2019.)

This chapter describes the experiences and lessons learned with the development and the integration of smart wastewater monitoring, process control and modelling systems by SIAAP, the public wastewater utility in the French city of Paris and its surrounding municipalities.

Figure 2: the central control room managing flow via the 440 km sewer network towards the wastewater treatment plants and storm water retention basins. (© SIAAP)

2.2 Drivers for Innovation

The evolution of regulatory requirements in France has played a key role in the development and deployment of innovative measuring methods and technological treatment solutions, in particular due to the European Urban Waste Water Directive in 1991 (91/271/EEC) and the European Water Framework Directive (2000/60/EC) in 2000, which have been enacted through appropriate national legislation. These efforts have changed the design and operations of wastewater treatment with new or upgraded systems resulting in improved plant-performance, service reliability and public and environmental health. A great testimony is given by the water quality of the rivers Seine and Marne flowing through the metropolitan area of the French capital. Since the 2000's, a drastic decrease of organic carbon, phosphorus and ammonium concentrations and the return of an enriched biodiversity have been observed, which has resulted in a good ecological status (Rocher et al. 2017).

The wastewater sector has been searching for ways to optimize the treatment process of water and sludge in order to reduce its energy consumption and usage of chemical reagents, while meeting the regulatory requirements. The principal arguments are to lower the environmental impact of operations (e.g. energy usage for pumps), to meet the evolving stricter environmental discharge limits of treated effluent, including emerging micropollutants, and to lower the dependence on fossil fuels as an energy source. This has resulted in the development of innovative technologies and adapted treatment system designs to guarantee compliance to management standards and regulatory norms. These innovative treatment technologies, such as chemically enhanced settlers, membrane reactors and biological active filtration systems, contain wide range adaptive capacities with shorter treatment times. Accordingly, they require a smaller spatial footprint, which is an important feature for municipal facilities in densely urbanized settings. At the same time, the degrees of control have increased due to operational complexity, such as through the introduction of varying flow regulation and recirculation, the injection of air and the addition of chemical reagents as coagulants and flocculants.

Another important driver for implementing smart water management is the reduction of environmental impact of operations by reducing the emissions of greenhouse gasses, such as CO_2 and N_2O, lowering the energy consumption and limiting the usage of chemical reagents for treatment.

The current approach of compliance management however does not consider the dynamic conditions of the receiving natural environment. Presently, French environmental policy is defined on national, regional and local administrative levels by governmental regulatory bodies that dictate effluent discharge limits for main pollutants and nutrients as maximum concentrations in the (processed) water leaving a treatment facility. In general, these discharge limits are rigid/fixed values for one or more variables, such as total suspended solids (TSS), organic matter (COD, BOD5), nitrogen (N) and phosphorus (P) concentrations and daily, monthly and yearly averages for treatment performances under predefined conditions (e.g. 90% of daily measured effluent N-concentrations not exceeding the maximum allowable concentration on a yearly basis for those days during which the influent flow rates do not exceed the maximum capacity of a treatment plant).

The requirements of European environmental policy and in particular the assessment of the municipal wastewater sector is facilitating a shift towards a holistic approach placing emphasis upon the overall ecological quality of the receiving surface waterbody (e.g. a river or lake) into which the treated effluent is discharged ('integrated river basin management'). The water quality of the natural receiving river will become the principal

indicator for the performance of urban water management, including wastewater treatment. These developments require an in-depth understanding of the interactions between urban water management and natural water systems and call for an adaptive, smart and holistic approach of the urban water system (Corominas et al., 2013).

2.3 Research and Implementation of Smart Water Systems

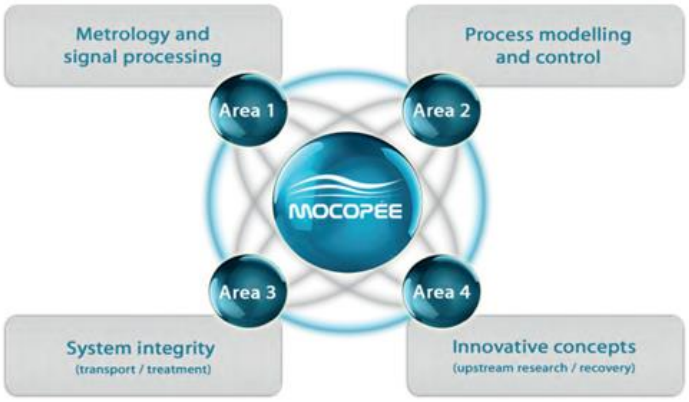
The prominent role of research and innovation in the Paris Wastewater Utility SIAAP goes back to the beginning of the 20th century with early field experiments on biological filtration and activated sludge systems. The creation of a Research Centre in 1980 has spurred a broader approach to innovation through dedicated programs in collaboration with the water industry and academic world, such as through the PIREN-Seine for research on the level of the river basin and the OPUR program, which is centered on pollution control in an urban context. Further examples of applied research programs are the Mocopée program for modeling and optimization of treatment processes at plant-level, and the MeSeine focusing on surface water monitoring. Partnerships and collaboration with public organizations in charge of water and waste management, as well as with the private sector, allow developing synergies and anticipating the technological evolution in the sanitation industry (SIAAP, 2020).

The 'SIAAP 2030' strategic plan provides the framework for the development of an environmental and economically sustainable and smart wastewater management plan. Launched in 2016, it calls for increased efficiency and optimization of treatment processes, for an augmented recovery of nutrients and energy-generation, for a transition to water management in the context of a smart, resilient and sustainable city, and for synergies with the domains of waste disposal and energy production. Importantly, it promotes research and innovation to develop new technologies and alternative treatment concepts (SIAAP, 2017b).

Figure 3: The 2018 Mocopée conference in 'La Cité de l'Eau et de l'Assainissement' in Paris. (© SIAAP Franck BELONCLE)

The Mocopée (*MOdelling, Control and Optimization of wastewater treatment ProcEssEs;* in French: *Modélisation, Contrôle et Optimisation des Procédés d'Épuration des Eaux],*) program offers operational and financial support for scientific research focused on the modelling, control and optimization of municipal wastewater treatment processes. This applied research program is managed by SIAAP and includes many partnerships with scientific institutions, industries, regulatory bodies and other stakeholders in wastewater management in the region. It actively seeks collaboration in order to increase knowledge-sharing (see figure 3). In its current phase (2018 – 2022), the focus lies on smart solutions for process control and optimization within the context of a circular economy and sustainable management of resources and asset management. It is organized around four key areas, as indicated in figure 4.

Figure 4: Focus areas of the Mocopée research program 2018 – 2022. (© SIAAP)



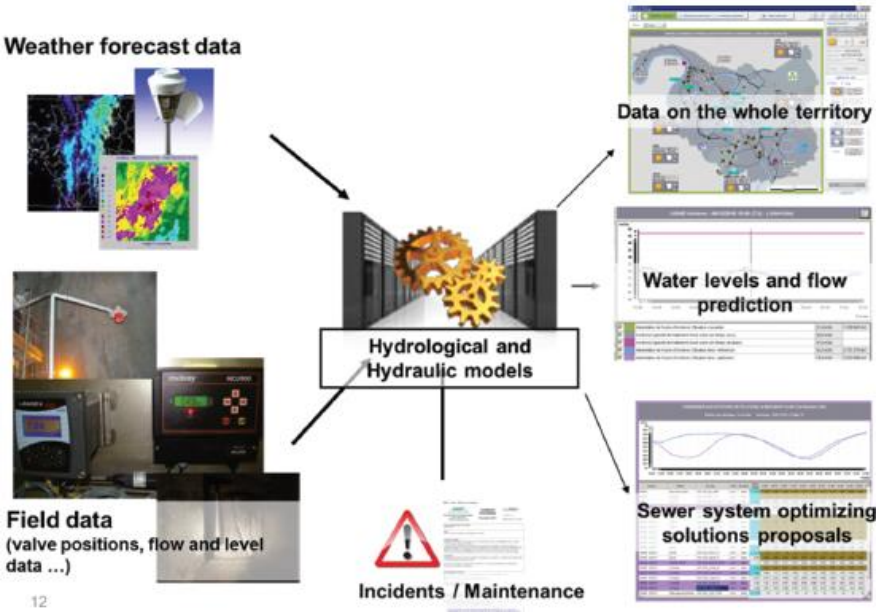
The program serves as a platform for scientists and operational staff to perform studies on practical questions encountered, in particular related to monitoring, process control and modelling. It functions as an exploratory ground for the development of new concepts and innovative solutions with multidisciplinary teams. It enables to experiment with new ideas outside the strict time constraints of traditional funding programs and permits testing solutions in large scale settings where technicians and operators provide feedback in an early phase (SIAAP, 2018).

2.4 Smart Flow Control and Monitoring in the Sewer System

Thanks to a supportive strategy of modernization and innovation, SIAAP operators rely on advanced monitoring and data processing systems for their day-to-day operations. The introduction of information and communication technologies in the 1970's has given way to advanced supervisory communication and data management systems. The daily transport of 2,5 million m^3 of collected wastewater through a network of interconnected main sewer lines with a total length of 440 km and its allocation to each of the six treatment stations is supported by a data-based management decision system named MAGES (*Modèle d'aide à la gestion des effluents du SIAAP*). It provides the operators with detailed flow predictions and suggestions for control strategies.

The MAGES system relies on monitoring and data processing tools which include 2 000 in-line flow rate sensors. As illustrated in figure 5, it combines these measurements with other field data such as valve positions and pumping station readings, observations made by operators (such as incidents or planned maintenance) and rain forecasts. The hydraulic status of flow rates and levels is displayed for 400 key points along the network, with a horizon of 24 hours in dry conditions and 6 hours in wet weather. The calculations are based on deterministic modelling of the transport and storage infrastructures consisting of 23 000 calculation points in a model combined with an optimization process to control wastewater and storm water flows supported by predictive models that detect potential critical situations. It then determines the optimized management scenario based on a multi-objective function to reflect the costs of flooding, storage and treatment. This enables indicating mitigation actions to avoid flooding of streets or to reduce the risks of excessively high flow rates at the entrance of the treatment stations. The MAGES system coordinates this information among the multiple stakeholders active at various management levels in the region. The final decision to intervene, for example by redirecting flows towards a storm water retention basin or to create additional storage volume in the vast sewer networks, is currently made by the operator at a local or departmental level. The MAGES system updates every five minutes to identify the current system state to permit to fully optimize the available capacities and temporary storage volumes located upstream of the wastewater treatment plants (Blanchet *et al*., 2008).

Figure 5: interactive structure of MAGES control system. (Tabuchi et al., 2020)



The development and implementation of the real-time control system MAGES has demonstrated that the operators are able to make early decisions in response to predicted changes in flow rates and therefore are able to optimize the available capacity in the network with preemptive transfer of volumes. This enables to stabilize the flow of water into the treatment plants, thereby reducing the probability of a critical situation impacting the treatment performance. It allows for a reduction in untreated discharges from the sewage network during periods with prolonged and heavy rains, which currently amounts up to 3 to 4% of the annual treated volume. In addition, the MAGES control system enables a strong coordination between the many actors and stakeholders in charge of urban sanitation in the Greater Paris region. It has also become a useful tool for planning and design for maintenance works and investments in the infrastructure and asset management (Tabuchi *et al.*, 2020).

2.5 Smart Monitoring Surface Water Quality

An advanced measurement system of the Seine river water quality has been developed within the framework of the MeSeine monitoring program. Frequently expanded since its first deployment in 1990, the monitoring system allows observing the quality in terms of physico-chemistry, bacteriology, micro-contamination and biological diversity. Three decades of collaboration with academic research institutes has resulted in the development of innovative monitoring strategies for surface water quality.

Sensors located at eight measuring stations along a 120 km long stretch of the Seine river measure the concentrations of dissolved oxygen and ambient water temperature. Furthermore, three stations are equipped with additional sensors for organic carbon, nitrogen, phosphorus, conductivity and chlorophyll in the water (see figure 6 left and center). These additional sensors are currently in a semi-development phase. The objective is to create an integrated network, of which the measurement data are transmitted and in real-time generating more than 1 million readings per year. In addition, at 14 sampling sites along the river, weekly water samples are taken for analysis in the laboratory to support the general state of river quality with up to 100 different parameters. These efforts allow nearly 15,000 analyzes per year on the presence of microorganisms and pathogens. Furthermore, fish surveys and inventories of aquatic life (fish, macro-invertebrates and diatoms) are regularly performed along a 110 km section of the river (SIAAP, 2020).

The analysis results are accessible to the public through an online platform (https://www.siaap.fr/metiers/qualite-de-leau-et-milieu-naturel) and are updated on a daily basis. This includes readings of oxygen level, a monthly status account of its quality in relation to regulatory thresholds and an annual report with a complete environmental quality assessment.

Figure 6: deployment of in-situ sensors and mobile sensors for real-time collection quality indicators. (© SIAAP Observatoire MeSeine [left and center], Fluidion DRONE [right] fluidion.com/en)

More recently, the water and wastewater utilities in the Paris region have decided to cooperate in the development of a real-time water quality monitoring tool and a machine-learning based forecasting tool to predict water quality at designated bathing sites. The development is actively supported by the regional and local water authorities, as well as the Organizing Committee for the Olympic and Paralympic Games (OCOG) in charge of the 2024 Olympic Games in Paris. The cooperation takes place in the framework of the digital-water.city (DWC) research project funded by the European Union's H2020 Research and Innovation Program and relies on the technical expertise of Fluidion (France) and Kompetenzzentrum Wasser Berlin (KWB) (Germany). As part of selected European cities, the DWC project in the greater Paris aims in particular to develop an early warning system based on a forecast of bathing water quality in the Seine and Marne river, to test an innovative technology for real-time in-situ bacterial monitoring as well as develop a mobile application to inform key stakeholders and citizens.

The prototype early warning system will be built to indicate contamination in specific bathing sites based on machine learning and statistical modelling using a set of local data, including indicators of water quality, measurements of flow rates, temperature and predictions of rainfall intensity. The statistical model approach uses an advanced machine learning model for robust early forecast of water quality developed by the German DWC project partner KWB (Kompetenzzentrum Wasser Berlin). It will be calibrated with data output from an existing numerical river model named ProSe as applied to selected sites of the rivers Seine and Marne along urbanized areas in the Paris region. This will take into consideration the structure and future changes of the municipal sanitation system, such as the points of discharge from wastewater treatment plants and the locations of combined sewer overflows upstream the bathing areas (digital-water.city, 2020).

The real-time monitoring of the bacterial concentrations in the surface water will be managed by the deployment of in-situ sensors, manufactured by the company Fluidion. The sensors can perform seven rapid bacterial quantifications on a single battery charge, are remotely controlled, and wirelessly send data in real time to a data analytics server. In addition to the in-situ sensors, a remotely-controllable battery-operated floating drone, also manufactured by Fluidion, can be deployed to perform sampling and subsequent analyses at multiple locations in the river, both mid-stream and near shore (figure 6 right). In the initial phase, the focus will be on a rapid quantification of E.coli and enterococci bacteria. Collection and quality control of data will be managed by a secure data analysis system, which will also link the communication of the real-time bathing water quality assessment to an online platform. Currently in a testing-phase at several pilot sites, it is expected to be operational for the Olympic & Paralympic Games organized in 2024 in the French capital.

On the longer term, there is the possibility to link this surface water quality assessment tool with the MAGES real-time control system of the regional sewage transport network and the plant-wide process models developed for the treatment plants. This should permit the overall optimization of urban wastewater management based on river water quality objectives.

2.6 Modelling Surface Water Quality

In order to assess treatment performance and provide forecasts on the quality of the river Seine, the river model ProSe (*Programme Seine*) has been developed by the École des Mines de Fontainebleau in collaboration with SIAAP within the framework of the Piren-Seine (*Programme interdisciplinaire de recherche sur l'environnement*

de la Seine) research program. This multidisciplinary research program started in 1989 with the aim to improve the understanding of the river system. It has since then been gradually developed into a basin-wide approach that supports the modelling of complex interactions between the surface water and aquifers (hydrosystem), the ecosystem (phytoplankton, bacteria, fish communities), the agronomic system (crops and soils), the river users (drinking water, navigation), and urban and industrial development (e.g., waste water treatment plants) (Flipo *et al*., 2020).

This allows studying the impact of anthropogenic pollutant discharges from wastewater treatment plants and the sewer network on the quality of the rivers Seine and the Marne crossing Paris. The model takes into account combined sewer overflow events during wet weather, as well as effluent outlets of municipal wastewater treatment plants and considers meteorological conditions in the river basin and catchment area. With the numerical model ProSe, it has become possible to study future scenarios of river basin management for the next 50 years based on the perspectives and claims of a diverse group of water users and stakeholders (Flipo *et al*., 2020).

The ProSe model simulations provide for projections of comprehensive surface water quality indicators taking into account the principal biogeochemical processes taking place in the river water and the sediments. It predicts key water quality variables, sediment deposit levels and benthic exchanges (Laborie et al., 2016). As an early assessment tool, ProSe is being used to assist in decision-making directing investments towards improving water quality. The model is applied in design and operational studies, such as for the pre-dimensioning of additional storage capacity for storm water and treatment processes, and for scenario analysis in impact studies

A simplified river water quality model ('SimSeine') is currently being developed by the model*EAU* (Université Laval) research team for the Seine River basin. Using a hydrological conceptual approach, the SimSeine model is meant to function as a fast alternative in real-time applications for operational support in municipal wastewater management. It will provide predictions of principal water quality indicators at critical locations along the river that are linked to the municipal wastewater network, such as in the case of combined sewer overflow events (CSO) and treatment plant effluents (SIAAP, 2021).

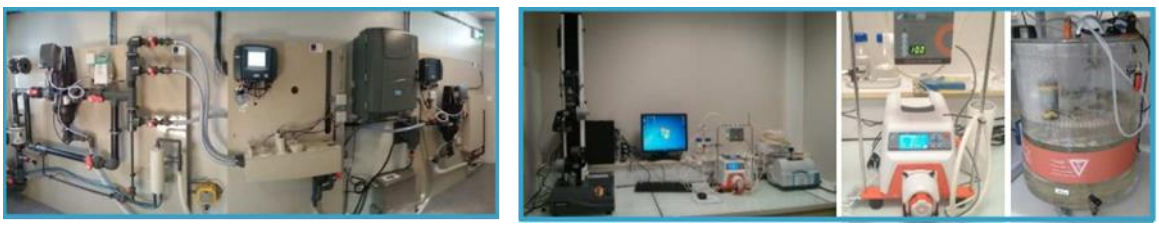
2.7 Smart Monitoring of Wastewater Concentrations at Plant-level

Monitoring of the main quality variables of wastewater and sludge is crucial for process control and optimization (Ingildsen and Olsson, 2016). Typically, important qualitative indicators that are measured are the Organic Matter concentration as Biological Oxygen Demand (BOD) and Chemical Oxygen Demand (COD), the suspended matter as Total Suspended Matter (TSS) and concentration levels of nitrogenous (N) and phosphoric (P) compounds. The fractionation between soluble and suspended matter in the wastewater has influence on the treatment options (Henze et al., 2008).

There has been a strong development of sensors capable to be placed in-line in wastewater streams to monitor dissolved oxygen, suspended solids and organic matter, as well as several nutrients (such as nitrogen and phosphorus) over the last few decades. Today, these sensors can provide reliable data and have become more affordable (Ingildsen and Olsson, 2016). Nevertheless, the importance of data quality checks, such as through autocalibration and built-in sensor checks, cleaning systems and reliable sample preparation units for data quality, has not diminished and these methods are still necessary as practical solutions to the frequently encountered problems with in-situ sensors in wastewater streams (e.g. Vanrolleghem and Lee (2003)).

Within the Mocopée program, innovative methods to measure organic and bacterial pollution of wastewater and to estimate the energy content of sludge have been developed to be deployed in treatment plant operations.

Figure 7: collection of real-time process data (left) supported by on-site laboratory analyses (right). (© SIAAP)



As mentioned, the organic matter content is an important indicator of water quality is, which negatively affects ecological quality of the receiving waters; it is therefore subject to strict emission standards. Measurements of organic matter are also necessary to improve the efficiency of the denitrification process, where the biodegradability of organic matter directly affects the nitrate and nitrite removal rates. Fast and accurate characterization of the qualities and amounts of organic matter should enable achieving significant savings in terms of addition of external carbon and energy consumption for recirculation flows and aeration.

Research by the LEESU laboratory in the framework of the Mocopée program has developed 3D-fluorescence spectroscopy for rapid measurement of dissolved organic matter in wastewaters entering the treatment plants (Goffin *et al.*, 2018). The study coupled optical 3D fluorescence spectrometry with the use of parallel factor analysis (PARAFAC) as a fast alternative to conventional indicators such as the biological oxygen demand (BOD_5) and the chemical oxygen demand (COD). The study showed that soluble COD concentrations in wastewater can be predicted via a simple linear model and that soluble BOD_5 can be predicted by a multiple linear model. The applied method is non-destructive, low cost, possible to automate and applicable to wastewater treatment allowing for online monitoring at high frequency (duration of the measurement of the order of a minute) of organic matter. These data can be used to reduce the energy costs (chemical reagents, pumping and aeration) and provide a better control of treatment processes (Goffin *et al.*, 2018).

