Développement de méthodes spatio-temporelles pour la prévision à court terme de la production photovoltaïque

Xwégnon Agoua

▶ **To cite this version:**

Xwégnon Agoua. Développement de méthodes spatio-temporelles pour la prévision à court terme de la production photovoltaïque. Energie électrique. Université Paris sciences et lettres, 2017. Français. NNT : 2017PSLEM066 . tel-01878943

HAL Id: tel-01878943
https://pastel.hal.science/tel-01878943

Submitted on 21 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à MINES ParisTech

Développement de méthodes spatio-temporelles pour la prévision à court-terme de la production photovoltaïque

École doctorale n°432

SCIENCES DES MÉTIERS DE L'INGÉNIEUR (SMI)

Spécialité ENERGÉTIQUE ET PROCÉDÉS

COMPOSITION DU JURY :

Valérie MONBET
Université de Rennes 1, Présidente

Denis ALLARD
INRA Avignon, Rapporteur

Philippe LAURET
Université de la Réunion, Rapporteur

Virginie DORDONNAT
RTE, Examinateur

Paul PONCET
ENGIE, Examinateur

Georges KARINIOTAKIS
MINES ParisTech, Examinateur

Robin GIRARD
MINES ParisTech, Examinateur

Soutenue par **Xwégnon Ghislain AGOUA**
le 20 décembre 2017

Dirigée par **Georges KARINIOTAKIS**
et **Robin GIRARD**

MINES ParisTech | PSL RESEARCH UNIVERSITY PARIS

Remerciements

Ce manuscrit est l'épilogue d'une aventure qui a commencé par un échange téléphonique, une discussion sur un banc du jardin du Luxembourg avec Robin et des échanges téléphoniques avec Georges. C'est pourquoi je voudrais en premier lieu remercier Robin Girard et Georges Kariniotakis pour la confiance qu'ils m'ont accordé et le soutien qu'ils m'ont apporté tout au long de ces trois ans et trois mois de thèse. Avec votre style respectif, vous avez contribué à la bonne marche de ces travaux par votre disponibilité, votre rigueur, votre capacité à me (re)motiver, votre propension à développer mon autonomie tout en vous assurant de m'éviter le hors piste. Les longues discussions qui commençaient par trois avis ou idées différentes et qui convergeaient tard dans la soirée ou le lendemain vers une idée consensuelle et les versions v.x ($x \geq 15$) des présentations ou articles préparés ne sont qu'une petite illustration de votre contribution au développement de ma curiosité scientifique, à l'amélioration de mon sens de la communication scientifique. Je vous en remercie. Ces années à vos côtés ont été une très belle expérience pour moi.
J'aimerais faire ici un petit saut dans le passé pour remercier Raphaël Nedellec qui m'a encadré pendant mon stage de fin d'études d'ingénieur et qui m'a ouvert à la recherche dans le domaine de l'énergie. Merci pour ton encadrement et tes conseils. Merci aussi à Yannig Goude pour les échanges sur mon stage et aussi pendant le choix de mon sujet de thèse.

Je tiens ensuite à remercier les membres de mon jury qui ont accepté de consacrer leur temps que je sais précieux à examiner, commenter, et analyser mon travail. Merci à mes rapporteurs Valérie Monbet qui était aussi présidente du jury, Denis Allard et Philippe Lauret pour les commentaires, remarques, questions, conseils qui ont significativement amélioré la qualité de ce manuscrit. Merci à mes examinateurs Paul Poncet et Virginie Dordonnat pour leurs commentaires et analyses mais aussi d'avoir partagé la richesse de leurs expériences d'industriels en vue d'améliorer ce travail.

Je remercie aussi l'ensemble des partenaires industriels qui ont contribué à la bonne marche de ces travaux de thèse à savoir Solais, Coruscant, Compagnie Nationale du Rhône et Hespul en nous fournissant les données de production. La richesse des données que vous avez bien voulu partager avec nous ont permis de pouvoir travailler avec sérénité. Les échanges que nous avons eu ont aussi permis de mieux appréhender les attentes opérationnelles en terme de prévision. Je remercie également ECMWF pour avoir fourni les prévisions météorologiques que nous avons utilisé dans le cadre de ces travaux. Je tiens aussi à remercier Prof. Philippe Blanc qui nous a fourni les données d'irradiation du modèle ESRA et les images satellites (Merci indirect aussi à TransValor) utilisées dans le cadre de cette thèse. Je le remercie aussi pour sa porte toujours ouverte qui a facilité les discussions formelles et informelles et les conseils qu'il m'a donné pendant cette thèse.

Faire une bonne thèse nécessite d'avoir aussi un bon accompagnement pour les différentes démarches administratives. Merci donc à Lyliane, Marie-Jeanne, Christine et Brigitte pour leur soutien pendant ces trois années. Grâce à vous je peux me targuer de dire que les démarches administratives c'est facile quand on est bien accompagné. Merci à Muriel de m'avoir aidé à dompter les outils bureautiques lors de l'impression de mon manuscrit et de l'envoi au jury (l'exemplaire pour La Réunion est bien arrivé à destination !).

La thèse c'est aussi de nouvelles et belles rencontres, des conversations sérieuses ou non, des amitiés qui naissent, des claviers souffre-douleur, des litres de thé, de super goûters, etc. Merci aux anciens : Arthur, Benjamin, William, Valentin, Sabri pour leurs conseils et les échanges. Le bureau R009b a été le siège des discussions les plus philosophiques au plus potaches ; Merci à Fiona, Etta, Thibaut, Lucia, Adrian et Antoine pour la bonne humeur, le soutien moral, les échanges, les repas, les soirées, les Friday songs. Merci à Guillaume et Maxime pour les soirées. A Fabien et Papa, je refais une apparition remarquée au foot quand vous voulez. Merci aux autres doctorants du labo : Romain, Kévin, Pedro, Di, SeungWhoo, Giovanni, Simon, Thomas, Amaury, pour les divers échanges scientifiques ou non. Merci au club Cemef 230 : Lucile et Gerry pour des moments de bus hors du commun. Yoann, un merci spécial pour ces breuvages rafraichissant dont tu as le secret qui ont permis de tenir la chaleur du sud.

Je remercie particulièrement Kévin Giron, la plus grande girouette que je connaisse. Merci pour ces deux super années de colocation teintées de bonne humeur et de soirées épiques. Merci à Abdou, Arthur, Charles, Julie pour votre amitié et votre soutien. Vous êtes les "best" ! Merci à Alice pour son soutien et sa disponibilité. Merci à Chirif, Collince et Kersane pour les délires en tout genre. Ceux que j'ai oublié de citer, ne m'en voulez pas, merci à vous aussi.

J'aimerais avoir une pensée pour ma défunte grand-mère, sache que prononcer « Xwégnon » a causé et cause des problèmes même hors du Bénin. Enfin je remercie mon père, ma mère et mes frères pour tout. La distance n'a jamais rien enlevé à votre soutien.

Les publications

Les articles publiés et en cours de préparation

- Probabilistic Models for Spatio-Temporal Photovoltaic Power Forecasting. Xwégnon Ghislain Agoua, Robin Girard, Georges Kariniotakis. A paraître dans IEEE Transactions on Sustainable Energy, 2018.
  https://doi.org/10.1109/TSTE.2018.2847558
  Preprint disponible dans l'annexe A.
- Short-Term Spatio-Temporal Forecasting of Photovoltaic Power Production. Xwégnon Ghislain Agoua, Robin Girard, Georges Kariniotakis. IEEE Transactions on Sustainable Energy, April 2018, Volume 9, Issue 2, Page(s) : 538-546.
  https://doi.org/10.1109/TSTE.2017.2747765
  Preprint disponible dans l'annexe A.
- Photovoltaic power forecasting : comparison of multiple spatio-temporal sources of data on the forecasts accuracy. Working paper.

Les conférences et séminaires

- A Stochastic Multi-Temporal Optimal Power Flow Approach for the Management of Grid Connected Storage. Etta Grover-Silva, Xwégnon Ghislain Agoua, Robin Girard, Georges Kariniotakis. CIRED 2017 (Congres International des Réseaux Electriques de Distribution), Jun 2017, Glasgow, United Kingdom. pp.12 - 15
  https://hal-mines-paristech.archives-ouvertes.fr/hal-01500379
- Spatio-temporal forecasting of wind and PV production. Workshop on Energy Forecasting and Applications, ENGIE-PERSEE, Sept 2017, Sophia-Antipolis
- Spatio-temporal forecasting of PV production. DTU-ERSEI Seminar, Oct 2016, Sophia-Antipolis
- Spatio-temporal models for photovoltaic power short-term forecasting. Solar Integration workshop 2015, Oct 2015, Brussels, Belgium.
  https://hal-mines-paristech.archives-ouvertes.fr/hal-01220321
- Spatio-temporal wind power forecasting with distributed wind farms as sensors. Spatial Statistics Conference, Avignon, June 2015

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction générale

1.1 Le secteur de l'électricité dans le monde

La demande totale en énergie primaire dans le monde s'élevait en 2016 à 13 300 Mtep [1] avec près de 80% de cette demande couverte par les ressources fossiles à savoir le pétrole, le charbon et le gaz naturel [3]. L'Agence Internationale de l'Énergie (AIE) dans son principal scénario d'évolution de la demande en énergie prévoit une augmentation de la demande de l'ordre de 30% à l'horizon 2040. Cette hausse de la demande en énergie concerne toutes les sources d'énergies.

Parallèlement à cette augmentation de la demande en énergie, la grande majorité des états de la planète ont pris conscience de l'effet de la production énergétique sur le réchauffement climatique. L'accord de Paris sur le climat [4], entré en vigueur en novembre 2016, a placé le secteur de l'énergie au cœur des discussions. En effet, ce secteur, responsable d'au moins deux tiers des émissions de gaz à effet de serre doit être transformé pour atteindre les objectifs de l'accord de Paris notamment sur les valeurs seuils d'élévation de la température de la planète. La nécessité de préserver l'environnement, couplée à l'impératif de réponse à une demande en énergie grandissante conduit à ce qu'on appelle la transition énergétique ou la refonte du secteur de l'énergie. Cette transition se matérialise par la poussée des investissements dans les sources d'énergie faiblement carbonées, l'accroissement de l'efficacité énergétique et l'amélioration de l'intensité énergétique à l'échelle de l'économie mondiale. Cette transition énergétique se traduit aussi par la réduction des investissements dans les secteurs pétroliers et gaziers essentiellement portée par la refonte des politiques de subvention. La transition énergétique touche aussi le secteur du charbon, dont l'utilisation tend à plafonner sous le poids des nombreuses préoccupations environnementales. Quant au secteur du nucléaire, la réduction de son importance est envisagée par la plupart des pays utilisateurs à part la Chine qui porte quasiment à elle seule l'augmentation des investissements de la filière.