Advanced sensors are being applied in the treatment trains for the monitoring of the principal qualitative indicators. For example; the concentrations of nitrates and nitrites in the water are monitored in real-time by in-line sensors using the amount of absorption of ultraviolet light as a measure of the NOx concentration. The levels of suspended solids (TSS) are measured using an infrared light source (880 nm) by an optical beam scattering method, which consists of an infrared emitting source and two photoelectric cells that function as receptors inclined at 90 ° in relation to the incoming light beams.

2.8 Process Modelling of Wastewater Treatment

The use of computer-aided mathematical modelling to simulate physical, chemical and microbial processes in municipal wastewater treatment is advancing at a fast pace. Numerical models based on empirical and phenomenological descriptions are providing deeper insight in the biochemical and hydraulic processes and have found their way to support performance evaluation and decision-making for planning, design, control and optimization (Rieger *et al.*, 2012).

For the Parisian region, SIAAP has put in place several research and development projects focusing on process modelling and process control. The first objectives are to obtain well calibrated and validated simulation models of each of the principal treatment procedures, including the chemically enhanced settling tanks and the biological activated filtration, as well as for the anaerobic sludge digestion. In a second phase, these individual models should be connected in order to allow plant-wide simulation. These will serve as support tools for operators and allow for process optimization. As part of the Mocopée program, one collaborative research project with model*EAU* (Université Laval) aims to further improve the use of models and to validate their application on more large-scale processes. This will permit automated integration with (near) real-time measurement data and statistical analysis for data management and data reconciliation, as well as to enable forecasting based on observations and an inclusive assessment of treatment costs, energy consumption and environmental impact.

Modelling of Treatment by Chemically Enhanced Settling

A phenomenological model using mathematical functions has been developed to simulate the processes taking place in chemically enhanced settling tanks (Zhu, 2020). This model is implemented in the WEST® modelling environment (DHI) and includes estimations of the average settling velocity and non-settleable fraction suspended matter as a function of chemical dosage and calculations of the environmental costs and greenhouse gas emissions of operations. It is based on an adaptation of the SimDec model (Bernier *et al.*, 2016) made suitable for a chemically enhanced settling tank as well as to allow the use of different coagulants to be added to the treatment process, such as ferric chloride ($FeCl_3$) in the primary and/or tertiary treatment phase. Another promising application that is being studied, is the usage of Alufer (a mix of iron chloride and aluminum iron) to improve sludge incineration.

Consequently, a sub-model for the reagent injection has been added to better control the quantities of chemicals injected. For the settling unit – which is equipped with tubed lamellas – modifications had to be made to the

original model regarding the fraction of non-settleable particles at the inlet of the lamellar settling tank and to the average particle settling velocity (Zhu, 2020).

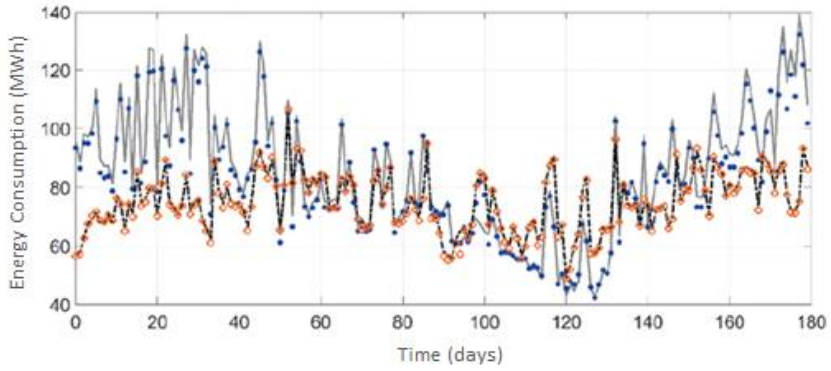
Modelling of Biologically Active Filtration Systems (BAF)

The development of a well calibrated and validated model for the biological removal of ammonia and other nitrogen compounds by multi-stage biofiltration in the Paris wastewater treatment plants has been a pursued vision for over 10 years. Commercially available modelling software does often not offer biofiltration as a standard (sub-) model. Consequently, in collaboration with Université Laval, a mechanistic model ('SimBio') was developed that simulates the secondary biological treatment stage at the Seine-Aval plant. It was built in Matlab/Simulink (MathWorks) and considers the most relevant mass transport and microbial processes, including a two-step nitrification and denitrification processes for the removal of nitrogen and phosphorus assimilation (Bernier *et al.*, 2014).

Furthermore, researchers are studying ways to reduce the emissions of greenhouse gases from nitrifying biologically active filters (BAFs), which have been found to be important emitters of nitrous oxide (N_2O), a powerful greenhouse gas. To that end, the SimBio model was extended with a gas-phase that allowed describing full-scale N_2O emissions. Research performed in collaboration with INRAE showed that the improved gas-liquid transfer model improved the predictions of nitrifying performance. The model successfully described nitrification and the mass transfer of O_2 and N_2O (Fiat *et al.*, 2019).

More recently, in the framework of PhD research co-supervised by Université de Technologie de Compiègne and model*EAU* (Université Laval), an enhanced biofiltration model for the 5,7 million PE Seine-Aval plant was implemented in the WEST® (DHI) modelling software. It allowed integrating the earlier developments into a complete biofiltration model of a full-scale secondary treatment that includes the three-stage treatment (pre-denitrification – nitrification – post-denitrification). This model now also considers the production of nitrites and nitrous oxide during nitrification and denitrification. It includes process control systems such as for aeration and allows for an assessment of environmental and economic costs, as shown in figure 8 (Zhu, 2020).

Figure 8: Simulation results of the energy consumption of the nitrification stage in the WEST Biofiltration model (blue dots: aeration observations, gray line: aeration simulation, orange circles: pumping observations, black dotted line: pumping simulation). (Zhu, 2018)



Simulation of the biofilm processes has been improved by using a higher amount of biofilm layers and a dynamic boundary layer with variable thickness. An improved media exchange between media layers is added which helps to maintain proper description of the growth and decay of biomass. The production of nitrites, nitrogen monoxide and nitrous oxide during treatment is now also considered and the nitritation reaction model was expanded to two-pathway N_2O production by nitrifying organisms. The aeration control strategy, which is controlled by Feed-Forward (FF) and Feed-Back (FB) control rules based on the observed ammonium concentrations, was added to predict the injected air flow. The model has been calibrated and successfully validated with high frequency data collected over long periods by in-line sensors supported by laboratory composite samples at the Seine-Aval plant (Zhu, 2020).

This phenomenological model allows calculating the production of greenhouse gasses during treatment as well as estimating the energy consumption and economic costs related to aeration and hydraulic pumping. The integration of the particular aeration control system allows realizing scenario analyses, such as aeration increase

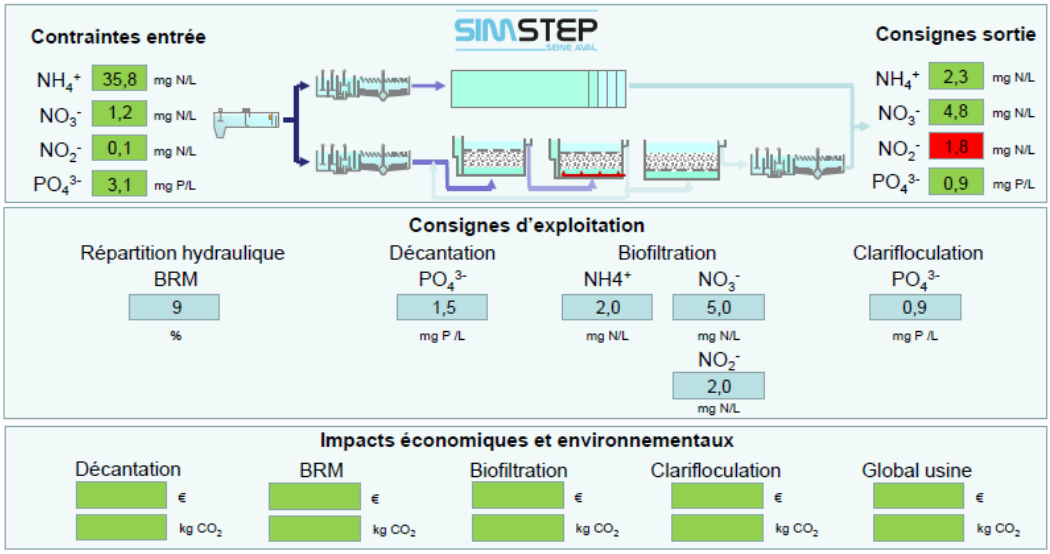
or reduction by applying different ammonium set points and the estimation of the required energy. It also allows testing different aeration control strategies during the nitrification process by replacing the current FF and FB control with alternative control strategies. Similar controls have been added for the injection of organic matter (such as methanol) during pre- and post-denitrification. These allow providing support for process optimization and the development of control strategies. The model will provide the possibility to evaluate the economic and environmental impact of different operating scenarios of the biofilters at the Seine-Aval plant (Zhu, 2018).

Plant-wide Modelling and AI-driven plant models

Models of individual treatment phases cannot indicate the best possible overall treatment. For that, it is necessary to look at plant-wide operations using a well calibrated model covering the entirety of processes taking place. When it is to be used as an aid for sustainable management, it needs to also include the economic and ecological costs of treatment. Plant-wide simulation models (PWM) are being developed for process control able to mirror the real-time operational status of the wastewater plants, make predictions and assist in decision support for (preemptive) maintenance. These digital twin models will mimic the physical plant conditions and provide a performance evaluation to predict best possible treatment with an indication of the costs of treatment and environmental impact, such as energy consumption and GHG emissions. It can become a meaningful tool for operator training and communication among the many levels of operation and management. These dynamic process models will enhance overall system understanding and support decision making of ongoing operations and improve planning for future scenarios. It's expected that it will significantly reduce downtime and support the decisions for stable and secure investments in equipment and infrastructure.

Once the 'upgraded' models describing the individual processes have been fully calibrated and validated with recent measurement data, the objective is to interconnect the primary treatment, the three-stage biofiltration and tertiary treatment models. During the Mocopée 2018 colloquium, a demonstration was given by SIAAP researchers to illustrate a possible graphical user interface of a plant-wide process simulation model – shown in figure 9. In a clear and concise manner, the demo-screen indicates the real-time status of the principal treatment installations, such as of the primary settler tanks and the secondary biological filtration units. It will indicate by color codes if the real-time capacity of these installations is sufficient to meet the requirements on the effluent quality as calculated by the model. It will indicate the routing and quality guidelines within the processing lanes (in the center part of the screen) and evaluate the economic and environmental impacts of the proposed operating settings (Zhu, 2018b).

Figure 9: illustration of a possible GUI of a plant-wide model showing inflow & outflow concentrations (upper section), indications for operational limits for the different treatment stages (middle section) and a window for calculated financial and environmental costs (lower section). Red colored boxes indicate alerts. Figure forms part of a presentation given at the Mocopée colloquium in Colombes, France on the 4/12/2018 (© SIAAP)



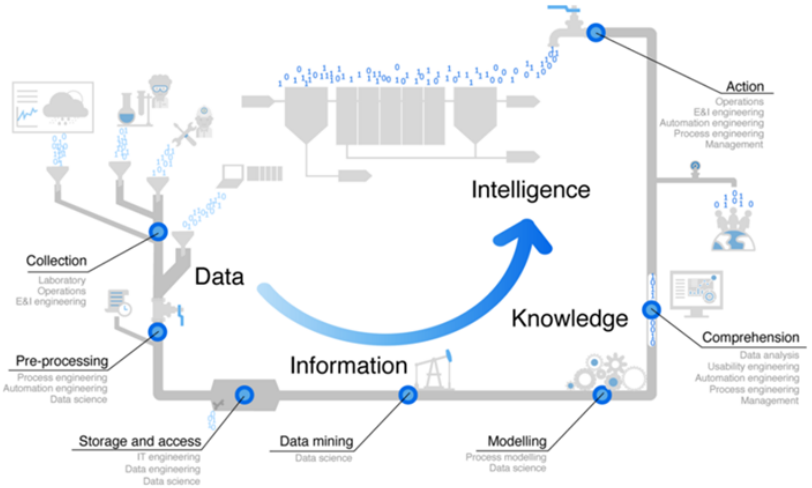
A first digital twin model is under development for the Seine-Aval treatment plant connecting individual models of each treatment stage, including the above mentioned models of the chemically enhanced settler and biofiltration, all implemented in the WEST modelling environment. The final form is still being discussed, but the

expected capabilities of this digital twin should include the hydraulic and qualitative aspects of the wastewater, and be able to base optimization rules on the economic and ecological costs of operation. It should be capable of simulating the impact of distribution scenarios for the influent over the different treatment stages, as well as simulate the internal recirculation rules that are based on the type of influent (dry weather profile vs. storm weather for example), the effluent quality standards to achieve, the available and required equipment and the optimization objectives. For each treatment stage, the model should propose the most efficient control settings for the process variables, such as aeration, recirculation, addition of coagulants and reagents. Accordingly, it could become a decision-making support system for operators; as well as provide better understanding of the inter-relationships and thus become a training tool. In 'training-mode', it will allow performing detailed analyses of extreme operational conditions, including equipment failures, without placing the real plant's performance at risk.

In cooperation with the model*EAU* research team at Université Laval, studies are underway on the various data processing and management components of digital twins, such as aspects related to automated data quality control, data structure and accessibility, parameter sensitivity and model calibration, all part of the data-pipeline (Therrien *et al.*, 2020). Data management aims to evaluate and augment data quality in order to transform raw data into actionable insights and use that to make intelligent decisions with confidence. As indicated in figure 10, many data processing steps are required to create an understanding and intelligence from raw measurements.

This begins with automated data collection, which is quite often hampered by the harsh operational conditions prevailing in wastewater treatment. It necessitates data cleaning to remove faulty and erroneous readings ('data preprocessing'). Going further along the data 'pipeline', this process includes proper data storage and easy accessibility in order for data to become useful and timely information (Therrien et al., 2020). The cleaned, stored and accessible datasets form the variables used by numerical models for simulations translating data into knowledge. Models can provide forecasts and be used for scenario analyses to deepen our understanding of the real processes. Statistical modeling approaches could provide additional support, such as for the extrapolation of processes that aren't fundamentally understood or captured by (mechanistic) models. Such hybrid semi-parametric models allow model predictions beyond experimentally tested process conditions (Glassey and Von Stosch, 2018). Finally, this knowledge or intelligence acquired may lead to redefine operational settings or procedures in order to mitigate identified problems ahead of time.

Figure 10: From data to intelligence. For each step in the pipeline, the most essential elements are listed. (Therrien et al., 2020)



The hybridization of data analysis methods and intelligent forecasting modelling systems with process control is expected to provide many benefits. A digital twin model of a treatment plant can evaluate the status of a process, predict changes and prescribe mitigation actions to keep performance within norms and in compliance. Recent bibliometric analyses (e.g. Zhao *et al.*, 2020) and reviews (e.g. Corominas *et al.*, 2018) indicate the progression of academic studies applying an artificial intelligence-based data-driven approach to improve the reliability of wastewater systems, increase water quality and lower energy consumption.

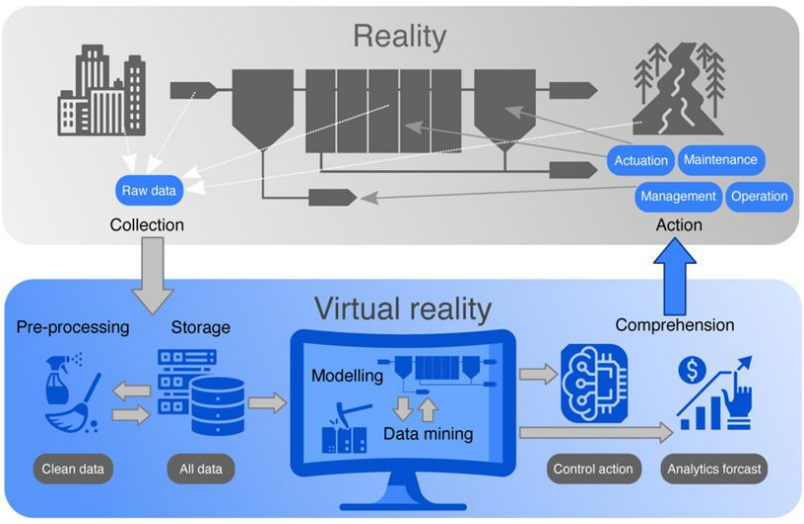
An ongoing Ph.D. research project (2020 – 2023) within the Mocopée framework in collaboration with the W-SMART Association, LEESU and model*EAU* aims to develop a hybrid process control system that is based on

knowledge of processes, but driven by data collection and interpretation. It aims to integrate the strengths of plant-wide simulation models to evaluate scenarios with the strength of data-driven AI models to recognize patterns and update the model parameters to maintain the alignment of the model with reality. It will integrate automatic data collection and quality control from high-frequency measurements from in-situ sensors and laboratory analyses (originating from LIMS and SCADA systems) to remove faulty readings and outliers and allow interpolation to replace missing data. These time series will then be uploaded to the dynamic process model to perform multiple simulation runs. Examples could be with scenarios of chemical dose rate injections or flow distribution patterns over the different treatment trains. The prediction horizon will vary from a few hours to several days depending on the time range of input data available (e.g. meteorological forecasts).

This adaptive hybrid digital twin will make use of machine learning techniques to continuously check and update the model parameters characterizing physical, biological and chemical reactions happening during treatment processes. As illustrated in figure 11, this requires an interactive exchange of data in the form of measurements and instructions between the real plant and the virtual reality. The digital twin will make use of data processing, modelling and data mining techniques to decide if adjustments of manipulated variables are recommended in the ongoing processes to maintain performance within thresholds. This digital twin model will be able to recognize type changes or anomalies in the operating conditions or parameter values of the plant model. The concept is an intermediate approach to combine a more conventional theory-based ('white-box') process model used for design and scenario analyses with a data-based ('black-box') application to provide early detection and prediction. The idea is that such a hybrid digital twin model ('grey box') will follow in real-time the variable wastewater conditions in the different treatment stages, detect important changes early on, make predictions and suggest actions for mitigation to assure optimal treatment from the perspectives of water quality, economics and environmental protection, at minimal downtime. It could also be used for process design or upgrade and optimization by running different scenarios and evaluating the modeled outputs.

The data-driven component of this smart system will function in support of the process model to monitor and correct the deterministic model. In comparison with fully data-driven models, such hybrid systems generally have lower data requirements, because it can function with a shorter history of time series data and higher prediction accuracy (Glassey and Von Stosch, 2018). The ongoing research will contribute to the integration of a hybrid AI model as part of the functional decision support system. In that study, the objective is to demonstrate the feasibility of a merger of the strengths of empirical knowledge with those strengths of data analytics and introduce a framework for large scale industrial application.

Figure 11: interaction between physical and digital world. (From: Therrien et al., 2020).



A pilot study by W-SMART (Tinelli and Juran, 2019) presented an automated statistical model and artificial intelligence supported algorithms that allow early detection of anomalies in data collected at the physical-chemical clarifier of the Seine-Aval plant. The developed algorithms exploit the concept of expert pattern recognition of deviations from normal conditions and assess the probability of an event or an anomaly, filtering false alarms.

The study evaluated the removal rates of organic matter and suspended solids and the coagulant reagents used in a chemically enhanced settler and provided a risk-based classification of potential anomalies. The study applied both Supported Vector Machines (SVM) and Artificial Neural Networks (ANN). The SVM technique showed a classification error of approximately 2%, increasing to approximately 4% using the ANN. Overall; this study presented an efficient method for anomaly detection and risk-based classification, as well the ability to isolate anomalies according to a risk severity scale (Tinelli and Juran, 2019).

Modelling of Integrated Urban Wastewater Systems

The many interactions and dependencies within and between wastewater and sludge treatment processes, as well as the sequential relationship between the sewer network, the treatment plant and the receiving natural environment compel an integrated system-wide approach. This should cover the entire chain of processes taking place within the catchment area, including sewage collection and transport, the conditions in the receiving natural environment, rainfall and other meteorological conditions, both in real-time and forecasts within the next 6 to 12 hours. The latter is in particular of importance during extreme weather events, such as for the predictions of storm water run-off volumes.

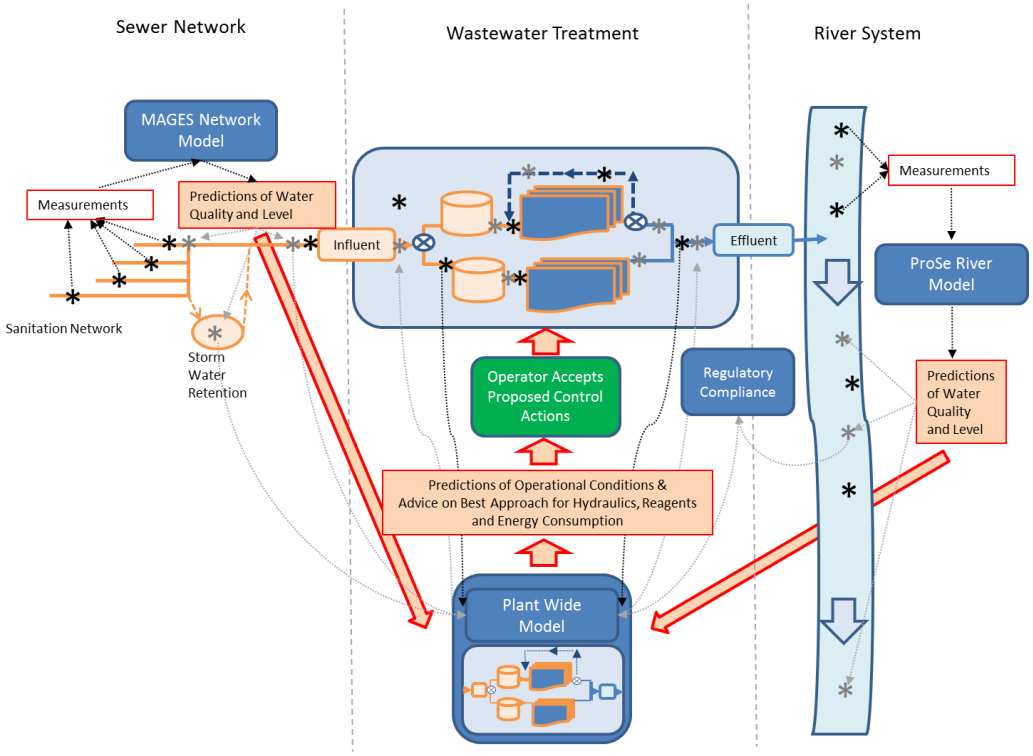
As is demonstrated in figure 12, an integrated approach would compel the monitoring and collection of principal indicators in each sub-system that would feed a numerical model in order to provide for predictions at the scale of the sub-system. That would take place both at the level of the river system and the sewer network respectively by means of the ProSe model and the MAGES model. The river model would take measurement data coming from in-situ sensors located upstream and downstream of the wastewater plant outlet, as well as meteorological forecasts and then produce predictions of water quality and water levels in the river section downstream from the wastewater treatment plant outlet. Based upon these predictions, different scenarios could be run to simulate and define the optimum scenario or most efficient treatment performance, which could include recommendations to adjust the treatment processes (temporarily) to better meet the river quality conditions.

On the upstream side of the urban water cycle, the sewer model would provide predictions of sewage volumes and storm water run-off in a time horizon of 6 to 12 hours ideally. The further development is foreseen to include the adaptation of the MAGES monitoring and forecasting system to be able to include data monitoring on key pollutants as well as flow rates at selected locations in the transport network.

Both predictions of the river model and the sewer model will serve as inputs to the plant-wide process model. It will use these to run scenario's and define actions to ensure best possible treatment given the river water quality and predicted sewage volumes. . Scenario definition includes the search for optimum treatment during highly dynamic conditions which might disrupt treatment performance. This could be the case with high flow rates and/or high nutrients and pollutant loads of the influent, specifically when this occurs in a situation of a low dilution capacity of the river. It can also be applied to situations where the constraints for treatment are less strong, such as under low influent flow rates with stable pollutant loads and a high dilution capacity of the river. For each scenario, the model predictions will include the performance levels for wastewater and sludge treatment, cost estimations of energy and reagents used during treatment, as well as an estimation of the by-products formed (e.g., biogas) and the resources or materials that are recoverable. As presented during the Mocopée colloquium in 2018, it is foreseen that the plant operator holds the final decision regarding the proposed modifications (Zhu, 2018b).

Figure 12: conceptual representation of an IUWS that integrates the data stemming from the sewer system (left) and the river (right) with a model of a wastewater treatment plant (center). The black lines represent the flow of monitoring data, the grey lines model predictions and the red lines the exchange of the acquired model-based knowledge between different

models and the operator. Figure is based on a presentation given during the Mocopée colloquium in Colombes, France on the 4/12/2018 (© SIAAP).



On the longer term, the development of an integrated monitoring and modelling system of the complete urban water cycle for the Paris region incorporating the sewer network, the river as well as all of the six wastewater treatment plants and storm water retention basins operated by SIAAP is being discussed. Such an integrated approach will be necessary to place the river at the center of the performance assessment of wastewater processing. The expected outcome of such a holistic approach will permit to create a powerful monitoring, data processing and process control tool of the municipal wastewater system. This will allow to better control plant performance while placing the focus on river water quality standards (Tabuchi et al., 2020).

2.9 Conclusion and Perspective

Decades of investments by the wastewater treatment industry in research and development have led to impressive progress in the fields of collection and treatment, with the introduction of stable and more reliably performing technologies, such as biofiltration, membrane bioreactors and chemically enhanced lamellar settling tanks. Furthermore, major advances have been made in the field of instrumentation for monitoring and process control with an expanded network of smart sensors connected in real-time to advanced supervisory communication and data systems. Management is up-scaled to the level of the urban water system, providing better coordination among stakeholder for operations and integrated planning for investments. The observed large improvements in the quality of natural waters and the repopulation of biotopes by a large variety of aquatic life forms in the rivers flowing through the Paris region are a great testimony to these efforts (Rocher et al., 2017).

The task is now to maximize the potential of existing infrastructures while minimizing the environmental impact of operations and continue to assure precise compliance to the norms for effluent quality. Achieving this, while being pressured by continuing urbanization, rising effects of climate change, the implementation of a circular economy with a transition to energy neutrality and a reuse of resources as well as growing public demands for clean and safe natural rivers for recreational usage. The modern and compact technologies installed in smaller decentralized treatment plants come with a higher operational complexity and are activated by biological and chemical processes that are not (yet) entirely understood. Consequently, the energy consumption of compact and highly reactive treatment systems is higher per unit treated water. The transition to a more sustainable economic and ecological management will necessitate a better understanding and control of these treatment processes.

Thanks to improved measuring and sensing technologies combined with advanced data management systems and computer based dynamic modeling tools, the level of understanding in the sanitation sector of the physical, biological and chemical processes taking place during sewage and sludge treatment has been increased (e.g. Henze et al, 2008 and Ingildsen and Olsson, 2016). This improved understanding helps identifying important process parameters and improve overall system control. It helps to build confidence in design and planning with dynamic process models that allow testing different control approaches and safely train operators without placing the real plant's performance at risk.

In the current dynamic wastewater world, the effects of climate change, population growth, the emergence of new types of pollutants (e.g., micropollutants and new forms of pathogens) and the desire to revalorize energy and other resources necessitate the application of increasingly sophisticated and comprehensive process control and data management systems. Such systems also allow dealing with the requirements to reduce the environmental impact (e.g., lower the carbon footprint) and the operational expenses (e.g., by reducing the electricity demand). These aspects are driving today's investments in the next generation model-based and Artificial Intelligence-based control systems.

The adaptation of AI-based control for process optimization could be the next paradigm shift in the sanitation industry and lead to a fundamental change in the approach to wastewater management. The next level hybridization of deterministic knowledge-based white-box models and black-box big data analytics systems has the potential to radically transform the operation of municipal wastewater management and bring in the fourth industrial revolution with smart water systems and a sustainable management of water resources (Garrido-Baserba *et al*., 2020). Smart systems build confidence in defining new control strategies and can assist in planning, design and operations. Potential critical conditions could be detected ahead of time and the maintenance management planning can be modified to avoid a decrease in performance or down-time of equipment. This will lead to stable and reliable removal of pollutants and helps to reach the full potential of resource recovery.

Nevertheless, a full-scale implementation of operational digital twins in daily wastewater plant management still faces many challenges to overcome. A major challenge that has been identified is the large amount of data being generated by intelligent monitoring equipment for flow and nutrients loads in wastewater. The integration of automatic controllers using real-time measurements necessitates comprehensive quality control of the data to avoid decisions being based on erroneous data. Enhanced systems for automated data collection, fault detection and the pre-processing of data have the potential to deliver regular, high frequent and trustworthy data, which is a concrete requirement to transform it into meaningful and actionable information in real-time (Therrien et al., 2020). With the advancements of highly accurate and self-calibrating in-situ sensing techniques, the amount of data rejected or lost, for example due to faulty readings or sensor drifts, are expected to lower significantly in the near future.

The classification of general standards and protocols for data management and data reconciliation, as well as a systematic methodology for process modelling could benefit the implementation of data-based control in the wastewater sector. Furthermore, the development of sufficient know-how for AI applications through expertise and skills training will require close collaboration with data-science experts and experts of other related fields. Research and innovation programs serving the sanitation sector, such as SIAAP's InnEAUvation program coordinating a holistic and inter-disciplinary approach incorporating engineers, scientists and policymakers, will support the on-going digitalization of the wastewater treatment sector.

The first results of these synergies as well as the lessons learned presented in this chapter, should be seen as an example for wastewater utilities worldwide and encourage each one to develop similar integrated and coordinative research partnership programs in preparation of becoming a smart water utility.

References

- Azimi S. and Rocher V. (2017). Energy consumption reduction in a wastewater treatment plant. Water Practice and Technology, 12(1), 104–116.
- Bernier J., Rocher V., Guérin S., Lessard P. (2014): Modelling the nitrification in a full-scale tertiary biological aerated filter unit. Bioprocess and Biosystems Engineering, 37(2), 289-300.
- Blanchet, B., Fradin, A., Tarif, P. (2008). Modèle d'aide à la gestion des effluents du SIAAP (Mages)-Outil d'aide à la gestion dynamique et coordonnée du système d'assainissement de la région parisienne. Techniques Sciences Méthodes, 12, 55-67.

- Blumensaat, F., Leitão, J.P., Ort, C., Rieckermann, J, Scheidegger, A., Vanrolleghem, P.A., Villez, K. (2019). How urban storm- and wastewater management prepares for emerging opportunities and threats: Digital transformation, ubiquitous sensing, new data sources and beyond – A horizon scan. Environmental Science and Technology, 53(15), 8488-8498.
- Corominas, L., Acuña, V., Ginebreda, A., & Poch, M. (2013). Integration of freshwater environmental policies and wastewater treatment plant management. Science of the total environment, 445, 185-191.
- Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U., & Poch, M. (2018). Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. Environmental modelling & software, 106, 89-103.
- Digital-water.city (2020). https://www.digital-water.city accessed January 2020.
- Fiat, J., Filali, A., Fayolle, Y., Bernier, J., Rocher, V., Spérandio, M., Gillot, S. (2019). Considering the plug-flow behavior of the gas phase in nitrifying BAF models significantly improves the prediction of N2O emissions. Water Research, 156, 337-346.
- Flipo N, Lestel L, Labadie P, Meybeck M and Garnier J (2020). Trajectories of the Seine River basin. In: Flipo N, Labadie P, Lestel L (eds) The Seine River basin. Handbook of environmental chemistry. Springer, Cham
- Garrido-Baserba, M., Corominas, L., Cortés, U., Rosso, D., Poch, M. (2020). The fourth revolution in the water sector encounters the digital revolution. Environmental Science and Technology, 54(8), 4698-4705.
- Glassey, J., and Von Stosch, M. (Eds.). (2018). Hybrid modeling in process industries. CRC Press, Boca Raton, FL, USA.
- Goffin, A., Guérin, S., Rocher, V., Varrault, G. (2018). Towards a better control of the wastewater treatment process: excitation-emission matrix fluorescence spectroscopy of dissolved organic matter as a predictive tool of soluble BOD_5 in influents of six Parisian wastewater treatment plants. Environmental Science and Pollution Research, 25(9), 8765-8776.
- Henze, M., van Loosdrecht, M. C., Ekama, G. A., & Brdjanovic, D. (Eds.). (2008). Biological wastewater treatment. IWA publishing, London, UK.
- Ingildsen, P. and Olsson, G. (2016). Smart Water Utilities. Complexity made simple. IWA Publishing, London, UK.
- Laborie, B., Rocher, V., Vilmin, L., Poulin, M., Raimonet, M., Benard, L., ... & Escoffier, N. (2016). Le modèle de prédiction de la qualité de la Seine ProSe Améliorations apportées par les récents travaux de recherche (Piren-Phase VI).
- Rieger, L., Gillot, S., Langergraber, G., Ohtsuki, T., Shaw, A., Takacs, I., Winkler, S. (2012). Guidelines for Using Activated Sludge Models. IWA publishing, London, UK.
- Rocher, V., Azimi, S., Paffoni, C., Rousselot, O., Tabuchi, J. P., Bernier, J. (2017). Evolution de la qualité de la Seine en lien avec les progrès de l'assainissement de 1970 à 2015. Editions Johannet, Paris, France.
- Sharma, A. K., Guildal, T., Thomsen, H. R., & Jacobsen, B. N. (2011). Energy savings by reduced mixing in aeration tanks: results from a full scale investigation and long term implementation at Avedoere wastewater treatment plant. Water Science and Technology, 64(5), 1089-1095.
- SIAAP (2017a). SIAAP'S strategy for an ecological transition towards sustainable development. SIAAP, Paris, France.
- SIAAP (2017b). Adapting the wastewater management of Greater Paris to its future challenges. Presentation of SIAAP's strategy for the future. SIAAP, Paris, France.
- SIAAP (2018). Innover dans les pratiques de monitoring et d'exploitation des stations d'épuration. Enseignements scientifiques et techniques tirés de la phase I du programme (2014-2017) Mocopée. SIAAP, Paris, France.
- SIAAP (2019). Rapport d'activité et de développement durable. SIAAP, Paris, France.
- SIAAP (2020). Programme scientifique innEAUvation coordonné par la Direction Innovation. SIAAP, Paris, France.
- Tabuchi, J. P., Benard, L., Blanchet, B., Esculier, F., Mouchel, J. M., Poulin, M., Saison, O. (2013). Gestion en temps réel du système d'assainissement de la région parisienne en fonction de la qualité de la Seine NOVATECH.
- Tabuchi, J.P, Tassin B., Blatrix C. (2016) Greater Paris Water and global change, Water megacities and global change, Portraits of 15 emblematic Cities of the World, UNESCO/ARCEAU.
- Tabuchi J.P., Blanchet B. and Rocher V. (2020) Integrated Smart Water Management of the sanitation system of the Greater Paris region, Water International, 45:6, 574-603.
- Therrien, J.D., Nicolaï, N., Vanrolleghem, P.A. (2020). A critical review of the data pipeline: how wastewater system operation flows from data to intelligence. Water Science and Technology, 82(12), 2613-2634.
- Tinelli, S. and Juran, I. (2019). Tertiary physico-chemical settling during wastewater treatment: multi-criteria optimization of chemical dosage. Artificial Intelligence Application for Wastewater Treatment Process (AIA-WWTP). Final report, W-SMART.
- Vanrolleghem, P. A., and Lee, D. S. (2003). On-line monitoring equipment for wastewater treatment processes: state of the art. Water Science and Technology, 47(2), 1-34.
- Zhao, L., Dai, T., Qiao, Z., Sun, P., Hao, J., & Yang, Y. (2020). Application of artificial intelligence to wastewater treatment: A bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse. Process Safety and Environmental Protection, 133, 169-182.
- Zhu, J., Bernier, J., Azimi, S., Pauss, A., Rocher, V., Vanrolleghem, P.A. (2018a) Comprehensive modelling of full-scale nitrifying biofilters and validation under different configurations. Presentation IWA Nutrient Removal and Recovery Conference, 18 - 21 November 2018, Brisbane, Australia.
- Zhu, J., Bernier, J., Azimi, S., Pauss, A., Rocher, V., Vanrolleghem, P.A. (2018b) Vers un modèle d'aide à la décision pour la station d'épuration Seine Aval Objectifs industriels et état d'avancement du projet NEXTSTEP. Presentation ASTEE phase I of the Mocopée program colloquium, 4 December 2018, Colombes, France.
- Zhu, J. (2020). Modélisation détaillée du fonctionnement de la filière complète de biofiltration de a station de traitement des eaux usées Seine Aval (PhD Thesis). Université de Technologie de Compiègne, Sorbonne Université.

3 Annex III Results of Global Sensitivity Analysis

3.1 Model Parameters

Parameters_PreDn_Model_MS_SAV_2019_2020_CAL							Parameters_DAM_Model_Filiere_JZ_SAV_2018			Difference
Name	Value	Unit	Default Value	Lower bound	Upper bound		Name	Value		%
Category: Manipulated Variables							**Category: Manipulated Variables**			
Group: Operational							**Group: Operational**			
pH	7	-	7	0	14		PH	7		100%
Kla	1800	1/d	750	0	5000		Kla	1800		100%
Airflow	10000	m3/d	750	0	+INF		Airflow	10000		100%
Temp	16,341887	degC	15	-273,15	+INF		Temp	16,341887		100%
Airflow_Real1	0	m3/d	750	0	+INF		Airflow_Real1	0		
Filter_Nb	58		84	-INF	+INF		Filter_Nb	58		100%
Pump_Nb	24		24	-INF	+INF		Pump_Nb	24		100%
Category: Parameters							**Category: Parameters**			
Group: Aeration							**Group: Aeration**			
OTR_Energy	1800	g/kWh	1800	-INF	+INF		OTR_Energy	1800		100%
m	0,85		0,85	-INF	+INF		m	0,85		100%
Period	1	d	1	0	+INF		Period	1		100%
Nretg	7		1	-INF	+INF		Nretg	7		100%
Flowfactor_A	0,753		0,753	-INF	+INF		Flowfactor_A	0,753		100%
Flowfactor_B	40		40	-INF	+INF		Flowfactor_B	40		100%
Differential_Pressure	101000	Pa	101000	0	+INF		Differential_Pressure	101000		100%
Temp_Environment	12	degC	12	-273,15	+INF		Temp_Environment	12		100%
Temp_Heat	10	degC	10	-273,15	+INF		Temp_Heat	10		100%
Temp_normal	0	degC	0	-273,15	+INF		Temp_normal	0		
Efficiency_Compressor	0,873		0,873	-INF	+INF		Efficiency_Compressor	0,873		100%
Efficiency_Pump	0,793		0,793	-INF	+INF		Efficiency_Pump	0,793		100%
Group: boundary							**Group: boundary**			
ShA	2		2	-INF	+INF		ShA	2		100%
ShK	0,51		0,51	-INF	+INF		ShK	0,51		100%
ShB	4,23		4,23	-INF	+INF		ShB	4,23		100%
ShM	0,833		0,833	-INF	+INF		ShM	0,833		100%
ShN	0,59		0,6	-INF	+INF		ShN	0,59		100%
ShBeta	0,339		0,333	-INF	+INF		ShBeta	0,339		100%
Viscocity	0,1129	m2/d	0,1129	-INF	+INF		Viscocity	0,1129		100%
Diameter	0,0045	m	0,004	0	+INF		Diameter	0,0045		100%
Porosity	0,35	dUnit/dUnit	0,365	-INF	+INF		Porosity	0,35		100%
Diffusion_Reduction	0,7	dUnit/dUnit	0,2	-INF	+INF		Diffusion_Reduction	0,7		100%
Group: Composition parameters							**Group: Composition parameters**			
i_X_B	0,109	gN/gCOD	0,086	0	0,2		i_X_B	0,098		111%
i_X_BP	0,0046	gP/gCOD	0,01	0	1		i_X_BP	0,0055		84%

i_X_P	0,06	gN/gCOD	0,06	0	0,2		i_X_P	0,06		100%
i_X_UP	0,008	gP/gCOD	0,01	0	1		i_X_UP	0,006		133%
Group: Conversion factors							**Group: Conversion factors**			
F_BOD_COD	0,65	-	0,65	0	1		F_BOD_COD	0,65		100%
F_TSS_COD	0,75	-	0,75	0	1		F_TSS_COD	0,75		100%
Group: Dimension							**Group: Dimension**			
d_max	0,0002	m	0,0005	0	+INF		d_max	0,0002		100%
d_min	0,00000001	m	5E-07	0	+INF		d_min	0,00000001		100%
d_maxth	0,0003	m	0,0003	0	+INF		d_maxth	0,0003		100%
A	74914,35	m2	173	0	+INF		A	74914,35		100%
d_ctrlmax	0,00004		4E-05	-INF	+INF		d_ctrlmax	0,00004		100%
d_ctrlwashmax	0,00002		2E-05	-INF	+INF		d_ctrlwashmax	0,00002		100%
V_R	1	m3	1	0	+INF		V_R	1		100%
L	13,15	m	1	0	+INF		L	13,15		100%
H	0,5	m	1	0	+INF		H	0,5		100%
W	13,15	m	1	0	+INF		W	13,15		100%
A_Sp_R	1333	1/m	400	0	+INF		A_Sp_R	1333		100%
f_void	0,35	dUnit/dUnit	0,5	-INF	+INF		f_void	0,35		100%
Group: filtering							**Group: filtering**			
geo	1		10	-INF	+INF		geo	1		100%
lami	0,43	1/m	0,0006	0	+INF		lami	0,4		108%
vmax	1	m3	1	0	+INF		vmax	1		100%
xx	1		1	-INF	+INF		xx	1		100%
yy	1,06		1	-INF	+INF		yy	0,83		128%
zz	1		1	-INF	+INF		zz	1		100%
Group: Gas-Liquid							**Group: Gas-Liquid**			
D_N2	2E-09	m2/s	2E-09	0	+INF		D_N2	2E-09		100%
D_N2O	1,77E-09	m2/s	1,77E-09	0	+INF		D_N2O	1,77E-09		100%
D_NO	2,21E-09	m2/s	2,21E-09	0	+INF		D_NO	2,21E-09		100%
D_O2	2,08E-09	m2/s	2,12E-09	0	+INF		D_O2	2,08E-09		100%
H_N2	0,77	atm.m3.Mol-1	0,77	0	+INF		H_N2	0,77		100%
H_N2O	0,02	atm.m3.Mol-1	0,02	0	+INF		H_N2O	0,02		100%
H_NO	0,53	atm.m3.Mol-1	0,53	0	+INF		H_NO	0,53		100%
H_O2	0,77	atm.m3.Mol-1	0,77	0	+INF		H_O2	0,77		100%
P_Airflow	1		1	-INF	+INF		P_Airflow	1		100%
f_CO2	0,0004	-	0,0004	0	1		f_CO2	0,0004		100%
f_NO	0,0000005	-	5E-07	0	1		f_NO	0,0000005		100%
f_N2	0,78	-	0,78	0	1		f_N2	0,78		100%
f_N2O	3,28E-07	-	3,28E-07	0	1		f_N2O	3,28E-07		100%
f_O2	0,21	-	0,21	0	1		f_O2	0,21		100%