Le développement des énergies renouvelables est la principale source de transformation du secteur de l'électricité. En 2015, la source d'énergie renouvelable la plus utilisée est le charbon de bois et les biocarburants solides car ils sont très prisés dans les pays en développement pour la cuisson et le chauffage. L'énergie hydraulique est la deuxième source d'énergie renouvelable suivie par l'ensemble constitué de la géothermie, les biogaz, les biocarburants liquides, le solaire, l'éolien et la marée [5]. Dans le cas spécifique des pays de l'Organisation de Coopération et de Développement Économiques (OCDE), la principale source d'énergie renouvelable est l'hydroélectricité suivie par l'éolien, les biocarburants

1. Million de tonnes équivalent pétrole

et le solaire. Ces différentes sources d'énergie représentaient en 2016 environ 24% de la production totale d'électricité de l'OCDE, la plus grande part de renouvelables jamais observée depuis 1990 [5]. Des efforts supplémentaires sont attendus sur l'augmentation de la part des énergies renouvelables dans la production d'électricité des pays de l'OCDE. La figure 1.1 montre la puissance photovoltaïque connectée dans l'Union Européenne en 2016. L'Allemagne, l'Italie et le Royaume-Uni présentent les puissances installées les plus importantes. La France arrive en quatrième position.

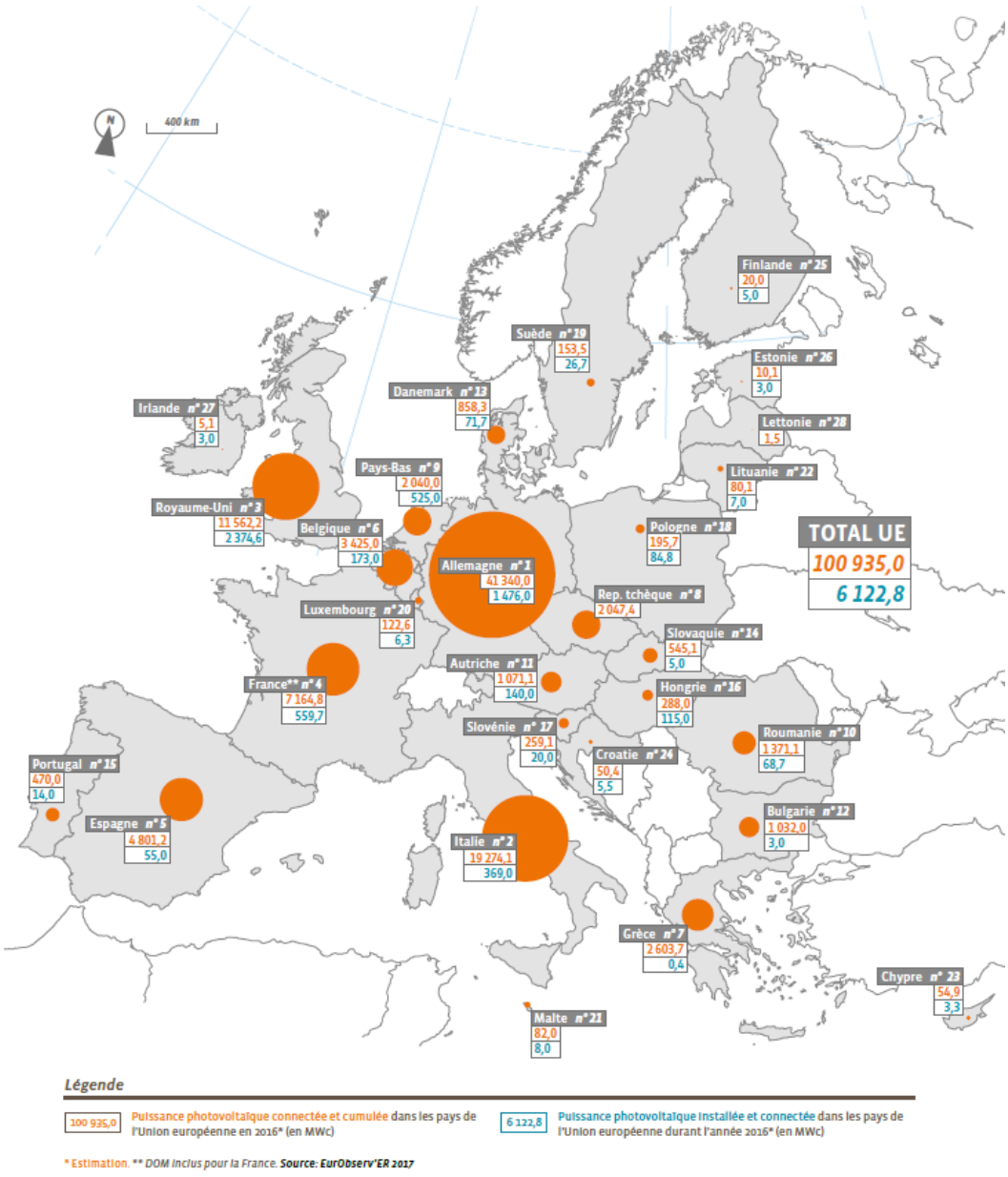


FIGURE 1.1 – Puissance photovoltaïque connectée et cumulée dans l'Union européenne en 2016 (en MWc). Source : EurObserv'ER 2017

1.2 Les énergies renouvelables en France et l'enjeu de la prévision

En France, les objectifs spécifiques à atteindre pour réussir la transition énergétique sont consignés dans la loi relative à la transition énergétique pour la croissance verte (loi n 2015-992 du 17 août 2015) [6]. Le premier article du titre premier stipule que la politique énergétique de la France doit favoriser l'émergence d'une économie compétitive, assurer la sécurité d'approvisionnement et réduire la dépendance aux importations. Elle doit aussi préserver la santé humaine et l'environnement, en particulier en luttant contre l'aggravation de l'effet de serre. La politique énergétique doit aussi contribuer à la mise en place d'une Union Européenne de l'énergie, qui vise à garantir la sécurité d'approvisionnement et à construire une économie décarbonnée et compétitive, au moyen du développement des énergies renouvelables, des interconnexions physiques, du soutien à l'amélioration de l'efficacité énergétique et de la mise en place d'instruments de coordination des politiques nationales. Les énergies renouvelables ont donc une place de choix dans la transition énergétique en France. Les axes prioritaires et les différentes stratégies à mettre en place pour atteindre ces objectifs sont définis dans la programmation pluriannuelle de l'énergie (PPE) portée par le Ministère de l'environnement, de l'énergie et de la mer, en charge des relations internationales sur le climat (dénomination de 2016). Dans la PPE l'objectif fixé pour 2023 pour le développement des énergies renouvelables est d'augmenter de plus de 70% la capacité installée des énergies renouvelables par rapport au niveau de 2014 [7].

Le parc de production d'électricité renouvelable en France métropolitaine affichait au 31 mars 2017 une puissance installée de 46 392 MW répartis sur les différents réseaux des acteurs du secteur. Dans ce parc, la filière la plus importante est l'hydraulique suivie des filières éolienne et solaire. Au 31 mars 2017, le parc solaire installé en France métropolitaine atteint une capacité installée de 6 854 MW [2]. La figure 1.2 présente les niveaux de raccordement par région. Les régions du sud de la France à savoir la Nouvelle-Aquitaine, l'Occitanie présentent les puissances installées les plus importantes. Elles sont suivies des régions Provence-Alpes-Côte d'Azur et Auvergne-Rhône-Alpes. Ce sont aussi ces régions qui présentent les meilleurs niveaux d'ensoleillement en France. Ces chiffres sont très encourageants et montrent une bonne croissance des puissances solaires raccordées par rapport au niveau de 2015 (6 196 MW). Cette croissance peut s'expliquer par les efforts d'innovation pour réduire les coûts d'implantation de moyen de production et une meilleure structuration de la filière industrielle.

Pour atteindre les objectifs de la PPE, la seule réduction du coût des énergies renouvelables n'est pas suffisante. En effet, depuis 2007, la quasi-totalité de la puissance photovoltaïque installée en France est raccordée au réseau donnant ainsi naissance à des enjeux importants pour les gestionnaires de réseaux chargés de raccorder les centrales et de garantir la sécurité du système électrique. La conception du réseau électrique en France est basée sur une production « commandable », disponible et raccordée au réseau de transport ce qui n'est pas le cas des sources de production d'énergies renouvelables météo-dépendantes comme l'éolien et le solaire. Les réseaux électriques des autres pays de la zone Union Européenne sont aussi conçus suivant les mêmes principes. Les réseaux sont soumis à la variabilité et l'intermittence de la production renouvelable qui compliquent l'équation de l'équilibre entre l'offre et la demande qui doit être maintenu en permanence. Avec l'accroissement de la puissance photovoltaïque installée, il est donc impératif d'avoir

2. Les données sont extraites du "Panorama de l'électricité renouvelable au 31 Mars 2017" qui est un état des lieux détaillé des principales filières de production d'électricité de source renouvelable publié par RTE, le SER, Enedis et l'ADEeF.

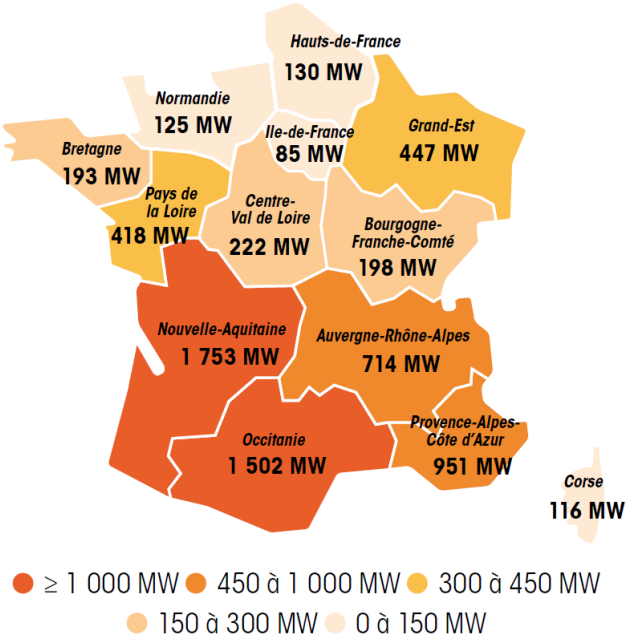


FIGURE 1.2 – Puissance solaire raccordée par région au 31 mars 2017 en France. Les régions Nouvelle-Aquitaine et Occitanie constituent les plus importants pôles de production d'électricité photovoltaïque. Source : RTE

des prévisions fiables de la production à court-terme, c.à.d. pour les horizons de quelques minutes à quelques jours.