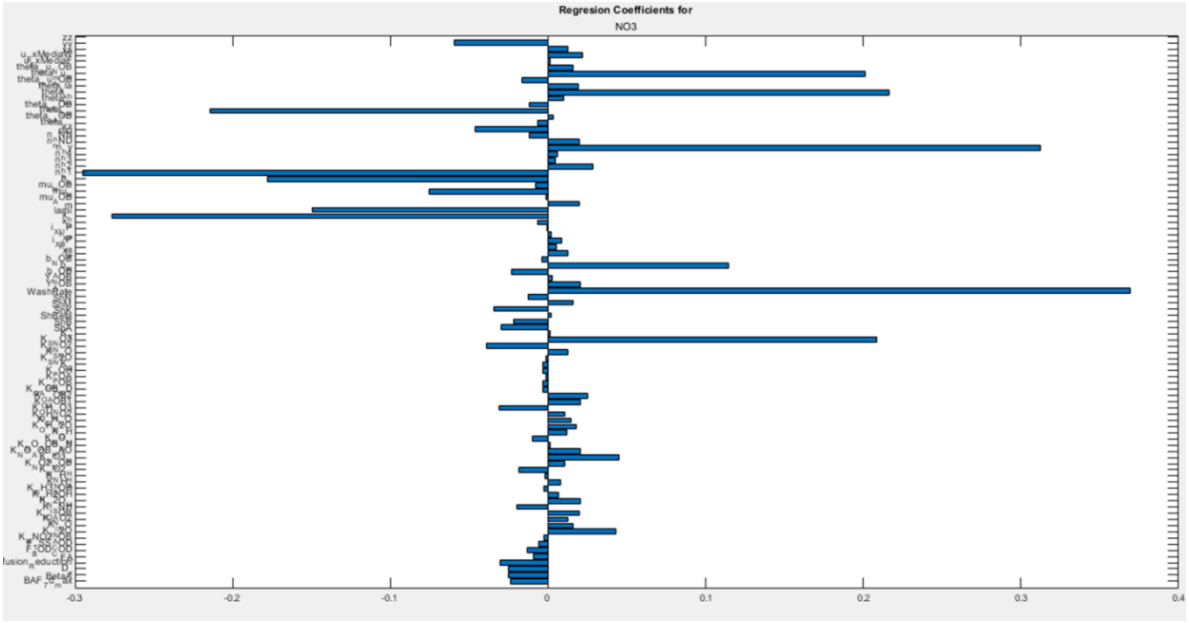
K_O2_CO2	1	dUnit/dUnit	1	-INF	+INF		K_O2_CO2	1		100%
MM_CO2	44	g/mol	44	0	+INF		MM_CO2	44		100%
MM_O2	32	g/mol	32	0	+INF		MM_O2	32		100%
MM_NO	30	g/mol	30	0	+INF		MM_NO	30		100%
MM_N2O	44	g/mol	44	0	+INF		MM_N2O	44		100%
MM_N2	28	g/mol	28	0	+INF		MM_N2	28		100%
V_N_GAS	0,0224	m3	0,0224	0	+INF		V_N_GAS	0,0224		100%
Atm	101325	Pa	101325	0	+INF		Atm	101325		100%
V_Tunnel	260	m3	312	0	+INF		V_Tunnel	260		100%
H_Tunnel	1,5	m	1,5	0	+INF		H_Tunnel	1,5		100%
KlaR	0,032	dUnit/dUnit	0,032	-INF	+INF		KlaR	0,032		100%
Kla_ref	1800	1/d	1968	0	5000		Kla_ref	1800		100%
Air_ref	10000	m3/d	44980	0	+INF		Air_ref	10000		100%
Group: Kinetic							**Group: Kinetic**			
K_HNO2_AOB	0,004	gN/m3	0,004	0	100		K_HNO2_AOB	0,004		100%
K_I_N2O	0,075	gN/m3	0,075	0	100		K_I_N2O	0,075		100%
K_I_NO	0,31	gN/m3	0,3	0	100		K_I_NO	0,31		100%
K_I_NO2	0,5	gN/m3	0,5	0	100		K_I_NO2	0,5		100%
K_I_O_AOB	0,8	gO2/m3	0,8	0	100		K_I_O_AOB	0,8		100%
K_I_Red_NO3	0,00000005	gN/m3	0,05	0	100		K_I_Red_NO3	0,00000005		100%
K_I_SNH	1E-12	gN/m3	1E-12	0	1		K_I_SNH	1E-12		100%
K_N2O_H	0,05	gN/m3	0,05	0	100		K_N2O_H	0,05		100%
K_NH	1	gNH3-N/m3	1	0	10		K_NH	1		100%
K_NH2OH	0,018	m3/d	0,0148	0	100		K_NH2OH	0,018		100%
K_NH3_AOB	0,004	gN/m3	1,47	0	100		K_NH3_AOB	0,004		100%
K_NH_A	0,05	g/m3	0,05	0	100		K_NH_A	0,05		100%
K_NH_H	0,05	g/m3	0,05	0	100		K_NH_H	0,05		100%
K_NO	0,5	gNO3-N/m3	0,5	0	2		K_NO	0,5		100%
K_NO2_H	0,2	gN/m3	0,2	0	100		K_NO2_H	0,2		100%
K_NO2_NOB	0,2	gN/m3	0,2	0	100		K_NO2_NOB	0,2		100%
K_NO3_H	0,2	gN/m3	0,2	0	100		K_NO3_H	0,2		100%
K_NO_AOB_HAO	0,0003	gN/m3	0,0003	0	100		K_NO_AOB_HAO	0,0003		100%
K_NO_AOB_NN	0,008	gN/m3	0,008	0	100		K_NO_AOB_NN	0,008		100%
K_NO_H	0,053	gN/m3	0,05	0	100		K_NO_H	0,053		100%
K_OA	0,4	gO2/m3	0,4	0	10		K_OA	0,4		100%
K_OH	0,1	gO2/m3	0,2	0	10		K_OH	0,1		100%
K_OH_N2O	0,1	gO2/m3	0,2	0	10		K_OH_N2O	0,1		100%
K_OH_NO	0,1	gO2/m3	0,2	0	10		K_OH_NO	0,1		100%
K_OH_NO2	0,1	gO2/m3	0,2	0	10		K_OH_NO2	0,1		100%
K_OH_NO3	0,1	gO2/m3	0,2	0	10		K_OH_NO3	0,1		100%
K_O_AOB1	1,5	gO2/m3	1,5	0	100		K_O_AOB1	1,5		100%
K_O_AOB2	0,3	gO2/m3	0,3	0	100		K_O_AOB2	0,3		100%
K_O_AOB_ND	0,5	gO2/m3	0,5	0	100		K_O_AOB_ND	0,5		100%
K_O_NOB	0,6	gO2/m3	0,6	0	100		K_O_NOB	0,6		100%
K_POA	0,01	g/m3	0,01	0	10		K_POA	0,01		100%
K_POH	0,01	g/m3	0,01	0	10		K_POH	0,01		100%
K_S	20	gCOD/m3	20	0	100		K_S	20		100%
K_S_N2O	40	gCOD/m3	40	0	100		K_S_N2O	40		100%

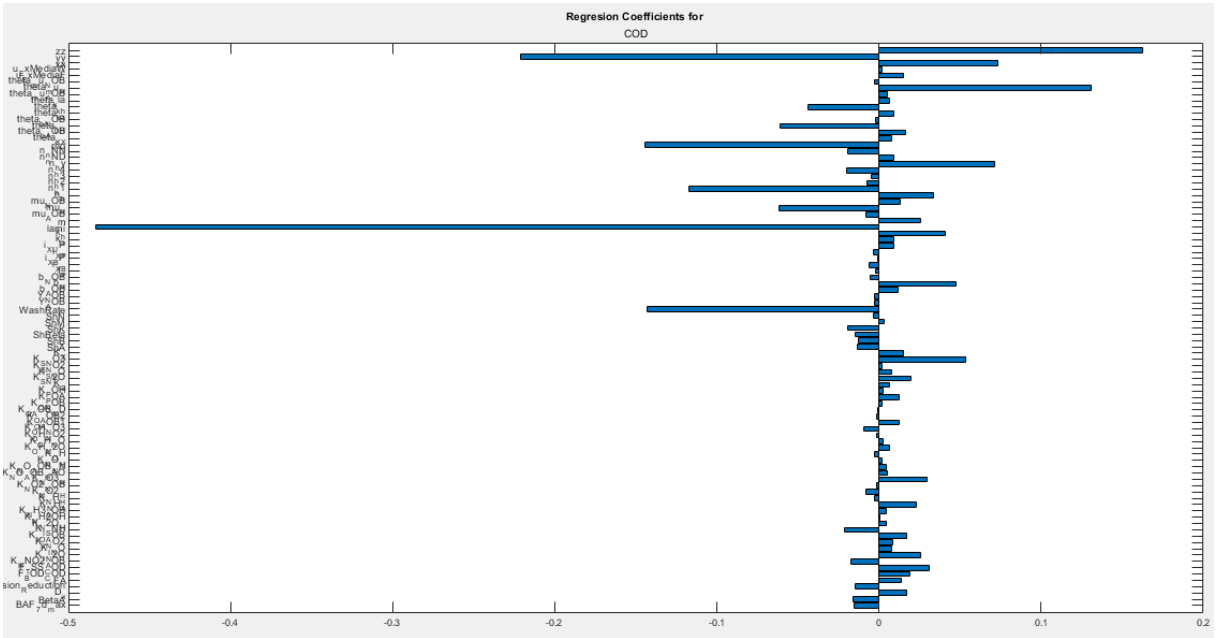
K_S_NO	17,7	gCOD/m3	20	0	100		K_S_NO	17,7		100%
K_S_NO2	16,06	gCOD/m3	20	0	100		K_S_NO2	13,8		116%
K_S_NO3	41,37	gCOD/m3	20	0	100		K_S_NO3	28,6		145%
K_X	0,03	gCOD/gCOD	0,03	0	100		K_X	0,03		100%
K_red_NO3	0,1	gN/m3	0,1	0	100		K_red_NO3	0,1		100%
Temp_Ref	20	degC	20	-273,15	+INF		Temp_Ref	20		100%
b_AOB	0,17	1/d	0,17	0	25		b_AOB	0,17		100%
b_H	0,73	1/d	0,62	0	25		b_H	0,7		104%
b_NOB	0,17	1/d	0,17	0	25		b_NOB	0,17		100%
k_a	0,065	m3/(gCOD*d)	0,08	0	0,25		k_a	0,075		87%
k_h	2,6	gCOD/(gCOD*d)	3	0	25		k_h	2,5		104%
kr	0	m3/(gCOD*d)	0,1	-INF	+INF		kr	0		
mu_AOB	0,8	1/d	0,8	0	5		mu_AOB	0,8		100%
mu_H	4,74	1/d	6	0	20		mu_H	5,18		92%
mu_NOB	1	1/d	1	0	5		mu_NOB	1		100%
n_h	0,47	-	0,4	0	1		n_h	0,47		100%
n_h1	0,28	-	0,28	0	1		n_h1	0,28		100%
n_h2	0,19	-	0,16	0	1		n_h2	0,14		136%
n_h3	0,31	-	0,35	0	1		n_h3	0,47		66%
n_h4	0,35	-	0,35	0	1		n_h4	0,35		100%
n_hy	0,72	-	0,75	0	1		n_hy	0,78		92%
n_nND	0,25	-	0,25	0	1		n_nND	0,25		100%
n_nNN	0,0015	-	0,0015	0	1		n_nNN	0,0015		100%
Group: media							**Group: media**			
u_ExMediaF	2	d-1	0,001	0	+INF		u_ExMediaF	2		100%
u_ExMediaW	20	d-1	0,01	0	+INF		u_ExMediaW	20		100%
Group: Pump							**Group: Pump**			
Coef_Loss	0,061232		0,061232	-INF	+INF		Coef_Loss	0,061232		100%
Pump_Height	10,84	m	10,84	0	+INF		Pump_Height	10,84		100%
Group: Stoichiometry							**Group: Stoichiometry**			
Y_AOB	0,21	gCOD/gN	0,21	0	4,57		Y_AOB	0,21		100%
Y_H	0,72	gCOD/gCOD	0,67	0	1		Y_H	0,72		100%
Y_NOB	0,06	gCOD/gN	0,06	0	4,57		Y_NOB	0,06		100%
f_P	0,08	-	0,08	0	1		f_P	0,08		100%
Group: system							**Group: system**			
gra	9,8	m/s2	9,8	0	+INF		gra	9,8		100%
Group: System							**Group: System**			
BetaA	0,95	-	0,95	0	1		BetaA	0,95		100%
FA	1	-	1	0	1		FA	1		100%
alfaA	1	-	0,95	0	1		alfaA	1		100%
d_ini	0,00001	m	0,0002	0	+INF		d_ini	0,00001		100%
rho	133600	g/m3	40000	0	+INF		rho	133600		100%
k_detach	0,4	1/d	0,0001	0	+INF		k_detach	0,4		100%
k_detachwater	1	1/d	0,0001	0	+INF		k_detachwater	1		100%
p_filter	0,05	dUnit/dUnit	0,05	-INF	+INF		p_filter	0,05		100%

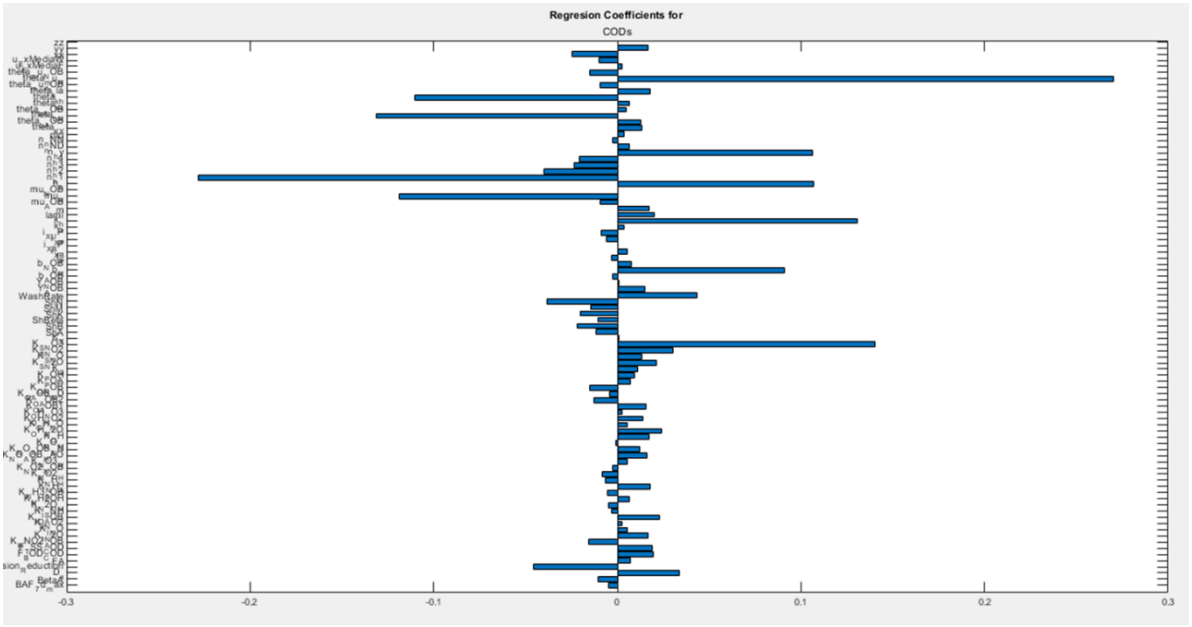
p_wash	0,05	dUnit/dUnit	0,6	-INF	+INF		p_wash	0,05		100%
Group: Temperature correction							**Group: Temperature correction**			
theta_K_X	1,116		1,116	-INF	+INF		theta_K_X	1,116		100%
theta_b_AOB	1,116		1,116	-INF	+INF		theta_b_AOB	1,116		100%
theta_b_H	1,12		1,12	-INF	+INF		theta_b_H	1,12		100%
theta_b_NOB	1,116		1,116	-INF	+INF		theta_b_NOB	1,116		100%
theta_k_a	1,072		1,072	-INF	+INF		theta_k_a	1,072		100%
theta_k_h	1,116		1,116	-INF	+INF		theta_k_h	1,116		100%
theta_kla	1,024		1,024	-INF	+INF		theta_kla	1,024		100%
theta_mu_AOB	1,103		1,103	-INF	+INF		theta_mu_AOB	1,103		100%
theta_mu_H	1,072		1,072	-INF	+INF		theta_mu_H	1,072		100%
theta_mu_NOB	1,103		1,103	-INF	+INF		theta_mu_NOB	1,103		100%
Group: Transport							**Group: Transport**			
D_x	1	1/d	1E-10	-INF	+INF		D_x	1		100%
DH2O	1E-10	m2/d	1E-10	-INF	+INF		DH2O	1E-10		100%
DS_I	0,0001	m2/d	0,0001	-INF	+INF		DS_I	0,0001		100%
DS_S	0,000058	m2/d	5,8E-05	-INF	+INF		DS_S	0,000058		100%
DS_O	0,000216	m2/d	0,000216	-INF	+INF		DS_O	0,000216		100%
DS_NO3	0,00017	m2/d	0,00017	-INF	+INF		DS_NO3	0,00017		100%
DS_NO2	0,00017	m2/d	0,00017	-INF	+INF		DS_NO2	0,00017		100%
DS_NO	0,000244	m2/d	0,000244	-INF	+INF		DS_NO	0,000244		100%
DS_N2O	0,000155	m2/d	0,000155	-INF	+INF		DS_N2O	0,000155		100%
DS_N2	0,00016416	m2/d	0,00016416	-INF	+INF		DS_N2	0,00016416		100%
DS_NH	0,00017	m2/d	0,00017	-INF	+INF		DS_NH	0,00017		100%
DS_NH2OH	0,0000238	m2/d	2,38E-05	-INF	+INF		DS_NH2OH	0,0000238		100%
DS_ND	0,00014	m2/d	0,00014	-INF	+INF		DS_ND	0,00014		100%
DS_ALK	0,00022	m2/d	0,00022	-INF	+INF		DS_ALK	0,00022		100%
DS_PO	0,00022	m2/d	0,00022	-INF	+INF		DS_PO	0,00022		100%
DX_I	0	m2/d	0	-INF	+INF		DX_I	0		
DX_S	0	m2/d	0	-INF	+INF		DX_S	0		
DX_BH	0	m2/d	0	-INF	+INF		DX_BH	0		
DX_AOB	0	m2/d	0	-INF	+INF		DX_AOB	0		
DX_NOB	0	m2/d	0	-INF	+INF		DX_NOB	0		
DX_P	0	m2/d	0	-INF	+INF		DX_P	0		
DX_ND	0	m2/d	0	-INF	+INF		DX_ND	0		
Dg_NO	0	m2/d	0	-INF	+INF		Dg_NO	0		
Dg_N2	0	m2/d	0	-INF	+INF		Dg_N2	0		
Dg_N2O	0	m2/d	0	-INF	+INF		Dg_N2O	0		
Dg_O2	0	m2/d	0	-INF	+INF		Dg_O2	0		
Dg_CO2	0	m2/d	0	-INF	+INF		Dg_CO2	0		
eta	1	dUnit/dUnit	1	-INF	+INF		eta	1		100%