La prévision de la production a aussi un fort intérêt économique car elle permet de réduire les coûts liés à la réserve [8], d'optimiser les décisions sur le marché de l'énergie, de réduire les coûts de régulation, d'accroitre la compétitivité des énergies renouvelables et de gérer les congestions [9]. De récentes études sur l'intégration des énergies renouvelables ont aussi montré le rôle prépondérant de la prévision court-terme dans la mise en œuvre des smart grids [8].

1.3 Etat de l'art de la prévision PV

Dans la littérature, il existe de nombreuses méthodes de prévision de la production PV [10, 11, 2, 12]. Ces méthodes peuvent être regroupées en trois grandes familles [11] : les méthodes statistiques, les méthodes physiques et les méthodes hybrides. Ces méthodes fournissent soit des prévisions de l'irradiation soit des prévisions directes de production. Le choix d'une méthode de prévision peut être guidé par plusieurs paramètres à savoir le besoin auquel doit répondre la prévision (planification, management de la réserve, participation au marché, etc.), l'horizon de prévision envisagé et aussi le type de données disponibles. Ces trois paramètres sont étroitement liés. En effet, le besoin à l'origine de la prévision permet de déterminer quels sont les horizons intéressants et donc quelles sont les données à utiliser. Il existe diverses sources de données utilisables dans le cadre de la prévision de production PV à savoir les mesures production et de variables météorologiques comme l'irradiation solaire, les prévisions météorologiques, les images de caméras ou de satellites. Une approche intéressante est de regrouper les modèles de prévisions par horizons croissants de quelques minutes à plusieurs jours.

Dans le cadre de la prévision PV on retrouve la classification en fonction des horizons de prévision suivante [13] :

— des prévisions intra-horaires (de 15 min à 2h avec un pas temps de 1min) ;

— des prévisions à très court-terme pour des horizons de quelques heures ($\leq$ 1h - 6h) ;

— des prévisions à court terme pour des horizons de quelques jours (1 jour - 3 jours) ;

— des prévisions à moyen-terme (1 semaine - 3/4 mois) ;

— des prévisions à long-terme ($\geq$ 1 an).

Les prévisions intra-horaires et très court-terme qui couvrent des horizons allant de moins de quelques minutes à quelques heures sont essentielles aux activités de traitement de la variabilité, de suivi de la production, d'ajustement de la charge et de gestion du stockage. La prévision à moyen terme est utilisée dans le cadre du management et du trading d'énergie. La prévision à long terme quant à elle permet une meilleure planification et optimisation des ressources. On retrouve dans la littérature des comparaisons de méthodes de prévision pour des horizons court et très court-terme [14, 15, 16] et des analyses détaillées de ces méthodes suivant le type de données d'entrées [2].

Dans la suite, nous proposons une revue des méthodes de prévision de l'irradiation solaire et de la production PV classifiées par thèmes. Les thèmes choisis ont pour but de regrouper les méthodes de prévision qui sont pertinentes dans le cadre de cette thèse.

Les modèles météorologiques

Les principaux modèles utilisés dans la littérature pour la prévision de l'irradiation solaire sont des modèles météorologiques. Ils sont dénommés modèles NWP (Numerical Weather Prediction) [17] à cause de leur caractère numérique et sont des modèles physiques complexes (résolution de système non linéaire d'équations différentielles) qui permettent de prévoir différentes variables météorologiques. Le principe général de ces méthodes repose sur une bonne connaissance au temps initial de l'état de l'atmosphère et des principes/lois physiques qui régissent les changements d'état de l'atmosphère [2]. Le fonctionnement des modèles NWP peut être décrit par la figure 1.3. On commence par choisir une zone (un domaine), qu'on discrétise ensuite spatialement suivant une résolution choisie. Enfin les modèles NWP prévoient les informations désirées en résolvant des équations de thermodynamique. Les modèles NWP ont beaucoup évolué et leur amélioration est très dépendante des capacités de calcul.

FIGURE 1.3 – Principe de fonctionnement de base des modèles NWP. La variable prédite ici est la température à la surface et l'horizon de prévision est 18 h. Source : Inman [2]

Les modèles NWP se divisent en modèles globaux ou régionaux. Parmi les modèles globaux, on peut citer le modèle ECMWF (European Centre for Medium-Range Weather

Forecasts) et le modèle GFS (Global Forecast System). Le modèle GFS a été développé aux Etats-Unis (National Centers for Environmental Prediction) et propose un accès gratuit aux données. Le modèle ECMWF développé par l'organisation intergouvernementale Européenne propose un accès payant aux données opérationnelles à travers les organismes météorologiques nationaux. Parmi les modèles régionaux on peut citer les modèles NAM (North American Mesoscale), RAP ou RAR (Rapid Refresh) qui offrent un accès gratuit aux données. Le tableau 1.1 présente les principales caractéristiques de ces modèles et la figure 1.4 montre le système de coordonnées utilisées pour le découpage.

Tableau 1.1 – Caractéristiques de quelques modèles NWP. Les modèles ECMWF et GFS sont des modèles globaux. Source : Kleissl [1]

Modèles	ECWF	GFS	NAM	RAP
Résolution spatiale horizontale (km)	16 (T1279)	50 (0.5)	12 (0.1)	13
Nombre de niveaux verticaux	91	47	42	50
Résolution temporelle en sortie (heures)	3	3	1	1
Horizon de prévision	6 jours	8 jours	36 heures	18 heures

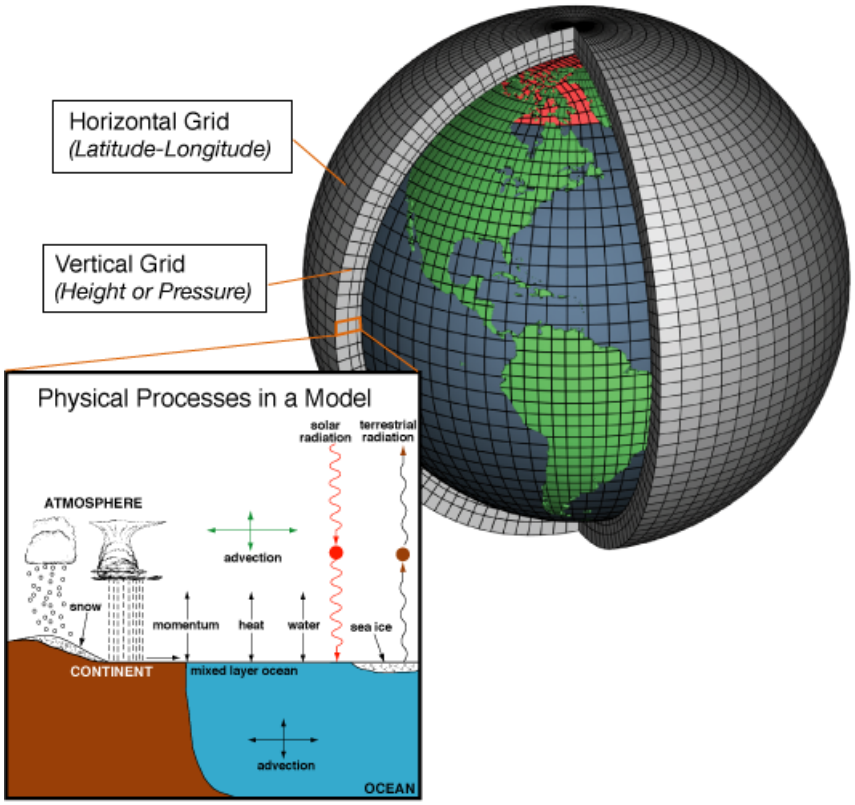


FIGURE 1.4 – Description des systèmes de coordonnées et des procédés physiques dans un modèle NWP. Source : Wikipedia

Les sorties des modèles NWP sont très rarement utilisées brutes. Elles peuvent être couplées à des observations d'images satellites ou de caméras hémisphériques qui permettent de décrire la couverture nuageuse au niveau des centrales par l'étude des mouvements des nuages [18, 19, 20]. Ces sorties peuvent aussi être injectées dans des modèles statistiques afin de prévoir l'irradiation ou la production PV. Ces modèles présentent plu-

sieurs limitations. La première concerne les temps de calcul (et de mise à jour) importants. En effet malgré les avancées dans le domaine pour construire des modèles plus rapides en réduisant le nombre de paramètres ré-estimés par exemple, les temps de calcul et de mise à jour restent des freins importants pour l'exploitation des produits de ces modèles dans le cadre de la prévision à court terme. Une autre limitation concerne la résolution spatiale de ces modèles qui rend impossible la descente à des niveaux microscopiques de la physique qui sont associés à la formation des nuages. Les séries d'irradiation fournies en sortie des modèles NWP sont souvent traitées par un second modèle "filtre" qui permet de les épurer de la variabilité due au cycle solaire. Cette variabilité étant connue avec une bonne précision ces modèles filtres sont intéressants en ce sens où ils permettent de se concentrer sur les autres sources de variabilité. Ces modèles "filtres" sont dits modèles de ciel-clair.

Les modèles ciel-clair

Le rayonnement solaire au sol est variable dans le temps (et dans l'espace). Cette variabilité est due à la position du soleil, à l'état de l'atmosphère (vapeur d'eau, ozone, aérosol) et enfin à la couverture nuageuse. Cette variabilité de la ressource solaire entraine celle de la production d'énergie grâce aux systèmes PV. Le rayonnement solaire au sol peut être modélisé par le produit du rayonnement ciel-clair par les effets d'atténuation dus aux nuages y compris l'albédo. Le rayonnement ciel-clair est observé lorsqu'il n'y aucun nuage. Il dépend donc essentiellement de la distance Terre/Soleil, de l'altitude et aussi de paramètres décrivant les gaz et particules de l'atmosphère même sans nuages. Il existe dans la littérature plusieurs modèles pour décrire le rayonnement ciel-clair [2, 21]. Le modèle auquel nous nous référerons ici est le modèle ESRA (European Solar Radiation Atlas) [22]. La variation de composition de l'atmosphère est prise en compte dans ce modèle par le trouble de Linke (Linke Turbidity Factor) calculé à partir d'une base de données en moyenne mensuelle. La figure 1.5 présente un exemple de série d'irradiation au sol (GHI, global horizontal irradiance) donnée par le modèle ciel-clair ESRA pour l'année 2014. On peut y observer les cycles journaliers et annuels.