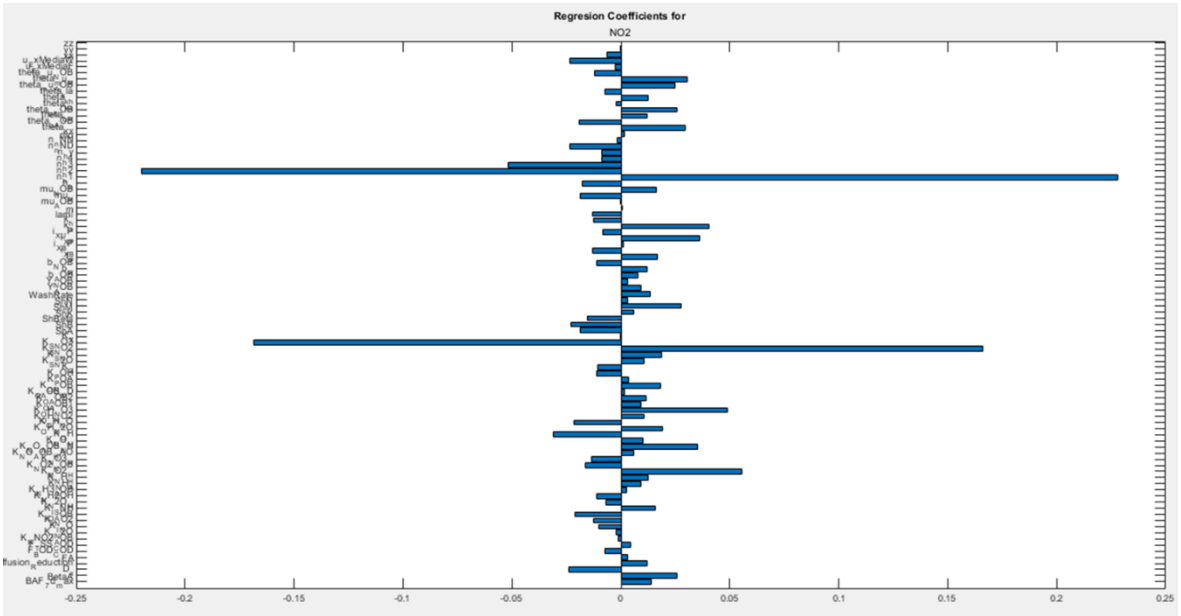
k_At	0		0,1	-INF	+INF		k_At	0		
Group: wash							**Group: wash**			
Washback_frequency	1	1/d	1	0	+INF		Washback_frequency	1		100%
Washback_Time	15	min	2,016	0	+INF		Washback_Time	15		100%
Filter_Time	1	d	1	0	+INF		Filter_Time	1		100%
Ensemencement_Time	0	d	2	0	+INF		Ensemencement_Time	0		
WashRate	0,55	dUnit/dUnit	0,5	-INF	+INF		WashRate	0,55		100%

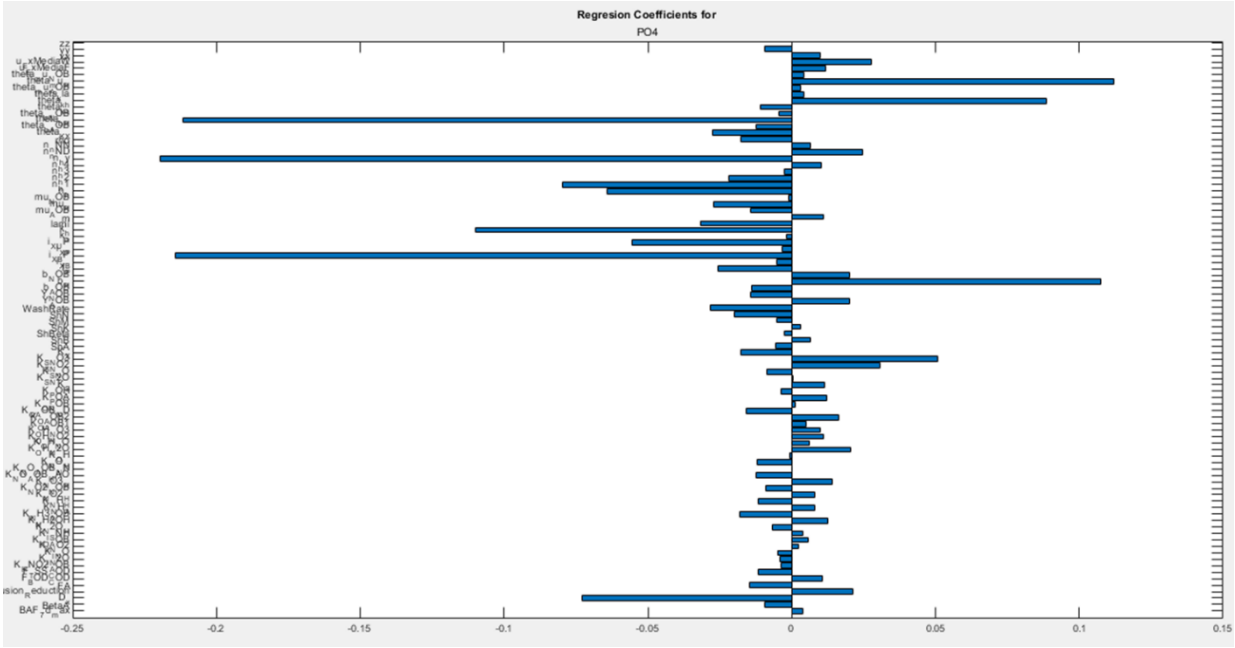
3.2 Plots of Regression Coefficients

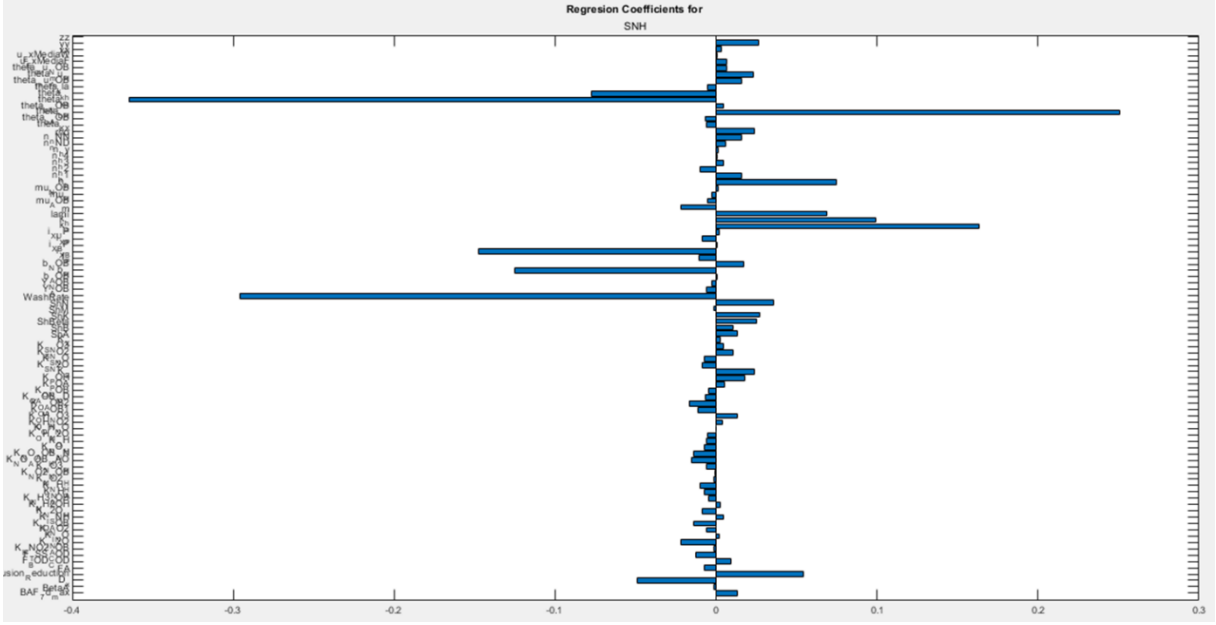


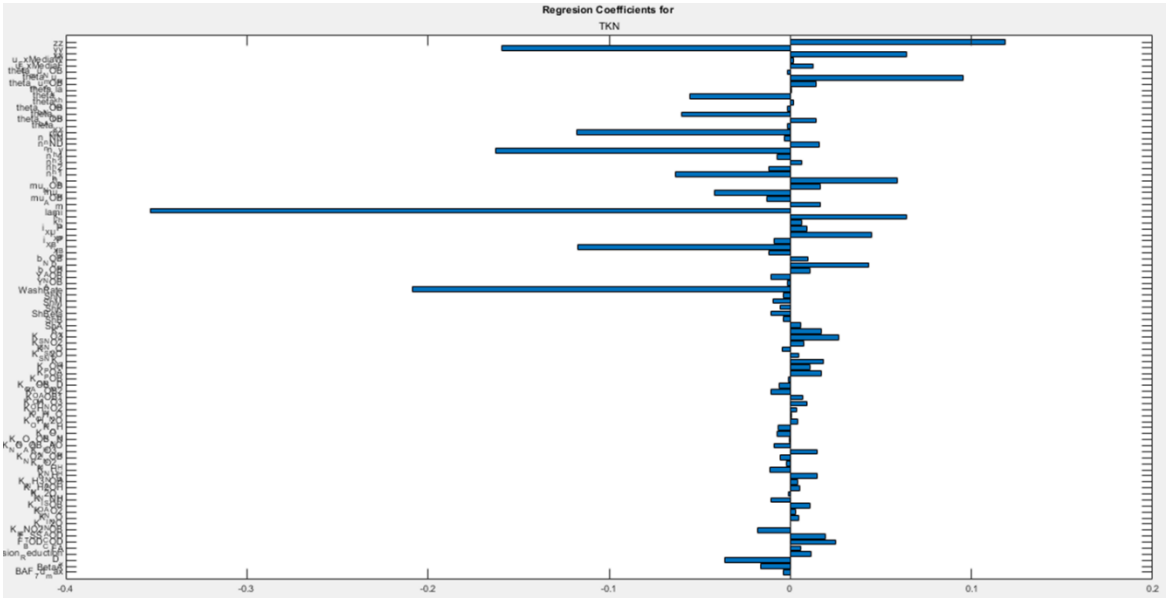


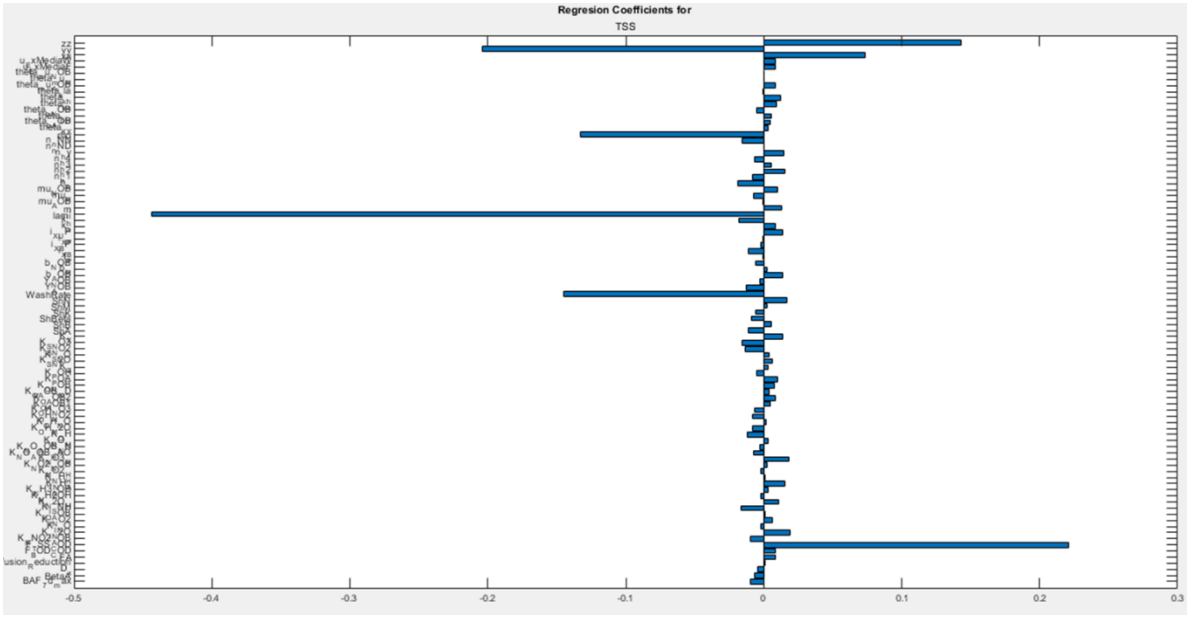

Regresion Coefficients for
CODs
-0.3
-0.2
-0.1
0
0.1
0.2
0.3


Regresion Coefficients for
NO2
-0.25
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25


Regresion Coefficients for
PO4
-0.25
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15


Regresion Coefficients for
SNH
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3


Regresion Coefficients for
TKN
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2


Regresion Coefficients for
TSS
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3

3.3 Tables of Regression Coefficients

Regression Coefficients for totalCOD effluent

regression coef COD	pm name	Unit	Description
-0,482897979	lami	1/m	Clean media filter coefficient
-0,221000419	yy	--	coefficient y of Ives equation 1
0,162396426	zz	--	coefficient z of Ives equation 1
-0,14412455	rho	g/m3	Biofilm density
-0,143201824	WashRate	dUnit/dUnit	Percentage washed during wash
0,130619686	theta_mu_H	--	Temperature correction factor for mu_H
-0,117408671	n_h1	--	Correction Factor for Anoxic Growth Of Heterotrophs NO3-
0,073604785	xx	--	coefficient x of Ives equation 1
0,071167829	n_hy	--	Anoxic yield factor
-0,061936728	mu_H	1/d	Maximum Specific Growth Rate for Heterotrophic Biomass
-0,061197319	theta_b_H	--	Temperature correction factor for b_H
0,053617688	K_S_NO3	gCOD/m3	Half-Saturation Coefficient for Heterotrophic Biomass NO3-
0,047809741	b_H	1/d	Decay Coefficient for Heterotrophic Biomass
-0,043884142	theta_k_h	--	Temperature correction factor for k_h
0,040576537	k_h	gCOD/(gCOD* d)	Maximum Specific Hydrolysis Rate
0,033821793	n_h	--	Correction Factor for Anoxic Hydrolysis
0,031162683	F_TSS_COD	--	Fraction TSS/COD
0,029900267	K_NO3_H	gN/m3	Saturation coeff of heterotrophs for NO3-
0,025896538	K_I_N2O	gN/m3	NO inhibition coefficient N2O
0,02560056	m	--	constant for diffuser and system configuration
0,023096449	K_NH_A	g/m3	Saturation coeff of autotrophs for ammonium
-0,021139555	K_I_SNH	gN/m3	growth limitation by ammonium
-0,019784374	n_h4	--	Correction Factor for Anoxic Growth Of Heterotrophs N2O
0,019471224	K_S_N2O	gCOD/m3	Half-Saturation Coefficient for Heterotrophic Biomass N2O
-0,019330503	ShK	--	coefficient K for number of Sherwood
-0,019084458	n_nNN	--	Reduction Factor for the NN pathway
0,018971247	F_BOD_COD	--	Conversion factor BOD/COD
-0,01754216	K_HNO2_AOB	gN/m3	Half Saturation coeff of heterotrophs for N2O
0,017317596	D_x	1/d	Diffusivity coefficient for particules components
0,016860395	K_I_O_AOB	gO2/m3	Inhibition constant by O2 on N2O production
0,01622703	theta_b_AOB	--	Temperature correction factor for b_AOB
-0,015712949	BetaA	--	reduction factor of solubility in wastewater
-0,015485678	BAF_7d_max	m	Maximum biofilm thickness
0,015162537	u_ExMediaF	1/d	Velocity of biofilm exchange of media when filtering
0,014864426	K_X	gCOD/gCOD	Half Saturation Coefficient for Hydrolysis Of Slowly Biodegradable Substrate
-0,014545796	Diffusion_Reduction	dUnit/dUnit	Reduction of diffusion coefficient in biofilm
-0,014512235	ShBeta	--	coefficient Beta for number of Sherwood
0,013579541	FA	--	Fouling factor for aeration

-0,01341231	ShA	--	coefficient A for number of Sherwood
0,01331272	mu_NOB	1/d	Maximum Specific Growth Rate for Autotrophic Biomass NOB
-0,012848353	ShB	--	coefficient B for number of Sherwood
0,012268497	K_POA	g/m3	Half Saturation coefficient for phosphate for Autotrophic Biomass
0,012237043	K_O_AOB1	gO2/m3	AOB affinity constant for O2 AMO reactions
0,011802037	b_AOB	1/d	Decay Coefficient for Autotrophic Biomass (AOB)
0,009350672	i_X_UP	gP/gCOD	Mass of Phosphore per Mass of COD in Products Formed added by JZ 06032017
-0,009202425	K_OH_NO3	gO2/m3	NO3- Oxygen Half-Saturation Coefficient for Heterotrophic Biomass
0,009169825	theta_k_a	--	Temperature correction factor for k_a
0,009096058	n_nND	--	Reduction Factor for the ND pathway
0,008897067	k_a	m3/(gCOD*d)	Maximum Specific Ammonification Rate
0,008730669	K_I_NO2	gN/m3	NO inhibition coefficient NO2-
-0,008302232	mu_AOB	1/d	Maximum Specific Growth Rate for Autotrophic Biomass AOB
-0,007932599	K_NO2_H	gN/m3	Saturation coeff of heterotrophs for NO2-
0,007868631	K_S_NO	gCOD/m3	Half-Saturation Coefficient for Heterotrophic Biomass NO
-0,007712566	n_h2	--	Correction Factor for Anoxic Growth Of Heterotrophs NO2-
0,007665218	theta_K_X	--	Temperature correction factor for K_X
0,007558317	K_I_NO	gN/m3	NO inhibition coefficient NO
0,006697361	theta_kla	--	Arrhenius parameter for kLa
0,006655822	K_OH_N2O	gO2/m3	N2O Oxygen Half-Saturation Coefficient for Heterotrophic Biomass
0,006311914	K_S	gCOD/m3	Half-Saturation Coefficient for Heterotrophic Biomass
-0,005929731	i_X_B	gN/gCOD	Mass of Nitrogen per Mass of COD in Biomass
-0,005596348	b_NOB	1/d	Decay Coefficient for Autotrophic Biomass (AOB)
0,005306356	K_NO_AOB_HAO	gN/m3	AOB affinity constant for NO (HAO)
0,005294766	theta_mu_AOB	--	Temperature correction factor for mu_AOB
-0,004797682	n_h3	--	Correction Factor for Anoxic Growth Of Heterotrophs NO
0,004707904	K_N2O_H	gN/m3	AOB Affinity constant for HNO2
0,004650053	K_NO_AOB_NN	gN/m3	AOB affinity constant for NO (NirK)
0,00433962	K_NH3_AOB	gN/m3	AOB affinity constant for NH3
-0,003313713	ShN	--	coefficient N for number of Sherwood
-0,003199289	i_X_P	gN/gCOD	Mass of Nitrogen per Mass of COD in Products Formed
0,003170287	ShM	--	coefficient M for number of Sherwood
-0,003053427	K_NH_H	g/m3	Saturation coeff of heterotrophs for ammonium
-0,003034377	Y_NOB	gCOD/gN	Yield for Autotrophic Biomass
-0,003002815	K_OH	gO2/m3	Oxygen Half-Saturation Coefficient for Heterotrophic Biomass
-0,002735976	Y_AOB	gCOD/gN	Yield for Autotrophic Biomass AOB
0,002671656	K_OH_NO	gO2/m3	NO Oxygen Half-Saturation Coefficient for Heterotrophic Biomass
-0,002484045	theta_mu_NOB	--	Temperature correction factor for mu_NOB
-0,002395046	theta_b_NOB	--	Temperature correction factor for b_NOB
-0,002382841	f_P	--	Fraction of Biomass Converted to Inert Matter