Dans le cadre de la prévision de l'irradiation (ou rayonnement), on définit un indice de ciel-clair (initialement défini comme indice de clarté ou de couverture nuageuse [23, 24, 25]). Cet indice k_c est le ratio entre l'irradiation globale mesurée GHI, et l'irradiation estimée par un modèle ciel-clair GHI_{clear} comme définie par $k_c = GHI/GHI_{clear}$. L'irradiation ciel-clair GHI_{clear} peut être fournie par l'un des modèles de la littérature énumérés au paragraphe précédent. L'indice de ciel-clair est une variable bornée (théoriquement) entre 0 et 1. La division par l'irradiation estimée sous conditions de ciel-clair permet ainsi d'extraire les cycles connus de l'irradiation pour se concentrer sur les variations liées à la couverture nuageuse.

Il existe aussi dans la littérature des indices ciel-clair purement statistiques utilisés pour normaliser non pas les séries d'irradiation mais celles de production PV. Ces modèles ne nécessitent aucune connaissance physique préalable des interactions dans l'atmosphère mais seulement les informations tirées des observations [26, 12]. Le modèle ciel-clair est défini comme une fonction du jour de l'année et du moment de la journée. Dans ces modèles ciel-clair statistiques, la régression par noyau et la régression quantile sont utilisés pour estimer les fonctions de lien adéquates entre la production et les variables temporelles. Ce modèle ciel-clair issu de l'apprentissage statistique permet de définir un indice ciel-clair pour la production PV [27]. Cet indice sert non seulement à normaliser les séries de production mais aussi de réduire leur non stationnarité. Les séries normalisées sont plus

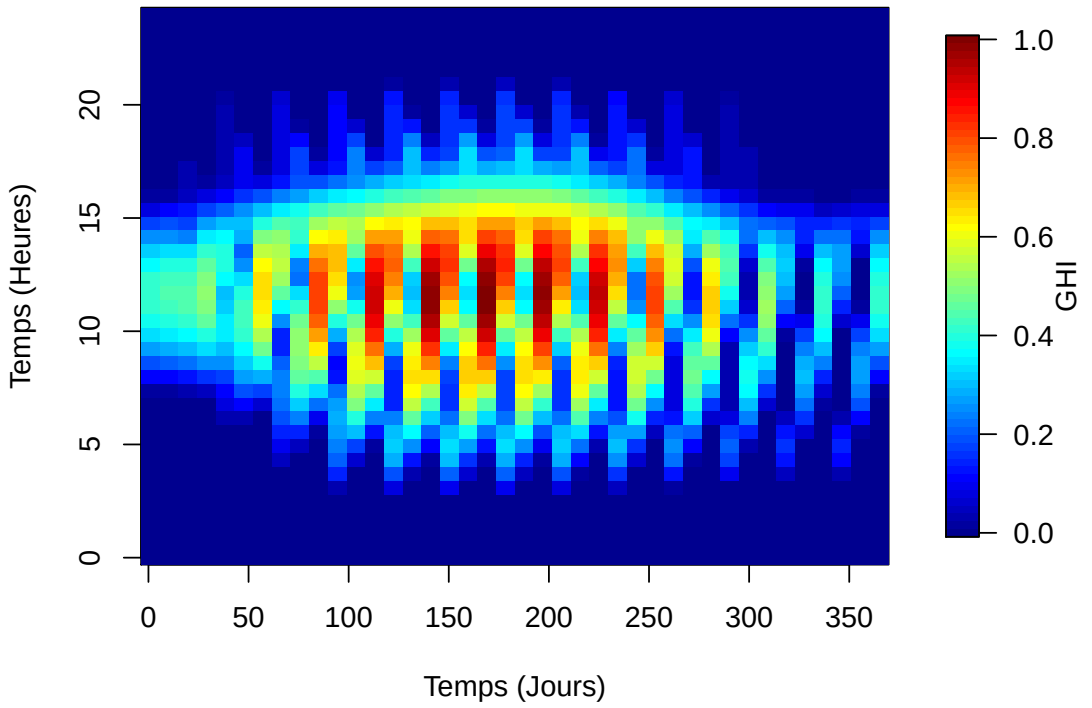


FIGURE 1.5 – Irradiation globale sur plan horizontal (GHI) fournie par le modèle ciel clair ESRA pour l'année 2014. Site d'évaluation : une centrale PV dans le sud-est de la France

proches des propriétés de stationnarité que les séries originelles ce qui facilite l'application de méthodes classiques de traitement de séries temporelles [28]. Cette modélisation ciel-clair statistique pose plusieurs problèmes. Le premier est l'absence dans la littérature de définition claire du processus d'estimation des fonctions de lien entre la production ciel-clair et le moment de la journée. De plus, dans la littérature, les références proposant une méthode de ciel-clair statistique pour normaliser les séries de production ne présentent pas d'analyse post-stationnarisation des propriétés des nouvelles séries que ce soit au sens des indicateurs de stationnarité que de la performance de ces séries pour la prévision.

La prévision déterministe de la production PV

Il existe plusieurs méthodes dans la littérature qui permettent de fournir des prévisions de la production PV. Dans cette partie nous décrivons les méthodes dites déterministes en ce sens qu'elles fournissent pour chaque instant de l'horizon de prévision une seule valeur qui est la production moyenne attendue. La production d'un système PV peut être prévue par couplage des prévisions météorologiques de type NWP avec les caractéristiques des centrales (position, orientation, puissance crête, etc.) [29]. Les prévisions de production peuvent aussi être obtenues grâce à des modèles de traitement de séries temporelles qui peuvent combiner les données de mesures avec des variables exogènes. Dans la littérature, on retrouve un large éventail de méthodes utilisables pour prévoir la production PV à un instant donné. La figure 1.6 présente une schématisation de la prévision déterministe de la production PV. Les modèles autorégressifs et de type Box et

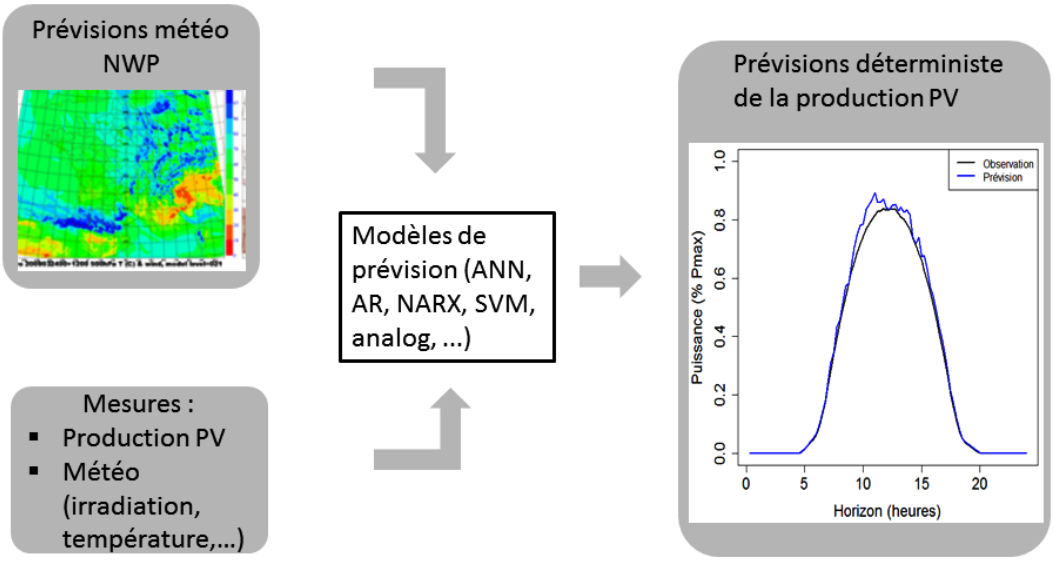


FIGURE 1.6 – Illustration des méthodes de prévision déterministe à court-terme de la production PV.

Jenkins [30, 12, 26] par exemple sont très couramment utilisés pour prévoir la production PV. Ces modèles exploitent les historiques de mesure sur site et de prévisions météorologiques en recherchant les similarités avec les observations passées. Dans la même optique, certains modèles fournissent des prévisions de production en construisant des classes de "situations" de production ou météorologiques semblables à celle observées au moment auquel la prévision est faite. Ces modèles sont appelés modèles "analogues" [15, 31, 32]. C'est ce même principe de recherche de situation semblable qui est utilisé dans [33] pour prévoir la production PV lorsque les panneaux sont couverts de neige. La production PV peut aussi être prévue par combinaison de différents modèles de prévisions [34, 35] ou par des approches non paramétriques [30].

Les méthodes d'intelligence artificielle sont aussi très utilisées dans la littérature pour prévoir la production PV. Les réseaux de neurones [36] sont les plus fréquemment utilisés. Il existe différents types de réseaux de neurones comme les Feed Forward Neural Network, les Radial Basis function Neural Network et les Recurrent Neural Network (RNN) par exemple. Ces méthodes peuvent être implémentées et comparées dans le but de choisir la meilleure fonction d'activation et les caractéristiques optimales du réseau pour la prévision de la production PV [37]. Les réseaux de neurones peuvent être combinés à diverses autres techniques afin d'obtenir des méthodes de prévision de la production PV plus performantes. On peut par exemple les associer à un modèle linéaire autorégressif avec variable exogène (NARX) [38] ou à des modèles physiques de rayonnement solaire [39, 40]. Les réseaux de neurones sont aussi utilisés dans des processus de modélisation à deux étapes où la première étape consiste à prévoir les variables météorologiques en utilisant les modèles NWP [31, 41, 42]. Les données de production PV peuvent aussi être classifiées selon les niveaux de couverture nuageuse (ciel dégagé, partiellement couvert, totalement couvert) et des réseaux de neurones spécifiques à chacune de ces classes peuvent être construits[43]. Des scénarios de la production PV peuvent aussi être construits et intégrés aux réseaux de neurones [44] pour la prévision.

Les Support Vector Machine (SVM) [45] sont aussi utilisés pour la prévision de la production PV. Il existe plusieurs formulations de l'algorithme des SVM mais ils peuvent

être interprétés comme une méthode d'estimation où le critère à estimer est la norme carrée de la fonction d'estimation dans un espace bien défini (Hilbertien reproductible). Les SVM peuvent être utilisés avec des prévisions météorologiques de type NWP pour prévoir la production PV [46, 47, 48]. L'utilisation des SVM pour prévoir la production PV peut aussi intervenir dans un processus à deux étapes où la première est une classification des situations météorologiques passées [49, 50]. Les SVM peuvent aussi être couplées à un modèle SARIMA [51] ou à des chaînes de Markov [52] pour améliorer les performances de prévision de la production PV. La méthode de gradient boosting [53] ou les algorithmes génétiques [54] sont aussi des méthodes utilisées pour la prévision de la production PV. La principale limitation de ces méthodes déterministes est qu'elles ne permettent pas de quantifier les incertitudes associées aux prévisions fournies.