0,002380552	K_POH	g/m3	Half Saturation coefficient for phosphate for Heterotrophic Biomass
0,002159606	K_S_NO2	gCOD/m3	Half-Saturation Coefficient for Heterotrophic Biomass NO2-
0,00215018	K_O_NOB	gO2/m3	NOB affinity constant for O2
0,001801762	K_NO_H	gN/m3	Saturation coeff of heterotrophs for NO
0,001671316	u_ExMediaW	1/d	Velocity of biofilm exchange of media when washing
-0,001627268	K_NO2_NOB	gN/m3	Half-saturation coefficient for NO2-,NOB
-0,001536692	K_O_AOB2	gO2/m3	AOB affinity constant for O2 HAO reactions
-0,001513935	K_OH_NO2	gO2/m3	NO2- Oxygen Half-Saturation Coefficient for Heterotrophic Biomass
-0,000998312	i_X_BP	gP/gCOD	Mass of Phosphorus per Mass in COD Biomass added by JZ 06032017
0,000635673	K_NH2OH	gN/m3	AOB affinity constant for NH2OH
-0,000561301	K_O_AOB_ND	gO2/m3	AOB affinity constant for O2 ND pathway

Regression Coefficients for NO_3-N effluent

regression coef NO3	pm name	Unit	Description
0,369524727	WashRate	dUnit/dUnit	Percentage washed during wash
0,312250828	n_hy	--	Anoxic yield factor
-0,295035492	n_h1	--	Correction Factor for Anoxic Growth Of Heterotrophs NO3-
-0,276197288	k_h	gCOD/(gCOD*d)	Maximum Specific Hydrolysis Rate
0,216182796	theta_k_h	--	Temperature correction factor for k_h
-0,214236786	theta_b_H	--	Temperature correction factor for b_H
0,208348433	K_S_NO3	gCOD/m3	Half-Saturation Coefficient for Heterotrophic Biomass NO3-
0,20108089	theta_mu_H	--	Temperature correction factor for mu_H
-0,177943409	n_h	--	Correction Factor for Anoxic Hydrolysis
-0,149785745	lami	1/m	Clean media filter coefficient
0,114813208	b_H	1/d	Decay Coefficient for Heterotrophic Biomass
-0,075529605	mu_H	1/d	Maximum Specific Growth Rate for Heterotrophic Biomass
-0,059370567	yy	--	coefficient y of Ives equation 1
0,056405363	zz	--	coefficient z of Ives equation 1
-0,04597675	rho	g/m3	Biofilm density
0,044823808	K_NO3_H	gN/m3	Saturation coeff of heterotrophs for NO3-
0,043167096	K_I_N2O	gN/m3	NO inhibition coefficient N2O
-0,038743058	K_S_NO2	gCOD/m3	Half-Saturation Coefficient for Heterotrophic Biomass NO2-
-0,03435073	ShK	--	coefficient K for number of Sherwood
-0,030895796	K_OH_NO3	gO2/m3	NO3- Oxygen Half-Saturation Coefficient for Heterotrophic Biomass
-0,030648872	Diffusion_Reduction	dUnit/dUnit	Reduction of diffusion coefficient in biofilm
-0,029517968	ShA	--	coefficient A for number of Sherwood
0,028477231	n_h2	--	Correction Factor for Anoxic Growth of Heterotrophs NO2-
0,025183397	K_O_AOB2	gO2/m3	AOB affinity constant for O2 HAO reactions

-0,025019854	BetaA	--	reduction factor of solubility in wastewater
-0,024997992	D_x	1/d	Diffusivity coefficient for particules components
-0,023568343	BAF_7d_max	m	Maximum biofilm thickness
-0,023389752	b_AOB	1/d	Decay Coefficient for Autotrophic Biomass (AOB)
-0,022108519	ShB	--	coefficient B for number of Sherwood
0,021688249	u_ExMediaW	1/d	Velocity of biofilm exchange of media when washing
0,02086102	Y_AOB	gCOD/gN	Yield for Autotrophic Biomass AOB
0,020764443	K_NO_AOB_HAO	gN/m3	AOB affinity constant for NO (HAO)
0,020469297	K_N2O_H	gN/m3	AOB Affinity constant for HNO2
0,020375379	K_O_AOB1	gO2/m3	AOB affinity constant for O2 AMO reactions
0,019796326	m	--	constant for diffuser and system configuration
-0,019736583	K_I_SNH	gN/m3	growth limitation by ammonium
0,019682293	K_I_O_AOB	gO2/m3	Inhibition constant by O2 on N2O production
0,019579447	n_nND	--	Reduction Factor for the ND pathway
0,018917744	theta_kla	--	Arrhenius parameter for kLa
-0,018512055	K_NO2_H	gN/m3	Saturation coeff of heterotrophs for NO2-
0,018055975	K_OH_N2O	gO2/m3	N2O Oxygen Half-Saturation Coefficient for Heterotrophic Biomass
-0,016764489	theta_mu_AOB	--	Temperature correction factor for mu_AOB
0,01613031	theta_mu_NOB	--	Temperature correction factor for mu_NOB
0,016022507	ShM	--	coefficient M for number of Sherwood
0,01593609	K_I_NO	gN/m3	NO inhibition coefficient NO
0,014442934	K_OH_NO	gO2/m3	NO Oxygen Half-Saturation Coefficient for Heterotrophic Biomass
-0,013039032	F_BOD_COD	--	Conversion factor BOD/COD
0,012935886	K_S_NO	gCOD/m3	Half-Saturation Coefficient for Heterotrophic Biomass NO
0,012669127	xx	--	coefficient x of Ives equation 1
-0,012613972	ShN	--	coefficient N for number of Sherwood
0,012606964	K_I_NO2	gN/m3	NO inhibition coefficient NO2-
0,01249979	f_P	--	Fraction of Biomass Converted to Inert Matter
0,012001678	K_OH	gO2/m3	Oxygen Half-Saturation Coefficient for Heterotrophic Biomass
-0,01188092	n_nNN	--	Reduction Factor for the NN pathway
-0,011715191	theta_b_NOB	--	Temperature correction factor for b_NOB
0,010894233	K_NO2_NOB	gN/m3	Half-saturation coefficient for NO2-,NOB
0,010839022	K_OH_NO2	gO2/m3	NO2- Oxygen Half-Saturation Coefficient for Heterotrophic Biomass
0,010068483	theta_k_a	--	Temperature correction factor for k_a
-0,009648722	K_NO_H	gN/m3	Saturation coeff of heterotrophs for NO
-0,009292596	FA	--	Fouling factor for aeration
0,008939662	i_X_BP	gP/gCOD	Mass of Phosphorus per Mass in COD Biomass added by JZ 06032017
-0,007919039	mu_NOB	1/d	Maximum Specific Growth Rate for Autotrophic Biomass NOB
0,007746423	K_NH_A	g/m3	Saturation coeff of autotrophs for ammonium
-0,006708084	k_a	m3/(gCOD*d)	Maximum Specific Ammonification Rate
0,006542163	K_NH2OH	gN/m3	AOB affinity constant for NH2OH
-0,006281263	theta_K_X	--	Temperature correction factor for K_X
-0,006066445	F_TSS_COD	--	Fraction TSS/COD
0,005712251	n_h4	--	Correction Factor For Anoxic Growth Of Heterotrophs N2O

0,005491002	i_X_B	gN/gCOD	Mass of Nitrogen per Mass of COD in Biomass
0,004854825	n_h3	--	Correction Factor for Anoxic Growth of Heterotrophs NO
-0,004086111	b_NOB	1/d	Decay Coefficient for Autotrophic Biomass (AOB)
-0,003542483	K_POH	g/m3	Half Saturation coefficient for phosphate for Heterotrophic Biomass
0,003320783	theta_b_AOB	--	Temperature correction factor for b_AOB
-0,003030133	K_O_AOB_ND	gO2/m3	AOB affinity constant for O2 ND pathway
-0,003014595	K_O_NOB	gO2/m3	NOB affinity constant for O2
0,002984203	Y_NOB	gCOD/gN	Yield for Autotrophic Biomass
-0,002980638	K_S	gCOD/m3	Half-Saturation Coefficient for Heterotrophic Biomass
-0,002916748	K_NH3_AOB	gN/m3	AOB affinity constant for NH3
-0,002807386	K_HNO2_AOB	gN/m3	Half Saturation coeff of heterotrophs for N2O
0,001781783	ShBeta	--	coefficient Beta for number of Sherwood
0,001771305	i_X_P	gN/gCOD	Mass of Nitrogen per Mass of COD in Products Formed
-0,001747412	K_NH_H	g/m3	Saturation coeff of heterotrophs for ammonium
0,001609758	K_NO_AOB_NN	gN/m3	AOB affinity constant for NO (NirK)
-0,001517532	K_S_N2O	gCOD/m3	Half-Saturation Coefficient for Heterotrophic Biomass N2O
-0,001503611	mu_AOB	1/d	Maximum Specific Growth Rate for Autotrophic Biomass AOB
0,001453616	u_ExMediaF	1/d	Velocity of biofilm exchange of media when filtering
0,001128718	K_X	gCOD/gCOD	Half Saturation Coefficient for Hydrolysis Of Slowly Biodegradable Substrate
-0,001035279	K_POA	g/m3	Half Saturation coefficient for phosphate for Autotrophic Biomass
-0,000659668	i_X_UP	gP/gCOD	Mass of Phosphorus per Mass of COD in Products Formed added by JZ 06032017

4 Programming Codes (Python) for the Hybrid Model Predictive Controller

#	Python Script	Application
1	**run_workflow_v2**	MPC & HMPC
2	'Coupling_python_west.py'	MPC & HMPC
2b	'run_NN_model_forloop_v2.py'	HMPC
3	'analyze_error_SP.py'	MPC & HMPC
4	'reinitialize_West_Reduced.py'	MPC & HMPC
5	'Prepare_BIG_Model.py'	MPC & HMPC
6	'Coupling_python_west_BIG_model_SD.py'	MPC & HMPC
7	'Calculate_new_setpoint_SSE.py'	MPC & HMPC
8	'Counter.py'	MPC & HMPC

Script 1 : run_workflow_v2

```
# this is the Beginning; this file will run all python scripts in consecutive order, according to the workflow type selected
# Choose between MPC (runs all scripts, but excludes run_NN_model.py) or HMPC (runs all scripts including run_NN_model.py)
# -*- coding: utf-8 -*-
"""
Created on Tue Apr 25 17:02:40 2023

@author: SERRAO-M & DANESHGAR-S
"""

import subprocess
import os
import time

path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2\\"
os.chdir(path)

program_list_mpc = ['create_scenarios_Mov_Time_v2.py', 'Coupling_python_west.py', 'analyze_error_SP.py', 'reinitialize_West_Reduced.py', 'Prepare_BIG_Model.py', 'Coupling_python_west_BIG_model_SD.py', 'Calculate_new_setpoint_SSE.py', 'Counter.py']
program_list_hmpc = ['create_scenarios_Mov_Time_v2.py', 'Coupling_python_west.py', 'run_NN_model_forloop_v2.py', 'analyze_error_SP.py', 'reinitialize_West_Reduced.py', 'Prepare_BIG_Model.py', 'Coupling_python_west_BIG_model_SD.py', 'Calculate_new_setpoint_SSE.py', 'Counter.py']

# Choose here: 'mpc' or 'hmpc'
mpc_type = 'mpc'
```

```
for i in range(0,100): # select here how many iterations, it includes the last nr. Each iteration simulates 15 minutes; 96 iterations = 24 hrs
   start_time = time.time()
   if mpc_type == 'mpc':
      for program in program_list_mpc:
         if os.path.isfile(program):
            subprocess.call(['python', program])
            print("Finished:" + program)
         #if program == 'reinitialize_West_Reduced.py':
                  #  print(time.time()-start_time, ' seconds to reach script for reinitialize_West_Reduced.py')
         else:
            alternative_directory = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\REALITY_P_M\\"
            alternative_path = os.path.join(alternative_directory, program)
            if os.path.isfile(alternative_path):
               subprocess.call(['python', alternative_path])
               print("Finished: " + alternative_path)
            else:
               print("Program not found: " + program)
   elif mpc_type == 'hmpc':
      for program in program_list_hmpc:
         if os.path.isfile(program):
            subprocess.call(['python', program])
            print("Finished :" + program)
         else:
            alternative_directory = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\REALITY_P_M\\"
            alternative_path = os.path.join(alternative_directory, program)
            if os.path.isfile(alternative_path):
               subprocess.call(['python', alternative_path])
               print("Finished: " + alternative_path)
            else:
               print("Program not found: " + program)

   print()
   print("Done! Executed all scripts in the workflow: " + mpc_type)

   with open('counter.txt', "r") as f2:  # reading the counter
      j = int(f2.read())
      print("This was iteration number: " + str(j-1)) # subtract 1, because at this stage, the counter has already been updated with +1

   with open("opt_value_1.txt", "r") as f3:
      p = int(f3.read())
      print("The optimal Q_Dose =  " + str(p))
      #print()
   with open("Q_Dose_Reality.txt", "r") as f3:
      q = float(f3.read())
      print("The optimal Q_Dose in Reality Model =  " + str(q))
      print()
```

```
      print('Time it took: ', time.time()-start_time, ' seconds, which is equal to ', (time.time()-start_time)/60, '
minutes to complete this iteration')
   print('----------------------------------------------------------')
```

Script 2 : 'create_scenarios_Mov_Time_v2.py'

```
# This script creates WEST simulation .xml files, based on the values noted in parameter.csv
# -*- coding: utf-8 -*-
"""
Created on Tue May  9 15:52:10 2023

@author: SERRAO-M
"""

# This script creates WEST simulation .xml files, based on the values noted in parameter.csv

import os
import pandas as pd
from lxml import etree
# the lxml is not installed on your system by default, you can install it by running "pip install lxml" in a
command line terminal


# path to the project folder where you have ObjEval.Exp.xml files
# make sure to end you path with \\ like "C:\\Program Files\\project\\"
path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2\\"
# path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_A\\"
# path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_B\\"
# path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_C\\"

file_name = 'BAFNIT16.Dynamic.ObjEval.Exp.xml'  # in the next iteration, this file will be replaced by
the xml file identified at the end of script reinitialize_West_Reduced.py


original_xml = os.path.join(path, file_name)
tree = etree.parse(original_xml)
root = tree.getroot()
Quantities = root[0][1][0][0][1][0][0]


org_start_time = 0
org_stop_time = 120/60/24 # 120/60/24 = 2 hrs

with open(path + 'counter.txt', "r") as f2:
```

```
    j = int(f2.read())
    j = j*15/60/24

time_settings = root[0][1][0][0][1][3]

T1_name = ".Timer_2.T1"
T5_name = ".Timer_2.T5"

# read the parameter file that is stored in the same folder as the project and create scenario files
parameters = pd.read_csv(path + 'parameters.csv', index_col='scenario')
for scenario in parameters.index:
    for q in time_settings.iter('Prop'):
        if q.get('Name') == 'StartTime':
            q.set('Value', str(org_start_time + j))
            #print('Start Time is:', new_start_time,'(days)')
            # Convert to float and format with 2 decimal places
            formatted_start_time = '{:.6f}'.format(float(org_start_time+j))
            print(f'Start Time is: {formatted_start_time} (days)')

        elif q.get('Name') == 'StopTime':
            q.set('Value', str(org_stop_time + j))
            # print('Stop Time is:', new_stop_time,'(days)')
            # Convert to float and format with 2 decimal places
            formatted_stop_time = '{:.6f}'.format(float(org_stop_time+j))
            print(f'Stop Time is: {formatted_stop_time} (days)')
    for name in parameters.columns:
        param_in_file = 0
        for q in Quantities.iter('Quantity'):
            param_name_in_file = q.get('Name')
            if ((not param_in_file) & (param_name_in_file == name)):
                q.set('Value', str(parameters[name][scenario]))
                param_value = q.get('Value')
                param_in_file = 1
        if (not param_in_file):
            new_param = etree.SubElement(Quantities, 'Quantity', dict(Name=name))
            new_param.set('Value', str(parameters[name][scenario]))
        for q in Quantities.iter('Quantity'):
            if q.get('Name') == T1_name:
                q.set('Value', str(org_start_time + j + 15/60/24))
            if q.get('Name') == T5_name:
                q.set('Value', str(org_stop_time + (j*2)))

    for output in root.iter('Output'):
        if output.get('Name') == "File_10":
            for var in output.iter('File'):
```

```
            var.set('Name', 'scen_'+str(scenario+1)+'_'+'.out.txt')

    tree.write(path+'scen_'+str(scenario+1)+'_'+file_name, encoding='UTF-8', xml_declaration=True)

print()
print('Done! '+str(scenario+1) +' scenario xml files created in '+ path)
```

Script 2 : 'Coupling_python_west.py'

```
# This script executes each scenario xml files defined in script create_scenarios.py,
# by running it through the WEST MPC model and stores output files per scenario as: scen_k_.out.txt
"""
@author: SERRAO-M & DANESHGAR-S
"""
#!/usr/bin/env python
# coding: utf-8

# In[ ]:

import matplotlib.pyplot as plt
import numpy as np
import os
import pandas as pd
import subprocess

# In[23]:

# note: don't change any double backslashes (\\) to slashes (\) when you copy your local paths
instead of the ones here
# so if your local path is C:\model\project in this file you should put it as C:\\model\\project

# this is the path to the west installation files, you should change this to your the installation
directory on your machine if it's not the same
path_to_west = "C:\\Program Files (x86)\\DHI\\2016"

# this is the path to the west project, the batch file should be created in the same directory as model
files
#path_to_project = "C:\\Users\\sdaneshg\\OneDrive - UGent\\Desktop\\WEST_model_latest
update\\WEST_model\\WEST\\data\\projects\\Influent_model\\Influent"
path_to_project =
"C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2"
# path_to_project = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_A\\"
# path_to_project = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_B\\"
# path_to_project = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_C\\"
```

```
# the name of the log file to save and the name of the ObjEval file to run the model
# these two will be used only in texec command

# this will read all .Dynamic.ObjEval.Exp.xml files in your directory and make a list out of their names
# useful if you have multiple of them for more scenarios
# the assumptption is that you have already gone through the procedure for making
.Dynamic.ObjEval.Exp.xml files
# for scenarios based on the parameters that need to be changed
exp_filenames = []
for files in os.listdir(path_to_project):
   if files.startswith("scen") and files.endswith(".Dynamic.ObjEval.Exp.xml"):
         exp_filenames.append(files)

# print to see if it works: it's OK
print()
print("The following scenario xml files are used: ", exp_filenames)

# here the batch file will be created
windows_batch_file = os.path.join(path_to_project, "run_model.bat")

# open the batch file and write the lines inside it
fbat = open(windows_batch_file, "w")
fbat.write("SET INCLUDE=%INCLUDE%;"+path_to_west+"\\WEST\\include\n")
fbat.write("SET LIB=%LIB%;"+path_to_west+"\\WEST\\lib\\win32-msvc\n")
fbat.write("SET
PATH="+path_to_west+"\\bin;"+path_to_west+"\\WEST\\third_party\\bcc5.5\\Bin;%PATH%\n")
fbat.write("SET TORNADO_CC_PLATFORM=win32-bcc5.5\n")
fbat.write("SET TORNADO_CC_PATH="+path_to_west+"\\WEST\\third_party\\bcc5.5\n")
fbat.write("SET TORNADO_DATA_PATH=C:\\Users\\Public\\Documents\\DHI\\WEST\n")
fbat.write("SET TORNADO_ROOT_PATH="+path_to_west+"\\WEST\n")
fbat.write("SET OPENMODELICAHOME="+path_to_west+"\\bin\n")

# up to here, the previous lines are for your machine to know where tornado files are stored

# here you can then add a command for tbuild or texec

# In[3]:

##########################################################################
# read this part only if you need to use tbuild command
# for creating model files to run on linux (like for running on a cluster)
##########################################################################

# example for tbuild command, xxx is the path to the model block library,
# you should keep External\External at the end of this path, just change the previous part
```

```
# for example it should be something like this: C:\example\blocks\MSL\External\External
# _ProjectName_ is only the name of the project and not the complete path, the complete
# path is not required since you are already in the directory of the project

# fbat.write("tbuild -S C:\\Users\\sdaneshg\\OneDrive - UGent\\Desktop\\WEST_model_latest
update\\WEST_model\\WEST\\data\\blocks\\WESTforKALLISTO - InfluentModel\\External\\External
-B "" _Influent_ > tbuild.log\n")

# In[24]:

##############################################################################
# this part is related to using texec for running simulations
##############################################################################

# example for texec command

# EXPLANATION if you need to run tbuild first
# when you first run this batch script, keep this line as a comment
# you need to run the tbuild first
# after you run this with tbuild command successfully then open this file again
# and comment the fbat.write("tbuild...") line and uncomment the line below
# then run the script again to create batch file fo simulations

# log_filename is the name of the log file (with .log extension) that you choose to save
# exp_filename is the .Dynamic.ObjEval.Exp.xml file you have in your project directory
# these two should be only names and not full path, you don't need the full path here
# output_name_orig is the name of the output file you choose to save in WEST environment
# this should work even if you have multiple ObjEval files for example:
scen_1.Dynamic.ObjEval.Exp.xml, scen_2.Dynamic.ObjEval.Exp.xml , ...
# this will create a texec command for each scenario to run and then rename the output file
accordingly
# so they will not be overwritten
# this is not probably the most efficient way to do this on windwos but it should work

output_name_orig = "Methanol_Dosing_Control.txt"

for f in exp_filenames:
    log_filename = f.split('.')[0] + ".log"
    new_output_name = f.split('.')[0] + ".out.txt"
    fbat.write("texec -l \"" + log_filename  + "\" \"" + f+"\"\n")
  #  fbat.write("ren \"" + output_name_orig + "\" \"" + new_output_name +"\"\n")

# write exit and close the file
fbat.write("EXIT")
fbat.close()
```

```
# # Outputfile importeren

# In[156]:

#output_name=path_to_project + "\\" + new_output_name

#output_data=pd.read_csv(output_name, sep="\t",header=[0])
#output_data=output_data.drop(0)
#output_data=output_data.set_index('#.t')
#display(output_data)

# In[157]:

#output_data.plot()
#print(output_data.dtypes)

# In[164]:

#output_data['.ASU_2.X_TSS']=output_data['.ASU_2.X_TSS'].astype(float)
#output_data['.ASU_4.C(S_O)']=output_data['.ASU_4.C(S_O)'].astype(float)
#output_data['.Qair_Basin_4']=output_data['.Qair_Basin_4'].astype(float)
#output_data['.SST_1.X_Out']=output_data['.SST_1.X_Out'].astype(float)
#output_data['.SST_1.X_Under']=output_data['.SST_1.X_Under'].astype(float)
#print(output_data.dtypes)
#output_data.iplot()

# In[168]:

#from statsmodels.tsa.seasonal import seasonal_decompose
#results=seasonal_decompose(output_data['.ASU_2.X_TSS'])
#results.plot()

os.chdir(path_to_project)
p = subprocess.Popen("run_model.bat")
#,shell=True, stdout=subprocess.PIPE, strerr=subprocess.PIPE
p.communicate()

print('Done! Scenarios executed and output files stored.')
```