La prévision probabiliste de la production PV

Le développement des modèles de prévision probabiliste a été porté par la volonté d'avoir plus d'informations sur la distribution future de la production PV et l'incertitude liée aux prévisions. Les méthodes de prévision probabiliste de la production PV peuvent se présenter sous deux approches : la prévision des niveaux de confiance associés à des prévisions déterministes et l'approche directe qui fournit une représentation de la densité de la production.

La majorité des méthodes de prévision probabiliste de la production PV repose sur la régression quantile [55]. Les quantiles sont ainsi considérés comme une bonne représentation de la densité future à prévoir. La régression quantile peut être utilisée sous différentes formes allant de simples modèles linaires à des modèles plus complexes associant des prévisions NWP [16]. La régression quantile peut être aussi couplée à des méthodes d'estimation non paramétrique [56, 9], à des méthodes de machine learning [57] ou à des méthodes de classification [58, 59] pour prévoir la production PV. Ces méthodes de classification peuvent aussi être utilisées indépendamment de la régression quantile pour obtenir des prévisions probabilistes de la production PV. On peut citer les méthodes analogues [60], les arbres de régression [16, 61] et les plus proches voisins (k-nearest neighbors, kNN) [62, 53, 58].

Il existe aussi des méthodes de prévision qui permettent d'obtenir non pas des quantiles mais directement la distribution future complète. Les densités de probabilité peuvent ainsi être obtenues en utilisant comme entrée des prévisions d'ensemble issus d'un modèle NWP [63, 64, 34, 65]. Les méthodes d'estimation non paramétrique peuvent aussi être utilisées pour prévoir la densité future de la production [66, 67]. L'approche paramétrique ou technique d'adéquation à une loi qui consiste à faire une hypothèse sur la distribution future et à en estimer les paramètres [68] peut aussi être utilisée.

Il existe aussi des méthodes de prévision probabilistes basées sur du bootstrap et couplées aux chaînes de Markov [69], aux équations différentielles stochastiques [70, 71]. L'estimation fonctionnelle des densités de probabilité par les ondelettes peut aussi être utilisée pour prévoir la production PV [72, 48]. La simulation de Monte Carlo [69, 66] et l'inférence bayésienne [73, 74] peuvent aussi être utilisées pour prévoir la production PV. Enfin on peut aussi combiner plusieurs méthodes de prévision probabiliste et les faire voter en associant à chaque estimateur des pondérations [61]. Les méthodes de prévisions précédemment définies sont portées par les prévisions météorologiques qui rappelons-le sont caractérisées par des temps de mise à jour et de calcul longs. Il existe donc un besoin d'amélioration des prévisions de production à court terme qui suppose l'utilisation de nouvelles méthodes plus rapides à implémenter et à mettre à jour.

La prévision spatio-temporelle

Les limitations des méthodes de prévision basées sur les données de modèles numériques de prévision ont conduit au développement de méthodes qui exploitent les informations spatiales dans la prévision temporelle de la production PV. Les premières méthodes proposées pour cette modélisation spatiale et temporelle ont été introduites dans le cadre de la prévision de production éolienne notamment pour la prévision des vitesse et direction de vent [75], l'étude de la propagation spatiale et temporelle des erreurs de prévision [76, 77], l'identification de région à fort potentiel de production [78] et la prise en compte de la variabilité spatiale [79, 80, 81]. Ces méthodes ont ouvert la voie à la mise en œuvre de méthodes de prévision exploitant à la fois les informations spatiales et temporelles pour la prévision de la production PV. Dans [82], on retrouve une revue des méthodes spatio-temporelles de prévision de production d'énergie renouvelable mais aussi de la demande. Le principe de base de ces modèles est qu'ils exploitent les corrélations entre les conditions atmosphériques observées sur plusieurs sites au même moment mais aussi avec des décalages temporels. L'utilisation de ces corrélations permet d'améliorer la qualité des prévisions.

La plupart des références sur la prévision spatio-temporelle de production PV traitent de la prévision de l'irradiation solaire. La spatialité des modèles provient souvent de l'utilisation d'images de caméras ou de satellites. La résolution spatio-temporelle de l'irradiation globale au sol peut être améliorée par combinaison des mesures d'irradiation et des observations fournies par les images satellites [83]. Les prévisions d'irradiation peuvent être aussi améliorées par le calcul de la vitesse de déplacement des nuages au-dessus des centrales PV [84, 33]. Ces informations sur le déplacement des nuages sont consignées dans un vecteur appelé cloud motion vector (CMV). L'impact de la vitesse de déplacement des nuages sur la variabilité des mesures d'irradiation peut être calculé en utilisant par exemple des transformations en ondelettes [85] ou des algorithmes de traitement d'images [86]. Les cloud motion vector et les prévisions météorologiques sont combinées pour fournir de meilleures prévisions à moyen-terme de l'irradiation [87, 19]. La plupart des modèles autorégressifs dans le temps ou l'espace (AR, ARMA, ARX, ARIMA, krigeage, etc) peuvent aussi être utilisés pour la prévision spatio-temporelle de l'irradiation [88, 89] de même que les modèles géostatistiques [86, 90, 83]. Les méthodes de classification [91] et la prévision semi-paramétrique [92] sont aussi utilisées dans le cadre de la prévision spatio-temporelle de l'irradiation. Les prévisions par des modèles spatio-temporels de l'irradiation peuvent certes être transformées en prévision de production PV par des modèles physiques qui prennent en compte les caractéristiques des systèmes PV mais cela correspond à l'introduction d'un niveau d'incertitude (ou d'erreur) supplémentaire.

Il existe peu de références sur les méthodes de prévision spatio-temporelles directes de la production PV. La plus ancienne référence est la méthode de régression (analogue) qui évalue la faisabilité de la collaboration entre sites de production dans le but de gérer un flux continu de données de grande dimension [32]. On retrouve ensuite des méthodes autorégressives avec des variables exogènes qui sont des mesures météorologiques au voisinage de la centrale d'intérêt [93, 94, 95, 96]. L'évaluation des trajectoires spatio-temporelles des erreurs de prévision [97] est aussi une technique utilisée pour exploiter les informations spatio-temporelles dans le but de l'amélioration de la prédictibilité. Contrairement à la prévision day-ahead, on retrouve très peu de modèles pour la prévision spatio-temporelle probabiliste à court-terme de la production PV. Dans [98, 99] on retrouve des modèles basés sur le gradient boosting et les modèles vectoriels autorégressifs. Une autre approche consiste à utiliser des méthodes non paramétriques de machine learning [57].

1.4 Objectifs et démarche de la thèse

L'objectif principal de la thèse est d'améliorer la prédictibilité à très court-terme de la production PV en utilisant les corrélations spatio-temporelles entre les données de production. Les horizons temporels qui nous intéressent varient de quelques minutes à quelques heures. Contrairement aux méthodes de prévision qui exploitent les données de production d'un site ou d'une agrégation de sites mais rarement l'ensemble des informations des sites voisins, cette thèse propose des méthodes de prévision qui exploitent les relations spatiales et temporelles entre les productions de différents sites géographiquement distribués. Les ensembles de centrales voisines sont utilisés comme des réseaux de capteurs qui apportent de l'information pour améliorer la qualité des prévisions pour chaque centrale.

Les erreurs dans la prévision de la production PV proviennent entre autres des perturbations météorologiques. Ces phénomènes sont majoritairement corrélés à l'échelle spatiale : épisodes climatiques régionaux, mouvements des nuages, précipitations (pluies ou neige). Nous nous proposons dans un premier temps, d'analyser les dépendances spatiales et temporelles entre les productions de différents sites d'une région donnée. L'objectif est de mettre en évidence l'existence d'une corrélation spatiale et temporelle entre ces productions. La mise en évidence de telles corrélations passe par le traitement des séries de production pour en extraire les tendances liées à la course du soleil. Puisque la variabilité liée à la course du soleil est connue avec une bonne précision, l'extraire des séries permet de se concentrer sur les autre sources de variabilité notamment celles météorologiques. Nous proposons à cet effet une nouvelle méthode de stationnarisation des séries de production PV.

La deuxième partie de ce travail est consacrée à la prévision de la production pour des horizons très courts, de quelques minutes à 5-6h ne nécessitant pas de prévisions météorologiques. Il a été démontré que pour des horizons de l'ordre de quelques minutes à 5-6h, les données historiques de production sont suffisantes pour la construction des modèles [26]. Cela s'explique par la complexité des modèles météorologiques actuels dont les procédures d'assimilation de données, de calcul et d'actualisation sont longues. Les avancées en termes de puissance de calcul permettent certes des modèles avec des résolutions spatiales et temporelles plus importantes mais les mises à jour ne sont pas fréquentes. Nous proposons donc un modèle de prévision spatio-temporel déterministe basé uniquement sur les données historiques de production. Nous investiguons aussi la possibilité d'intégrer des informations sur les conditions météorologiques locales (lieu du site de prévision) et son intérêt en termes de réduction des erreurs de prévision. Les objectifs en termes d'augmentation de la pénétration des énergies renouvelables induisent une augmentation des moyens de production et donc un grand nombre de données disponibles. Pour notre modèle spatio-temporel, ce grand nombre de données disponibles peut créer un problème de dimension. Pour pallier ce problème, nous proposons une méthode de sélection de variables qui permet d'extraire de façon automatique de la masse de données disponibles, les sous-ensembles d'informations pertinents pour la prévision.

Nous construisons dans une troisième partie un modèle de prévision probabiliste de la production PV toujours dans le cadre spatio-temporel. Les horizons de prévision vont de quelques minutes à 12h. Ce travail est porté par la nature plus complète des prévisions probabilistes et les bonnes propriétés des prévisions NWP pour les horizons supérieurs à 5-6 heures. En effet, les prévisions probabilistes permettent une meilleure connaissance de la distribution de la production PV et fournissent plus d'indications sur les erreurs de prévision, par construction d'une région de confiance pour la distribution par exemple. De plus, dans les contextes de prise de décisions les comportements/risques extrêmes présentent

plus d'intérêt que le comportement moyen. Pour fournir ces prévisions probabilistes, nous proposons un modèle spatio-temporel basé sur la prévision de quantiles conditionnels. Les prévisions météorologiques de type NWP ont été utilisées dans le modèle probabiliste car elles permettent un meilleur couplage entre les prévisions à très court-terme et celles plus classiques. L'impact de l'utilisation de ces prévisions sur les performances du modèle a été évalué.