Script 2b: 'run_NN_model_forloop_v2.py

```
# -*- coding: utf-8 -*-
"""
```

```
Created on Tue May 16 19:00:43 2023

@author: SERRAO-M
"""
import os
import pandas as pd
import pickle
import numpy as np
from sklearn.preprocessing import MinMaxScaler
# from tensorflow.keras.models import Sequential
# from tensorflow.keras.layers import Dense
# from tensorflow.keras import regularizers
# import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from tqdm import tqdm

# path to the project folder where you have ObjEval.Exp.xml files
# make sure to end you path with \\ like "C:\\Program Files\\project\\"
path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2\\"
# path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_A\\"
# path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_B\\"
# path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_C\\"

# change directory
os.chdir(path)



with open('counter.txt', "r") as f2:  # reading the counter
   j = int(f2.read())
   k = j*15

# Count the number of files that match the pattern
count = sum(1 for file in os.listdir(path) if file.startswith('scen_') and file.endswith('_.out.txt'))



# start the for loop
# for file in os.listdir('.'):
for i, file in enumerate(tqdm(os.listdir(path))):
   if file.startswith('scen_') and file.endswith('_.out.txt'): # example: scen_1_.out.txt
      file_num = file.split('_')[1]  # extract the file number from the file name
      dfnew = pd.read_csv('new_data1min.txt', sep='\t')  # , skiprows=[0]) # read the file with features
data, but we don't need the row 1 with headers
      dfnew = dfnew[k:k+120] # read rows 0 to 120 (= 2hrs) (+ k each iteration; and k = j *15) # dfnew
= dfnew[j:j+120] # read rows 0 to 120 (= 2hrs) (+ j each iteration)
      dfnew.reset_index(inplace=True)
```

```
    print()
    print(f"Reading file {file} with file number {file_num}")

    # Load the trained NN model
    with open('model.pickle', 'rb') as f:
      model = pickle.load(f)

    # Define the predictor and response variable names
    predictor_names = ['WEEK', 'Temperature', 'Q_SW_BATC', 'Q_REC_BATC', 'NOx_IN_BATC',
'COD_IN_BATC', 'NOx_REC', 'TSS_SW', 'MeOH_INJ', 'NO3_SP', 'Weather', 'NO3_OUT_MOD']
    response_variable = 'ERROR_NO3'

    # In dfnew, replace the column 'NO3_OUT_MOD' with the NO3 predicted data from the West
reduced model
    # Load the predicted data from the file scenario file:
    filename_pred = file
    dfwest_pred = pd.read_csv(filename_pred, sep='\t', skiprows=[1])  # because we don't need the
row 2 with units

    # Get the column '.BAF_7.NO3_OUT' and in the dfnew overwrite the NO3_OUT_MOD column
with the values from '.BAF_7.NO3_OUT'
    dfnew['NO3_OUT_MOD'] = dfwest_pred['.BAF_7.NO3_OUT']
    # add a new row with index 0 and a value of g/m3 in the 'NO3_OUT_MOD' column
    #dfnew.loc[0, 'NO3_OUT_MOD'] = 'g/m3'
    # reset the index to ensure it starts at 0
    #dfnew = dfnew.reset_index(drop=True)
    # create instances of the StandardScaler and MinMaxScaler classes
    scaler_zscore = StandardScaler()
    minmax_scaler = MinMaxScaler(feature_range=(0, 1))

    # apply z-score to the DataFrame
    dfnew_z = pd.DataFrame(scaler_zscore.fit_transform(dfnew), columns=dfnew.columns,
index=dfnew.index)

    # apply min-max scaling to the DataFrame
    dfnew_scaled = pd.DataFrame(minmax_scaler.fit_transform(dfnew_z), columns=dfnew.columns,
index=dfnew.index)

    #print(dfnew_scaled)

    # make prdictions using the pretrained ANN model
    y_new = model.predict(dfnew_scaled[predictor_names]).flatten()
    print(y_new)

    df_pred = pd.DataFrame.copy(dfnew_scaled) # make a copy of the response
```

```
df_pred['ERROR_NO3'] = y_new  # here, we're adding the response variable as a 13th column to a copy (df_pred) of the existing dataframe dfnew_scaled

# Apply this to the y_new prediction:
# Reverse the scaling to obtain the denormalized data
dfpred_denorm_minmax = np.copy(df_pred)
dfpred_denorm_minmax = minmax_scaler.inverse_transform(dfpred_denorm_minmax)
#dfpred_denorm_minmax = minmax_scaler.inverse_transform(dfpred_denorm_minmax.reshape(-1, 1))
# Reverse the normalization to obtain the original values
dfpred_denorm_zscore = np.copy(dfpred_denorm_minmax)
# a = dfpred_denorm_zscore
# a_reshaped = a.reshape(624,)
dfpred_denorm_zscore = scaler_zscore.inverse_transform(dfpred_denorm_zscore)

df_pred_final = pd.DataFrame(dfpred_denorm_zscore, columns=dfnew.columns, index=dfnew.index)

# Print the original, normalized, and denormalized data
# print("Original data:")
# print(dfnew)
# print("\nNormalized data (z-score):")
# print(dfnew_z)
# print("\nNormalized data (z-score and min-max scaling):")
# print(dfnew_scaled)
# print("\nDenormalized data (min-max scaling only):")
# print(dfpred_denorm_minmax)
# print("\nDeNormalized data (min-max scaling and z-score):")
# print(dfpred_denorm_zscore)


# NO3_pred_HM = dfnew['NO3_OUT_MOD'] + dfnew['PREDICTED_ERROR_NO3']
# dfnew['NO3_pred_HM'] = NO3_pred_HM

NO3_pred_HM = dfwest_pred['.BAF_7.NO3_OUT'][:-1] + df_pred_final['ERROR_NO3']
# NO3_pred_HMb = NO3_pred_HM[k:k+120]

#=====================================================
# To save the NO3_pred_HM in the scenario output file:
#=====================================================

# Read the data from the scenario output file
data = pd.read_csv(file, delimiter='\t')

# Replace the values in the column
```

```
        #data['.BAF_7.NO3_OUT'].iloc[2:] = NO3_pred_HM
        data['.BAF_7.NO3_OUT'].iloc[0] ='g/m3'
        length = len(NO3_pred_HM)
        data['.BAF_7.NO3_OUT'].iloc[len(data['.BAF_7.NO3_OUT'])-length:] = NO3_pred_HM  # copy
from row 2, so that t = 0 is not predicted  data['column'].iloc[1:] = NO3_pred_HM[:len(data)-1]

        # Write the updated data back to the file
        data.to_csv(file, sep='\t', index=False)

        # Update the progress bar
        tqdm.write(f"Processed {i + 1}/{count} files.")

        print()
        print('Done! Scenario output file', filename_pred,' updated with NO3_pred_HM')
print()


print('We are done with the for loop and have updated', count, 'scen_k_out.txt files. Have a nice
day!')
```

Script 3: 'analyze_error_SP.py

```
# This script calculates the error SSE and relative ME between SETPOINT - Predicted for each scenario
output file stored in the previous script Couping_python_west.py and
# then indicates which scenario has the lowest SSE and the corresponding U_optimal_1 value of
methanol dosing for first time step (S_F in mg/L)

# -*- coding: utf-8 -*-
"""
Created on Fri Mar 10 15:13:22 2023

@author: SERRAO-M
"""

import os
import math
import numpy as np
import pandas as pd
import shutil

# Read observed variable:
filepath =
"C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2\\"
# filepath = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_A\\"
```

```
# filepath = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_B\\"
# filepath = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_C\\"

filename = "NO3_setpoint.txt"

with open('counter.txt', "r") as f2:  # reading the counter
   j = int(f2.read())

# Read the values from the "NO3_setpoint.txt file into a DataFrame
df_setpoint = pd.read_csv('NO3_setpoint.txt', header=None, names=['NO3_setpoint'])

# Display the DataFrame
# print(df_setpoint)

# Convert the values to a list of floats
NO3_sp = df_setpoint['NO3_setpoint'].values.tolist()

# Display the list of floats
# print(NO3_sp)

# with open(os.path.join(filepath, filename), "r") as f:
#     # Find the column number for the header we're looking for
#     headers = f.readline().strip().split("\t")
#     header_index = headers.index(header_to_find)

#     # Read in the data from the file
#     data = []
#     for line in f:
#         data.append(line.strip().split("\t"))

#     # Extract the data from the specified column
#     NO3_obs = [float(row[header_index].replace(",", ".")) for row in data[1:]]


# create a dictionary to store the results for each scenario that was run in previous script
stats_scores = {}

# loop through all the files in the working directory and select scenario output files
for filename in os.listdir('.'):
   if filename.startswith('scen_') and filename.endswith('.out.txt'):
      # open the file
         with open(filename, 'r') as f:
            # read the lines of the file
            lines = f.readlines()
            # find the index of the '.BAF_7.NO3_OUT' column header in the first row
```

```
header_index = lines[0].split().index('.BAF_7.NO3_OUT')

# extract the NO3_OUT data and the observed NO3_OUT data starting from row
NO3_pred = []
for line in lines[2:]:
   row = line.split()
   NO3_pred.append(float(row[header_index]))

   # Calculate statistical scores between the two variables:

   # find the minimum length of the two input lists
   min_len = min(len(NO3_sp), len(NO3_pred))

   # truncate the lists to the minimum length
   NO3_sp_trunc = NO3_sp[1:min_len]
   NO3_Pred_trunc = NO3_pred[1:min_len]

   # calculate the mean of NO3_sp and NO3_Pred
   mean_NO3_sp = np.mean(NO3_sp_trunc)
   mean_NO3_Pred = np.mean(NO3_Pred_trunc)

   # calculate the mean error between NO3_sp and NO3_Pred
   mean_error = mean_NO3_sp - mean_NO3_Pred

   # calculate the mean error / mean of NO3_sp
   relative_error = mean_error / mean_NO3_sp

   # calculate the sum of squared error between NO3_sp and NO3_Pred
   squared_error = np.sum((np.array(NO3_sp_trunc) - np.array(NO3_Pred_trunc)) ** 2)

   # Calculate RMSE
   root_mean_square_error = math.sqrt(squared_error)
   # print("RMSE: " + str(root_mean_square_error))

   # store the results in the dictionary
   stats_scores[filename] = {
   'mean_NO3_pred': mean_NO3_Pred,
   'mean_NO3_sp': mean_NO3_sp,
   'ME': mean_error,
   'RE': relative_error,
   'SSE': squared_error,
   'RMSE': root_mean_square_error,
   }
    # print some statistics for the extracted data
```

```
            # print(f"{filename}: mean_NO3_pred={mean_NO3_pred:.2f},
mean_NO3_obs={mean_NO3_obs:.2f}, ME={ME:.2f}, RE={RE:.2f}, SSE={SSE:.2f}")

# print the results dictionary
#print(stats_scores)
#print(stats_scores, round(stats_scores, 5))  # to round the numbers to 5 decimals ; but this doesn't
work with dictionary
#my_dict = {'a': 1.2345, 'b': 2.3456, 'c': 3.4567}
#rounded_dict = {k: round(v, 2) for k, v in stats_scores.items()}
#print(rounded_dict)

# print("Mean of NO3_Obs:", round(mean_NO3_Obs, 5))
# print("Mean of NO3_Pred:", round(mean_NO3_Pred, 5))
# print("Mean error between NO3_Obs and NO3_Pred:", round(mean_error, 5))
 # print("Relative error between NO3_Obs and NO3_Pred:", round(relative_error, 5))
 # print("Sum of squared error between NO3_Obs and NO3_Pred:", round(squared_error, 5))

# find the key with the lowest SSE score
lowest_key = None
lowest_score = float('inf')

for key, value in stats_scores.items():
   if value['SSE'] < lowest_score:
      lowest_key = key
      lowest_score = value['SSE']

# extract k from the lowest_key
k = int(lowest_key.split('_')[1])
# Save the value of k to a text file
with open("k.txt", 'w') as f:
   f.write(str(k))

print()
print(f"Scenario with the lowest SSE score: {lowest_key}, SSE: {lowest_score:.2f}")
print(f"k value (= scenario number) for lowest_key: {k}")

# Define the filename based on the value of lowest_key
filename_opt = f"{lowest_key}"

# Load the file into a pandas DataFrame
data = pd.read_csv(filename_opt, delimiter="\t")

# To know the Optimal Value, extract the value in row 2 of the column named '.MeOH_1.Q_Dose'
# opt_value_1 = data.loc[1, ".MeOH_1.Outflow(S_S)"]
```

```
opt_value_1 = data.loc[1, ".MeOH_1.Q_Dose"] # retrieves the value at row 1 and column ".MeOH_1.Q_Dose".

with open("opt_value_1.txt", 'w') as my_file:
    my_file.write(opt_value_1) # add the "k" variable that corresponds to the opt_value_1 ; I didn't do this here; I have k stored as a variable instead

# Print the extracted value
# print("Optimal value U_opt_1 =" , opt_value_1, "(S_F in [mg/L])\n")
print("Optimal value U_opt_1 =" , opt_value_1, "Q_Dose [m3/day])\n")
print('Done! Statistical errors calculated and optimal dosing value U_opt_1 identified.')

# Saving the output file of the optimal scenario under new filename that includes ITER + {j}
new_filename = f"Best_{k}_ITER_{j}_.out.txt"

# Saving the file under the new name
shutil.copyfile(filename_opt, new_filename)

print(f"File '{filename_opt}' saved as '{new_filename}'.")
print()

# now prepare the Q_Dose_Reality_Model

def read_jth_row(file_path, j):
    with open(file_path, 'r') as file:
        lines = file.readlines()
        if j >= 0 and j < len(lines):
            row = lines[j].strip()
            return row
        else:
            return None

# Read the value of j from counter.txt
with open('counter.txt', 'r') as counter_file:
    j = int(counter_file.read())

# Read the j-th row from the file containing numbers
numbers_file_path = 'Active_Filters_PreDN_0106_09062020.txt'
Act_Filters = read_jth_row(numbers_file_path, j)

# Output the result
print(f"The {j}-th row of Active Filters is: {Act_Filters}")

# now, multiply opt_value_1 with Act_Filters and save it under ,
```

```
# Read the value of opt_value_1

# Multiply jth_row with opt_value_1
if Act_Filters is not None:
   result = float(Act_Filters) * float(opt_value_1)
else:
   result = None

# Save the result in Q_Dose_Reality.txt
if result is not None:
   with open('Q_Dose_Reality.txt', 'w') as output_file:
      output_file.write(str(result))
      print("Result saved in Q_Dose_Reality.txt")
else:
   print(f"Invalid row number: {j}")
```