La dernière partie de notre travail propose une technique d'intégration de données d'images satellites dans le modèle de prévision spatio-temporel déterministe. Cette intégration a pour but non seulement de prendre en compte des résolutions spatiales plus importantes mais aussi de réduire les erreurs de prévision qui pourraient être dues à l'absence d'informations au niveau des centrales voisines sur les évolutions météorologiques. De plus, dans un contexte de multiplicité des sources de données disponibles nous proposons une évaluation de l'impact de chaque source de données (centrales voisines, prévisions NWP, images satellites) sur les performances de prévision en fonction de l'horizon de prévision. A notre connaissance, cette analyse incrémentale est originale et permet d'analyser la contribution de chaque source d'informations.

1.5 Structure du manuscrit

Dans le chapitre 2 nous présentons les sources de variabilité de la production PV. Nous présentons ensuite les différentes étapes de la méthode de stationnarisation et évaluons l'efficacité de la méthode proposée. Les séries stationnarisées sont évaluées aussi bien au regard des critères formels de stationnarité que de leurs performances en prévision. Nous montrons ensuite l'existence de corrélations spatio-temporelles entre les mesures de production des différentes centrales. Ces corrélations sont calculées après stationnarisation des séries pour s'assurer de détecter le transfert d'informations relatives aux perturbations affectant la production et non la course du soleil.

Dans le chapitre 3 nous présentons le modèle de prévision spatio-temporel qui fournit des prévisions de production pour une centrale en utilisant les données des centrales voisines. Nous définissons aussi les modèles de référence, qui sont choisis parmi les modèles les plus performants de la littérature. Ces modèles de référence sont utilisés pour évaluer les performances fournies par le modèle spatio-temporel selon les critères d'évaluation que nous avons rappelé dans ce chapitre. Nous présentons aussi les développements complémentaires sur le modèle spatio-temporel. Le premier est l'intégration des conditions météorologiques locales dans le modèle de prévision. Nous présentons aussi la solution proposée au problème de dimension qui est une méthode automatique de sélection de variables. La dernière partie de ce chapitre présente les prévisions météorologiques NWP utilisées et la procédure d'intégration de cette source de données au modèle spatio-temporel. Les performances de ce modèle avec prévision NWP sont évaluées sur des horizons allant jusqu'à 12h.

Le chapitre 4 est consacré à la prévision probabiliste de la production PV. Nous présentons le modèle spatio-temporel proposé ainsi que les autres données utilisées. Un modèle spatio-temporel de référence est défini et utilisé pour évaluer les performances du modèle proposé. Nous présentons aussi une procédure de sélection de variables adaptée au modèle spatio-temporel proposé de même qu'une procédure de sélection des variables NWP intéressantes pour la prévision de la production PV. Les évaluations des modèles sont faites avec les critères spécifiques à la prévision probabiliste que nous rappelons dans le chapitre.

Le chapitre 5 traite de l'intégration des données d'images satellites aux modèles spatio-

temporel déterministe. Nous présentons les images satellites utilisées, la procédure de traitement des images satellites et de leur intégration au modèle de prévision spatio-temporel. Nous évaluons l'apport de l'utilisation de ces images sur la qualité des prévisions de production PV. Nous présentons aussi une évaluation de l'apport de chaque source de données sur les performances des prévisions en fonction de l'horizon.

Chapitre 2

Étude de la variabilité, Stationnarisation et corrélation spatio-temporelle

Dans ce chapitre, nous présentons dans un premier temps les différents jeux de données et leurs caractéristiques. Nous présentons ensuite une analyse de la variabilité de la production des centrales PV. La méthode de stationnarisation des séries de production est ensuite détaillée. Cette méthode a été présentée dans le premier article publié qui est fourni en annexe. Les séries stationnarisées par la méthode proposée sont évaluées par des tests de stationnarité et les performances de prévision. Enfin, nous mettons en évidence l'intérêt de la modélisation spatio-temporelle par le calcul des corrélations spatio-temporelles entre les mesures de production stationnarisées. La limite temporelle et spatiale de la propagation de ces corrélations est aussi évaluée.

2.1 Présentation des cas d'étude

Plusieurs données ont été exploitées dans le cadre de cette thèse. Nous présentons ici deux principaux jeux de données de centrales PV qui correspondent à des localisations, des puissances installées, des conditions climatiques et des répartitions géographiques différentes.

Les données "Coruscant"
Le premier jeu de données nommé d_1 dans toute la suite est constitué de séries temporelles de mesure de la production PV de 9 centrales situées dans le sud de la France (sauf une en région parisienne). Les centrales sont représentées sur la figure 2.1. Les puissances crête des centrales varient de 45 kWc à 5 MWc. La distance entre les centrales varie de 5 km à 783 km. Le tableau 2.1 présente les distances entre les différentes centrales labellisées $P_i, i = 1 \dots 9$. Les données couvrent une période de 20 mois à partir de juillet 2013 avec une résolution allant de 6 min à 15 min en fonction du site. Un contrôle de la qualité des données a été effectué pour enlever les valeurs aberrantes et procéder à une imputation des données manquantes. La production a été supposée nulle entre 22h et 5h du matin. Les données ont été ensuite interpolées à un pas de temps de 15 min pour la suite des travaux.

FIGURE 2.1 – Les centrales du jeu de données d_1. Les centrales sont situées dans le sud-est de la France (sauf P_3).

Tableau 2.1 – Distances (en km) entre centrales du jeu de données d_1

Distance (km)	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
P_1	0							
P_2	186	0						
P_3	680	783	0					
P_4	280	465	638	0				
P_5	10	190	670	280	0			
P_6	55	235	670	230	58	0		
P_7	6	185	690	280	15	55	0	
P_8	183	20	800	460	188	229	179	0
P_9	185	17	801	460	188	230	180	3

Les données "Hespul"

Le second jeu de données est un bon exemple d'un cas d'étude comportant un nombre important de centrales avec une répartition géographique très dense. Ce cas d'étude comporte les données de 905 onduleurs installés dans la région centre-ouest de la France. Les puissances crêtes varient de 3.2 kWc à 58 kWc. Ces onduleurs correspondent à 185 centrales photovoltaïques différentes. Les centrales sont présentées sur la figure 2.2. La distance entre les centrales varie de 1 km à 230 km. Les données de production couvrent la période de novembre 2014 à mars 2016. Le même traitement des données que celui du premier cas d'étude a été effectué sur ce second cas d'étude et les données ont été interpolées à une résolution commune de 15 min. Après traitement des données, 136 centrales ont été retenues.

2.2 Analyse des données de production

La production PV est caractérisée par une variabilité importante. Les principales sources de cette variabilité sont la saisonnalité et les conditions climatiques. Si les variations saisonnières sont assez bien prévisibles car liées à la course du soleil, celles liées au climat notamment à la couverture nuageuse, le sont beaucoup moins.

Impact de la saisonnalité

On peut caractériser l'impact de la saisonnalité sur la production PV par :

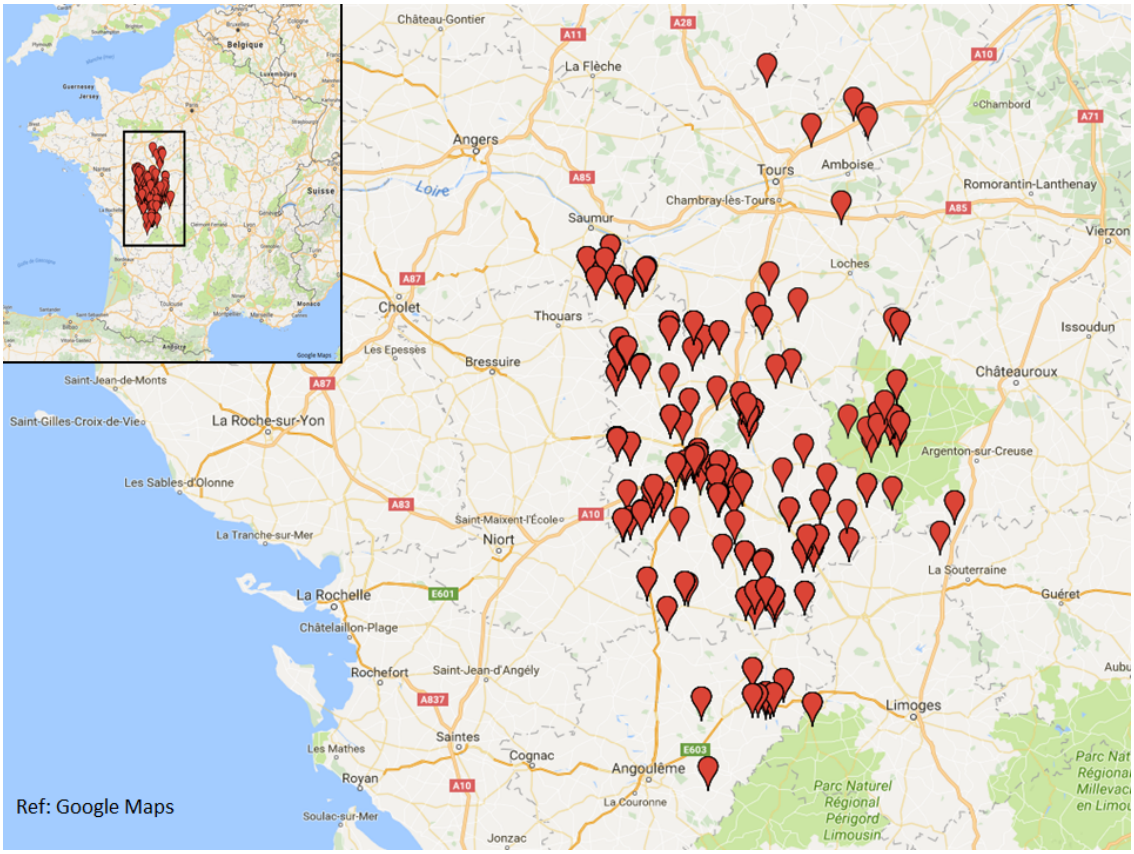


FIGURE 2.2 – Les centrales du jeu de données d_2. La distance entre les centrales varie de 1 km à 230 km. Les centrales sont situées dans le centre-ouest de la France.

— une production nulle la nuit ;

— un cycle journalier avec une pointe de production aux environs de midi (voir figure 2.4) ;

— un cycle annuel avec des niveaux de production élevés l'été et qui baissent significativement l'hiver (voir figure 2.3).

Les différentes variations que l'on peut observer sur la production journalière d'une centrale (voir figure 2.4) sont essentiellement dues aux variations météorologiques.

Variabilité de la production

La variabilité de la production d'une installation PV ou d'un réseau de plusieurs installations PV peut être analysée par différents critères. Dans [100], on retrouve un outil d'analyse de cette variabilité appelé "variabilité relative de la production". On définit la variabilité d'un ensemble de N installations PV par :

$$\sigma_{\Delta t}^{\Sigma N} = \left(\frac{1}{C^{Fleet}}\right)\sqrt{Var\left[\sum_{n=1}^{N}\Delta P_{\Delta t}^{n}\right]} \tag{2.1}$$

où C^{Fleet} est la capacité totale installée de l'ensemble des N installations PV. $\Delta P_{\Delta t}^{n}$ est la série temporelle des évolutions de la production pour la centrale n :

$$\Delta P_{\Delta t}^{n} = \left\{(t_1, \Delta P_{t_1,\Delta t}^{n}), (t_2, \Delta P_{t_2,\Delta t}^{n}), \ldots, (t_T, \Delta P_{t_T,\Delta t}^{n})\right\} \tag{2.2}$$

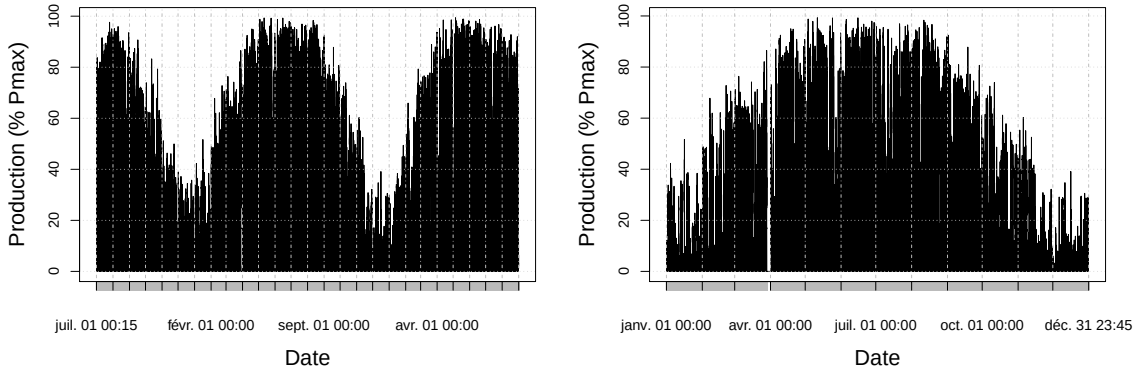


FIGURE 2.3 – Production photovoltaïque d'une centrale PV de juillet 2013 à août 2015 (gauche) et sur l'année 2014 (droite)

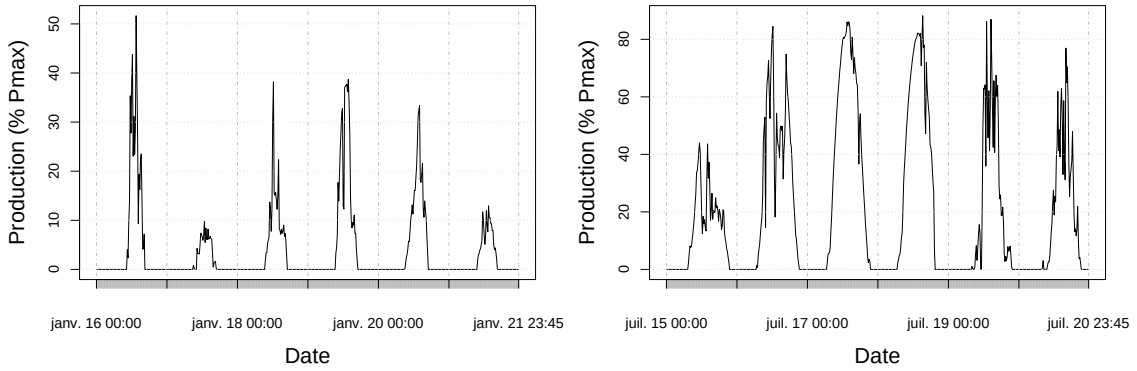


FIGURE 2.4 – Production photovoltaïque pour des semaines d'hiver (gauche) et d'été (droite)

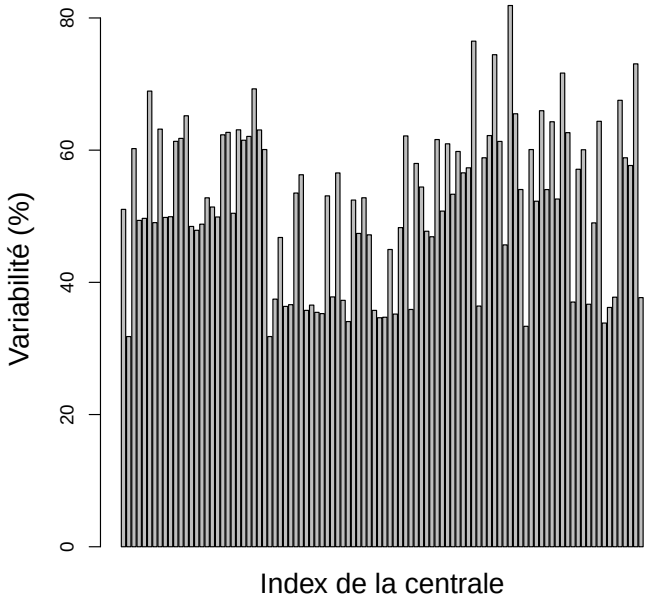
et $\Delta P^n_{t,\Delta t} = P^n_t - P^n_{t+\Delta t}$. La variabilité relative ou ROV (Relative Output Variability) d'une centrale i d'un ensemble de N centrales est définie comme :

$$ROV_i = \frac{\sigma^{\Sigma N}_{\Delta t}}{\sigma^{\Sigma i}_{\Delta t}}. \tag{2.3}$$

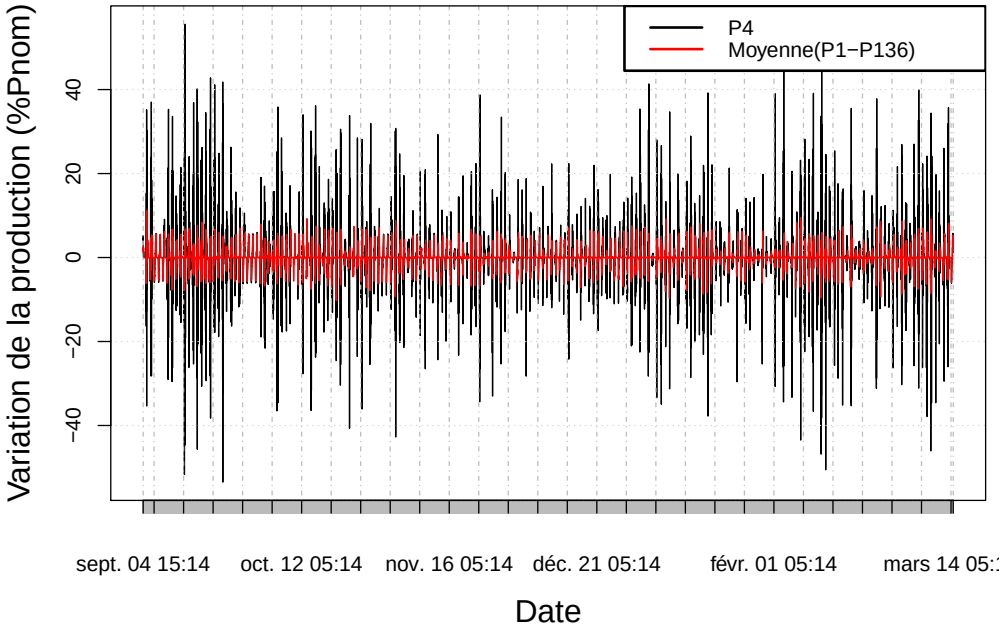
La figure 2.5a représente pour 100 centrales du jeu de données d_2 la variabilité relative en fonction de la centrale. Les valeurs de variabilité observées varient entre 30% et 80% traduisant ainsi une forte variabilité de la production PV. La variabilité peut être aussi analysée visuellement en s'inspirant de la série des différences proposée dans [100]. On définit donc une nouvelle série des différences P^d_t qui permet de visualiser les variations de la production PV au pas de temps k, $P^d_t = P_t - P_{t+k}$. La figure 2.5b représente la série des différences ainsi définie pour la centrale P_1 et l'ensemble (moyenné) des centrales du jeu de données d_2.

2.3 Méthode de stationnarisation proposée

Le développement des moyens de production d'énergie renouvelable at entrainé de nouveaux besoins en termes de prévision. Les méthodes de prévision basées sur les modèles



(a) Variabilité relative de 100 centrales du jeu de données d_2 sur la base d'un pas de temps de 15 min



(b) Variation (série P_t^d) de la production PV pour les centrales du jeu de données d_2. Le pas de temps est 15 min.