Script 4: reinitialize_West_Reduced.py

```
# This script finds the scen_k.xml input file that corresponds to the best fit among all the scenario xml files defined in script create_scenarios.py,
# then runs the WEST reduced model with it and stores output files per scenario as: scen_k_.out.txt

# -*- coding: utf-8 -*-
"""
Created on Tue Apr  4 18:28:40 2023

@author: SERRAO-M
"""

# placeholder for script 2B.py

# take variable filename from script analyze_error_3.py
# run it through WEST_Reduced model so that it has the correct scenario output as end_values for the next loop

import glob
import matplotlib.pyplot as plt
import numpy as np
import os
import pandas as pd
import subprocess

# path to the project folder where you have ObjEval.Exp.xml files
# make sure to end you path with \\ like "C:\\Program Files\\project\\"
path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2\\"
```

```
# path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_A\\"
# path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_B\\"
# path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_C\\"

# change directory
os.chdir(path)

# In[23]:

# note: don't change any double backslashes (\\) to slashes (\) when you copy your local paths
instead of the ones here
# so if your local path is C:\model\project in this file you should put it as C:\\model\\project

# this is the path to the west installation files, you should change this to your the installation
directory on your machine if it's not the same
path_to_west = "C:\\Program Files (x86)\\DHI\\2016"

# this is the path to the west project, the batch file should be created in the same directory as model
files
#path_to_project = "C:\\Users\\sdaneshg\\OneDrive - UGent\\Desktop\\WEST_model_latest
update\\WEST_model\\WEST\\data\\projects\\Influent_model\\Influent"
path_to_project =
"C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2"
# path_to_project = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_A\\"
# path_to_project = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_B\\"
# path_to_project = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\MPC_C\\"

# the name of the log file to save and the name of the ObjEval file to run the model
# these two will be used only in texec command

# this will read all .Dynamic.ObjEval.Exp.xml files in your directory and make a list out of their names
# useful if you have multiple of them for more scenarios
# the assumptption is that you have already gone through the procedure for making
.Dynamic.ObjEval.Exp.xml files
# for scenarios based on the parameters that need to be changed

# from analyze_error_3 import filename_opt  # watch out! this import function will rerun the script!

# read txt file opt_value_1.txt
# take k and store in variable k

# Load the value of k from the k.txt file
with open("k.txt", 'r') as f:
    k = int(f.read())
```

```
# Find the file in the current directory that matches the pattern
"scen_{}_BAFNIT16.Dynamic.ObjEval.Exp.xml"
xml_files = glob.glob("scen_{}_BAFNIT16.Dynamic.ObjEval.Exp.xml".format(k))

# If no files were found, print an error message and exit
if len(xml_files) == 0:
   print("Error: No XML files found for k={}".format(k))
   exit()

# Use the first file in the list
xml_filename = xml_files[0]
filename_opt = xml_filename

#filename_opt = name of xml ;  eg: scen_1_BAFNIT16.Dynamic.ObjEval.Exp.xml
# filename_opt = "scen_1_BAFNIT16.Dynamic.ObjEval.Exp.xml"
# read the xml file from directory

# Use the filename variable in script analyze_error_3.py
# with open(filename_opt, 'r') as a:
#     Opt_Scenario_Filename = a.read()

# Process the data as needed
# print(Opt_Scenario_Filename) # attenion: prints complete file! , head(5) does not work!

# print to see if it works: it's OK
print()
print("The following scenario xml file is used: ", filename_opt)
print("The following scenario xml file is used: ", xml_filename)

# here the batch file will be created
windows_batch_file = os.path.join(path_to_project, "run_model_opt.bat")

# open the batch file and write the lines inside it
fbat = open(windows_batch_file, "w")
fbat.write("SET INCLUDE=%INCLUDE%;"+path_to_west+"\\WEST\\include\n")
fbat.write("SET LIB=%LIB%;"+path_to_west+"\\WEST\\lib\\win32-msvc\n")
fbat.write("SET
PATH="+path_to_west+"\\bin;"+path_to_west+"\\WEST\\third_party\\bcc5.5\\Bin;%PATH%\n")
fbat.write("SET TORNADO_CC_PLATFORM=win32-bcc5.5\n")
fbat.write("SET TORNADO_CC_PATH="+path_to_west+"\\WEST\\third_party\\bcc5.5\n")
fbat.write("SET TORNADO_DATA_PATH=C:\\Users\\Public\\Documents\\DHI\\WEST\n")
fbat.write("SET TORNADO_ROOT_PATH="+path_to_west+"\\WEST\n")
fbat.write("SET OPENMODELICAHOME="+path_to_west+"\\bin\n")

# up to here, the previous lines are for your machine to know where tornado files are stored
```

```
# here you can then add a command for tbuild or texec

# do we need a texec file? YES we do, because without it, the bat file at the end of the script wil not know that you want to execute a model run
# for f in Opt_Scenario_Filename: # for loop that iterates through each item in the Opt_Scenario_Filename list, and assigns it to the variable f on each iteration.
#     log_filename = f.split('.')[0] + ".log" # splits the filename f on the '.' character, and takes the first element of the resulting list (which is the part of the filename before the '.'), and adds the string ".log" to the end. This creates a new string that represents the name of a log file based on the original filename.
#     new_output_name = f.split('.')[0] + ".out.txt" # creates a new string that represents the name of an output file.
#     fbat.write("texec -l \"" + log_filename  +  " -cp \" \"" + 'BAFNIT16.Dynamic.ObjEval.Exp.xml' + f+"\"\n") # writes a string to a file called fbat. The -l option specifies that texec should write its output to the log file, and the -cp option specifies that the output should be copied to the output file. It basically saves the end values of the current run in a new file named 'End_Values.' + overwrites the xml file with the end_values as initial values
##  fbat.write("texec -l \"" + log_filename  +  "-cp \" \"" + 'LINEBiofriltration.Dynamic.ObjEval.Exp.xml' + f+"\"\n") # saves the end values of the current run in a new file named 'End_Values.' + overwrites the xml file with the end_values as initial values

with
open('C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2\\counter.txt', "r") as f2:  # reading the counter
   j = int(f2.read())

log_filename = filename_opt.split('.')[0] + ".log" # splits the filename f on the '.' character, and takes the first element of the resulting list (which is the part of the filename before the '.'), and adds the string ".log" to the end. This creates a new string that represents the name of a log file based on the original filename.
new_output_name = filename_opt.split('.')[0] + ".out.txt" # creates a new string that represents the name of an output file.
#new_output_name = filename_opt.split('.')[0] + 'Iter_' + str(j) + ".out.txt" # creates a new string that represents the name of an output file.
fbat.write("texec -l \"" + log_filename  +  " -cp \" \"" + 'BAFNIT16.Dynamic.ObjEval.Exp.xml' + filename_opt+"\"\n") # writes a string to a file called fbat. The -l option specifies that texec should write its output to the log file, and the -cp option specifies that the output should be copied to the output file. It basically saves the end values of the current run in a new file named 'End_Values.' + overwrites the xml file with the end_values as initial values
# writes a string to a file called fbat. The -l option specifies that texec should write its output to the log file, and the -cp option specifies that the output should be copied to the output file. It basically saves the end values of the current run in a new file named 'End_Values.' + overwrites the xml file with the end_values as initial values
```

```
# write exit and close the file
fbat.write("EXIT")
fbat.close()

##===== explaination by Saba:==========================================
# the following commands at the end are for running the bat file.
# It first checks if there is a texec command in the bat file that states to run the model.
# If there's no texec command, the bat file will be run but there is no command inside it to tell the
tornado to run the model; so the model does NOT run.
##=====================================================================
os.chdir(path_to_project)
p = subprocess.Popen("run_model_opt.bat")
#,shell=True, stdout=subprocess.PIPE, strerr=subprocess.PIPE
p.communicate()

print('Done!' , filename_opt, ' executed and output file stored.')
```

Script 5: 'Prepare_BIG_Model.py

```
# -*- coding: utf-8 -*-
"""
Created on Tue Apr 18 18:51:40 2023

@author: SERRAO-M
"""

# This script prepares the input files for the large WEST model with Start Time and Stop Time and the
Opt_value_1 (Q_Dose [m3/d]) as indicated below
# and reads the opt_value_1 as the Q_Dose into the methanol dosing submodel MeOH_1 in the Big
model

import os
import pandas as pd
from lxml import etree # the lxml is not installed on your system by default, you can install it by
running "pip install lxml" in a command line terminal
import xml.etree.ElementTree as ET

# path to the project folder where you have ObjEval.Exp.xml files
# make sure to end you path with \\ like "C:\\Program Files\\project\\"
path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\REALITY_P_M\\"
path_to_reducedModel =
"C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2\\"
# change directory
os.chdir(path)
```

```
# Set the name of the file
file_name = 'LINEBiofriltration.Dynamic.ObjEval.Exp.xml' #spelling error in "filtration" is correct

# Get optimal value
# with
open("C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2\\opt_value_1.txt", 'r') as my_file:
  # opt_value_1 = my_file.read()
with
open("C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2\\Q_Dose_Reality.txt", 'r') as my_file:
      opt_value_Reality = my_file.read()

original_xml = os.path.join(path, file_name)
tree = etree.parse(original_xml)
root = tree.getroot()
Quantities = root[0][1][0][0][1][0][0]

# opt_value_Reality = str(240) = to test
opt_value_Reality = str(opt_value_Reality) # set variable into integer to allow for absolute value to be taken
# opt_value_1 = abs(opt_value_1) * 10 # to convert from S_S in [mg/L] to S_S in [g/d] we multiply with 10 m3/day of H2O flow rate
# not needed if opt_value = Q (m3/d)

print()

org_start_time = 0
org_stop_time = 15/60/24

with open(path_to_reducedModel + 'counter.txt', "r") as f2:
   j = int(f2.read())
   j = j*15/60/24

time_settings = root[0][1][0][0][1][3]

for q in time_settings.iter('Prop'):
   if q.get('Name') == 'StartTime':
      q.set('Value', str(org_start_time + j))
      #print('Start Time is:', new_start_time,'(days)')
      # Convert to float and format with 2 decimal places
      formatted_start_time = '{:.6f}'.format(float(org_start_time+j))
      print(f'Start Time is: {formatted_start_time} (days)')

   elif q.get('Name') == 'StopTime':
```

```
        q.set('Value', str(org_stop_time + j))
        # print('Stop Time is:', new_stop_time,'(days)')
        # Convert to float and format with 2 decimal places
        formatted_stop_time = '{:.6f}'.format(float(org_stop_time+j))
        print(f'Stop Time is: {formatted_stop_time} (days)')

# add the Opt_value_1 (= Q_Dose [m3/d]) into the line of the LINE.xml file
# that corresponds to the new MeOH_1 dosing submodel that is inserted in the BIG WEST model
for q in Quantities.iter('Quantity'):
    if q.get('Name') == '.MeOH_1.Q_Dose':
        q.set('Value', opt_value_Reality)

# every time you change something in the xml file you should use the tree.write to save the changes
into the file otherwise it remains only within the python env.
tree.write(path + file_name)
print("Path:", path)
print("The following xml file has been updated:", file_name)
```

Script 6: 'Coupling_python_west_BIG_model_SD.py

```
# This script executes the BIG_WEST model according to the Input variables (Q_Dose, StartTime,
StopTime) defined in the LINE...xml file updated in previous script Input_BIG_Model.py
# It runs the LINE...xml file through the BIG WEST model and saves all the end-values in a new file
named "End_Values.Dynamic.ObjEval.Exp.xml"
# # saves the end values of the current run in a new file named 'End_Values.' + overwrites the xml file
with the end_values as initial values

# -*- coding: utf-8 -*-
"""
Created on Thu Mar  9 15:33:47 2023

@author: SERRAO-M
"""

#!/usr/bin/env python
# coding: utf-8

# In[ ]:

import matplotlib.pyplot as plt
import numpy as np
import os
import pandas as pd
import subprocess
```

```
from lxml import etree
import shutil

# In[23]:
# note: don't change any double backslashes (\\) to slashes (\) when you copy your local paths
instead of the ones here
# so if your local path is C:\model\project in this file you should put it as C:\\model\\project

# this is the path to the west installation files, you should change this to your the installation
directory on your machine if it's not the same
path_to_west = "C:\\Program Files (x86)\\DHI\\2016"

# this is the path to the west project, the batch file should be created in the same directory as model
files
#path_to_project = "C:\\Users\\sdaneshg\\OneDrive - UGent\\Desktop\\WEST_model_latest
update\\WEST_model\\WEST\\data\\projects\\Influent_model\\Influent"
path_to_project = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\REALITY_P_M" #
changed this to directory of Big model
# change directory
os.chdir(path_to_project)
# the name of the log file to save and the name of the ObjEval file to run the model
# these two will be used only in texec command

# this will read all .Dynamic.ObjEval.Exp.xml files in your directory and make a list out of their names
# useful if you have multiple of them for more scenarios
# the assumptption is that you have already gone through the procedure for making
.Dynamic.ObjEval.Exp.xml files
# for scenarios based on the parameters that need to be changed
exp_filenames = []
for files in os.listdir(path_to_project):
   if files.endswith(".Dynamic.ObjEval.Exp.xml"):
         exp_filenames.append(files)

# print to see if it works: it's OK
print(exp_filenames)

original_xml = os.path.join(path_to_project, exp_filenames[0])
tree = etree.parse(original_xml)
root = tree.getroot()
time_settings = root[0][1][0][0][1][3]

for q in time_settings.iter('Prop'):
   if q.get('Name') == 'StartTime':
      print('start_time: ' + q.get('Value'))
      start_time = float(q.get('Value'))*24
```

```
    elif q.get('Name') == 'StopTime':
        print('stop_time: ' + q.get('Value'))
        stop_time = float(q.get('Value'))*24

# here the batch file will be created
windows_batch_file = os.path.join(path_to_project, "run_model.bat")

# open the batch file and write the lines inside it
fbat = open(windows_batch_file, "w")
fbat.write("SET INCLUDE=%INCLUDE%;"+path_to_west+"\\WEST\\include\n")
fbat.write("SET LIB=%LIB%;"+path_to_west+"\\WEST\\lib\\win32-msvc\n")
fbat.write("SET
PATH="+path_to_west+"\\bin;"+path_to_west+"\\WEST\\third_party\\bcc5.5\\Bin;%PATH%\n")
fbat.write("SET TORNADO_CC_PLATFORM=win32-bcc5.5\n")
fbat.write("SET TORNADO_CC_PATH="+path_to_west+"\\WEST\\third_party\\bcc5.5\n")
fbat.write("SET TORNADO_DATA_PATH=C:\\Users\\Public\\Documents\\DHI\\WEST\n")
fbat.write("SET TORNADO_ROOT_PATH="+path_to_west+"\\WEST\n")
fbat.write("SET OPENMODELICAHOME="+path_to_west+"\\bin\n")

# up to here, the previous lines are for your machine to know where tornado files are stored

# here you can then add a command for tbuild or texec

# In[3]:

##############################################################################
# read this part only if you need to use tbuild command
# for creating model files to run on linux (like for running on a cluster)
##############################################################################

# example for tbuild command, xxx is the path to the model block library,
# you should keep External\External at the end of this path, just change the previous part
# for example it should be something like this: C:\example\blocks\MSL\External\External
# _ProjectName_ is only the name of the project and not the complete path, the complete
# path is not required since you are already in the directory of the project

# fbat.write("tbuild -S C:\\Users\\sdaneshg\\OneDrive - UGent\\Desktop\\WEST_model_latest
update\\WEST_model\\WEST\\data\\blocks\\WESTforKALLISTO - InfluentModel\\External\\External
-B "" _Influent_ > tbuild.log\n")

# In[24]:

##############################################################################
# this part is related to using texec for running simulations
##############################################################################
```

```
# example for texec command

# EXPLANATION if you need to run tbuild first
# when you first run this batch script, keep this line as a comment
# you need to run the tbuild first
# after you run this with tbuild command successfully then open this file again
# and comment the fbat.write("tbuild...") line and uncomment the line below
# then run the script again to create batch file fo simulations

# log_filename is the name of the log file (with .log extension) that you choose to save
# exp_filename is the .Dynamic.ObjEval.Exp.xml file you have in your project directory
# these two should be only names and not full path, you don't need the full path here
# output_name_orig is the name of the output file you choose to save in WEST environment
# this should work even if you have multiple ObjEval files for example:
scen_1.Dynamic.ObjEval.Exp.xml, scen_2.Dynamic.ObjEval.Exp.xml , ...
# this will create a texec command for each scenario to run and then rename the output file
accordingly
# so they will not be overwritten
# this is not probably the most efficient way to do this on windwos but it should work

#output_name_orig = "PreDenit_Out.txt" # this is a test file, it really exists, but once we know from
Enrico (DHI) how to read the end values and use them as initail values, we will have to change the
filename accordingly
# eventhough we don't use this output any longer for setting the initial values in the reduced model;
keep this as it will store the State variables in concentrations, which we can use for plotting purposes

with
open('C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2\\cou
nter.txt', "r") as f2:  # reading the counter
   j = int(f2.read())

output_name_orig = "PRE_7_Mass_State_Var_out.txt" # this output file stores all Mass State
Variables from the Bulk Liquid in PRE_7
for f in exp_filenames:
   log_filename = f.split('.')[0] + ".log"
   new_output_name = str(start_time) + '_' + str(stop_time) + output_name_orig
   fbat.write("texec -l \"" + log_filename  + "\"" + " -cp " +
"\"LINEBiofriltration.Dynamic.ObjEval.Exp.xml\"" + " \"" + f + "\"\n") # saves the end values of the
current run in a new file named 'End_Values.' + overwrites the xml file with the end_values as initial
values
   fbat.write("ren \"" + output_name_orig + "\" \"" + new_output_name +"\"\n")
# write exit and close the file
fbat.write("EXIT")
fbat.close()
```

```
os.chdir(path_to_project)
p = subprocess.Popen("run_model.bat")
#,shell=True, stdout=subprocess.PIPE, strerr=subprocess.PIPE
p.communicate()

print('Done!')
print('Done! The WEST model has run a Dynamic Simulation from', start_time/24 , '(days) to ' ,
stop_time/24 ,'(days) and the output files have been stored in the directory.')
print('The LINE...xml file has been reinitialized (the end_values of M(S_S) etc. have been stored as
initial values for next iteration).')
print()
with
open("C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2\\Q_
Dose_Reality.txt", "r") as f3:
   q = float(f3.read())
   print("The optimal Q_Dose in Reality Model =  " + str(q), " (m3/d)")

# # Outputfile importeren

# In[156]:

#search for the output file named 'PreDenit_Out.txt'
filename = 'PreDenit_Out.txt'

# and create a modified filename
new_filename = f"PreDenit_Out_ITER_{j}_.txt"

# Saving the file under the new name
shutil.copyfile(filename, new_filename)

print(f"File '{filename}' saved as '{new_filename}'.")

#output_name=path_to_project + "\\" + new_output_name

#output_data=pd.read_csv(output_name, sep="\t",header=[0])
#output_data=output_data.drop(0)
#output_data=output_data.set_index('#.t')
#display(output_data)

# In[157]:

#output_data.plot()
#print(output_data.dtypes)
```

```
# In[164]:

#output_data['.ASU_2.X_TSS']=output_data['.ASU_2.X_TSS'].astype(float)
#output_data['.ASU_4.C(S_O)']=output_data['.ASU_4.C(S_O)'].astype(float)
#output_data['.Qair_Basin_4']=output_data['.Qair_Basin_4'].astype(float)
#output_data['.SST_1.X_Out']=output_data['.SST_1.X_Out'].astype(float)
#output_data['.SST_1.X_Under']=output_data['.SST_1.X_Under'].astype(float)
#print(output_data.dtypes)
#output_data.iplot()

# In[168]:

#from statsmodels.tsa.seasonal import seasonal_decompose
#results=seasonal_decompose(output_data['.ASU_2.X_TSS'])
#results.plot()

# os.chdir(path_to_project)
# p = subprocess.Popen("run_model.bat")
# #,shell=True, stdout=subprocess.PIPE, strerr=subprocess.PIPE
# p.communicate()

# print()
# print('Done! The WEST model has run a Dynamic Simulation from {formatted_start_time} to
{formatted_stop_time} and the output files have been stored in the directory.')
```

Script 7: 'Calculate_new_setpoint_SSE.py

```
# This script reads the end_values of the simulation of the Process Model; stores the last value of
NO3_modelled.
# Then we calculate the difference: Setpoint - NO3_modelled = delta
# then we add delta to the setpoint: setpoint + delta and store it in the setpoint_array first position
# we do this so that in the next iteration, the optimizer in the MPC loop will compare the MPC model
effluent to the modified setpoint
# Idea is that this way the MPC knows what is happening in the real world and corrects for any
divergence

# -*- coding: utf-8 -*-
"""
Created on Tue May  9 11:12:44 2023

@author: SERRAO-M
"""
import os
import pandas as pd
```

```
import numpy as np

# Read observed variable:
filepath = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\REALITY_P_M\\"
filename = "PreDenit_Out.txt"
header_to_find = ".PRE_7.NO3_OUT"

# change directory
os.chdir(filepath)

# Open the PreDenit_Out.txt that contains the end_values of the Process Model

with open(filename, 'r') as file:
   # Read all lines from the file
   lines = file.readlines()

# Get the column names from the header
header = lines[0].split()

# Get the index of the ".PRE_7.NO3_OUT" column
pre_column_index = header.index(".PRE_7.NO3_OUT")

# Initialize an empty list to store the values in the ".PRE_7.NO3_OUT" column
pre_values = []

# Iterate through the lines (excluding the header and units row) and extract the values in the
".PRE_7.NO3_OUT" column
for line in lines[2:]:
   values = line.split()
   pre_values.append(float(values[pre_column_index]))

# Get the last value in the ".PRE_7.NO3_OUT" column
pre_value = pre_values[-1]

# Create a DataFrame with the extracted value
df = pd.DataFrame({'.PRE_7.NO3_OUT': [pre_value]})

# Calculate the difference with the set point
setpoint = 1.3
delta = setpoint - df['.PRE_7.NO3_OUT']

# Create an array of 121 values for 2 hours of datapoints per minute (120 points) ; 1500 points = 24
hrs (1440) + 1 hr (60)
setpoint_array = np.empty(1500)
```

```
# select either option 1 or 2:

# option 1:
# setpoint_array[0] = (setpoint + delta) # set first value in array to (setpoint + delta)
# if setpoint_array[0] < 0:  # to avoid negative values
#    setpoint_array[0] = 0
# setpoint_array[1:] = setpoint # value of line 1 onwards is set to setpoint value

# option 2:
for i in range(len(setpoint_array)): # entire array is changed
  setpoint_array[i] = (setpoint + delta)

# Display the array
print(setpoint_array)

# now, to write the setpoint_array into the NO3_setpoint.txt file in the directorty of the MPC Model:
filepath =
"C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2\\"
# change directory
os.chdir(filepath)

# Save the array to the "NO3_setpoint.txt" file
np.savetxt('NO3_setpoint.txt', setpoint_array, fmt='%.6f')

# Confirmation message
print("NO3_setpoint.txt file has been updated with new values.")

# now also calculate PERFORMANCE of REALITY Model compared to the SETPOINT
# create a second array of setpoints for the Performance evaluation of the Reality Model (using first
16 points)
Reality_setpoint_array = np.empty(16)

# calculate the sum of squared error between NO3_Obs and NO3_Pred
squared_error_Reality = np.sum((np.array(Reality_setpoint_array)[1:] - np.array(pre_values)[1:]) **
2)
print("The SSE of the Reality Model with the desired set point over the next 15 minutes: ",
squared_error_Reality)

# save the SSE of the Reality Model in a txt file:
filepath = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\REALITY_P_M\\"
# change directory
os.chdir(filepath)
# Open the file in append mode and write the squared_error_Reality value
with open('Reality_SSE.txt', 'a') as file:
   file.write(str(squared_error_Reality) + '\n')
```

Script 8: Counter.py

```
# -*- coding: utf-8 -*-
"""
Created on Tue Apr 25 16:36:17 2023

@author: SERRAO-M
"""
import os

path = "C:\\Users\\Public\\Documents\\DHI\\WEST\\data\\projects\\June2020_MeOH_Control2"
# change directory
os.chdir(path)

with open('counter.txt', "r") as f2:  # reading the counter
   j = int(f2.read())
   print("This was iteration number: " + str(j))
   print("The new value of the counter is: " + str(j+1))

with open('counter.txt', "w") as f2:  # write counter j + 1
   f2.write(str(j+1))
```