FIGURE 2.5 – Analyse de la variabilité de la production PV

numériques ne permettent pas de répondre efficacement à la maitrise de la variabilité de la production notamment à très court-terme (quelques minutes à quelques heures). Hormis la variabilité due à la course du soleil, les erreurs observées dans le cadre de la prévision de production des énergies renouvelables sont liées à la faible performance des modèles pour l'anticipation des perturbations météorologiques. Ces perturbations sont la plupart du temps des phénomènes qui se propagent spatialement. La prédictibilité de

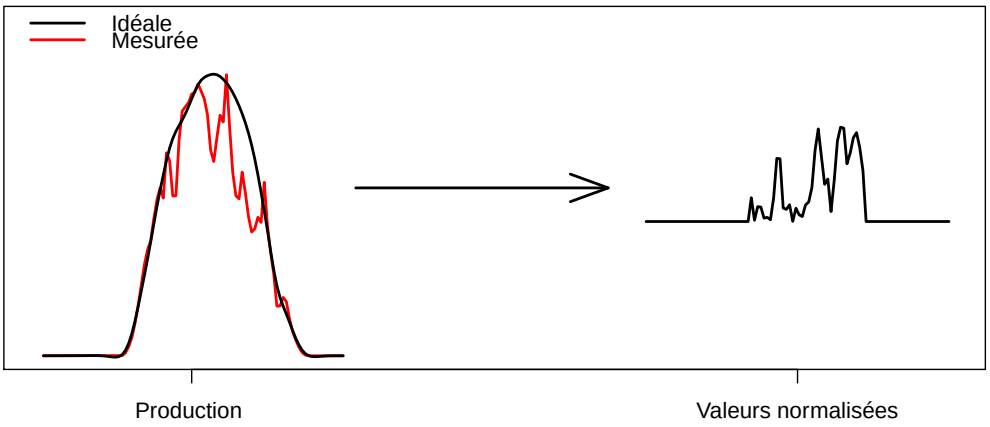


FIGURE 2.6 – Illustration du principe pratique de la stationnarisation

la production d'énergie renouvelable peut donc être améliorée par la prise en compte de ces perturbations spatiales en complément de l'analyse de la variabilité temporelle. C'est dans ce cadre que nous nous proposons de développer un modèle spatio-temporel qui exploite les informations spatiales et temporelles entre les mesures de production de différents sites de production PV. L'utilisation des mesures d'un ensemble de site de production PV pour mieux anticiper les phénomènes météorologiques nécessite au préalable de s'affranchir de la variabilité due à la course du soleil pour se concentrer sur l'effet des autres sources de variabilité de la production, notamment les nuages. L'extraction de la variabilité due à la course du soleil passe par la stationnarisation des séries de production. Cette stationnarisation permettra de répondre à la question de l'existence de corrélations spatio-temporelles non liées à la course du soleil entre les mesures de production et ainsi de l'opportunité de la mise en œuvre d'un modèle spatio-temporel. De plus, cette stationnarisation est utile pour la mise en oeuvre de certaines méthodes de prévisions classiques (non spatio-temporelles) qui font des hypothèses de stationnarité sur les données d'entrée.

En pratique, la stationnarisation consiste à "aplatir" la cloche (due à la course du soleil) dans les séries journalières de production pour ne garder que les autres variations qui sont liées aux perturbations météorologiques. La figure 2.6 montre l'objectif pratique visé par la sationnarisation. Sur la gauche de la figure, on voit deux courbes de production PV normalisées : celle idéale sans perturbation météorologique (courbe en noir) et celle réalisée (courbe en rouge). La production mesurée (réalisée) présente des variations dues aux perturbations météorologiques que l'on veut isoler par rapport à la courbe idéale. On veut donc obtenir la courbe de droite qui comporte uniquement les variations qui ne sont pas dues à la course du soleil.

2.3.1 Présentation de la méthode

Nous proposons ici une nouvelle méthode de stationnarisation des séries de production PV. Nous n'utilisons pas ici la différenciation des séries de production car elle nécessite une intégration a posteriori des erreurs. Cette intégration entraine une explosion des intervalles de confiance. Nous présentons dans la suite le cheminement qui a conduit à la méthode de stationnarisation proposée ainsi que les différentes analyses menées pour évaluer l'efficacité de ladite méthode.

Définition d'un indice de ciel clair pour la production PV
La méthode que nous proposons s'inspire de l'indice du ciel clair pour le rayonnement solaire [27, 22, 101]. Cet indice représente la façon dont l'atmosphère atténue la lumière d'une heure à l'autre ou au jour le jour en fonction du mouvement de la Terre autour du Soleil. Nous le définissons ici comme le rapport entre les mesures d'irradiation I_t^{meas} et celles estimées par modèle de ciel clair I_t^{sim} à l'instant t :

$$k_t^{irr} = \frac{I_t^{meas}}{I_t^{sim}}. \tag{2.4}$$

Dans le même esprit, nous définissons un indice de ciel clair pour la production photovoltaïque k_t^{pv} tel que :

$$k_t^{pv} = \frac{P_t^{meas}}{P_t^{sim}}, \tag{2.5}$$

avec P_t^{meas} la production PV mesurée à l'instant t, et P_t^{sim} la production sous hypothèses de ciel clair obtenue par transformation de l'irradiation fournie par un modèle ciel-clair en production pour le temps t. En pratique, P_t^{sim} est construit comme le produit du facteur d'efficacité η du système PV et de l'irradiation simulée I_t^{sim} sous des conditions ciel-clair :

$$P_t^{sim} = I_t^{sim} \times \eta. \tag{2.6}$$

Le paramètre η intègre le rendement du module, la surface active et les pertes.

L'indice ciel-clair pour la production k_t^{pv} ayant été défini, notre première idée pour l'appliquer a été d'utiliser les séries d'irradiation ToA (Top of Atmosphere) pour la stationnarisation. Cette irradiation est celle reçue par une surface horizontale à l'extérieur de l'atmosphère. Le choix de cette série d'irradiation est porté par le fait que l'irradiation ToA n'est affectée par aucun effet atmosphérique. Elle porte donc exclusivement la variabilité due à la course du soleil si on fait l'hypothèse de négliger les effets extra-terrestres. Les valeurs attendues sont en théorie comprises entre 0 et 1. La figure 2.7 montre la série résultant de la normalisation par le ToA pour l'année 2014 pour une centrale dans le sud de la France. En pratique, on peut observer une part non négligeable de valeurs plus grandes que 1, surtout pour les périodes avec les niveaux de production les plus faibles. Le traitement des périodes à faible niveau d'irradiation n'est pas satisfaisant. L'examen des autocorrélations de la série présentées dans la figure 2.8 montre une décroissance rapide qui confirme la non-stationnarité des séries.

L'utilisation des séries d'irradiation ToA ne permettant pas d'obtenir des séries stationnaires, il a été envisagé d'utiliser les séries d'irradiation ciel-clair. Le modèle ciel-clair retenu est le modèle ESRA [22]. Ce modèle, plus complet, permet de prendre en compte les paramètres de l'atmosphère à savoir l'effet des aérosols et de l'absorption des gaz. La figure 2.9 illustre les différences entre les irradiations ToA et ESRA.

Les limites de la normalisation par le modèle ciel-clair
Comme présenté dans l'état de l'art sur les modèles de ciel clair (partie 1.3), les méthodes de stationnarisation des séries de production PV (pas celle d'irradiation) par modèle de ciel-clair proposées dans la littérature sont très peu détaillées, ce qui les rend difficiles à reproduire. De plus, dans la littérature on ne retrouve pas d'analyse formelle sur l'efficacité des méthodes de stationnarisation, surtout avec des critères de vérification de stationnarisation. La normalisation par l'irradiation ciel-clair présente des limitations. La figure 2.10 présente pour une centrale située dans le sud-est de la France, les séries de production

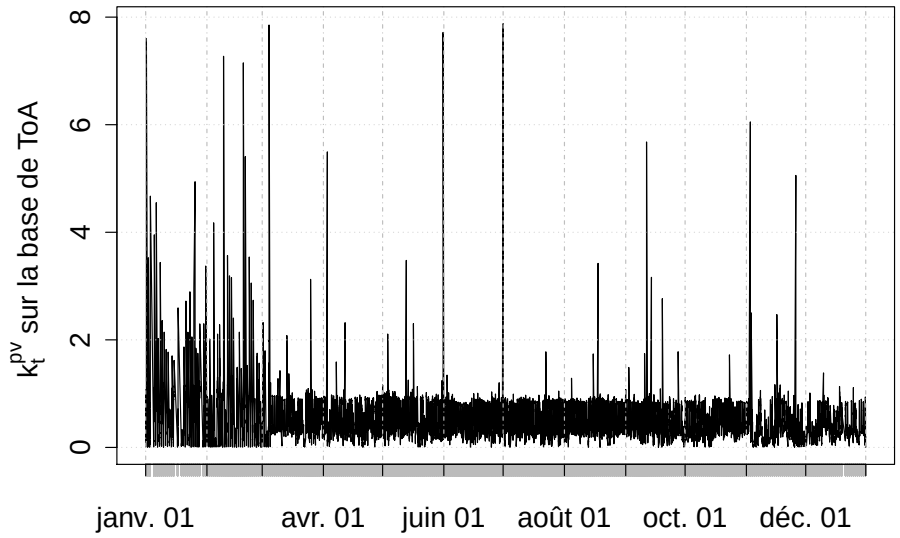


FIGURE 2.7 – Série normalisée sur la base de l'irradiation ToA d'une centrale située dans le sud-est de la France pour l'année 2014.

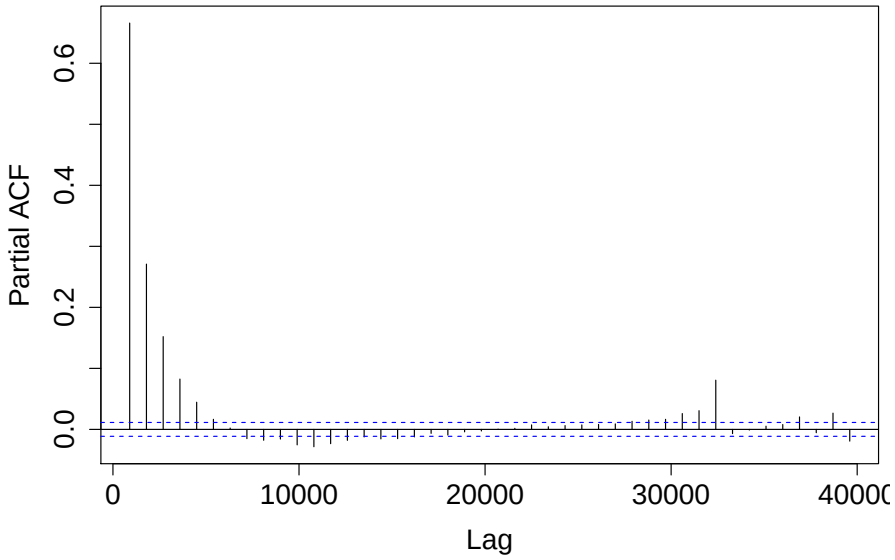


FIGURE 2.8 – PACF de la série normalisée sur la base de l'irradiation ToA d'une centrale située dans le sud-est de la France pour l'année 2014.

normalisées par la puissance maximale et celles normalisées par une série d'irradiation ciel-clair obtenue par le modèle ESRA. La simulation de l'irradiation a été effectuée sous l'hypothèse d'une surface horizontale en admettant que la variation de la production en sortie due au niveau d'inclinaison est assimilée par le facteur d'efficacité η. La principale limitation de cette stationnarisation concerne les faibles valeurs de production observées en début et fin de journée qui ne sont pas traitées. Ce constat est le même pour la méthode basée uniquement sur l'apprentissage statistique proposée par Bacher [26] qui exclut tout simplement les faibles valeurs d'irradiation.

Normalisation par une fonction de l'irradiation ciel-clair

La figure 2.11 présente la relation entre les valeurs normalisées de la production et l'irradia-

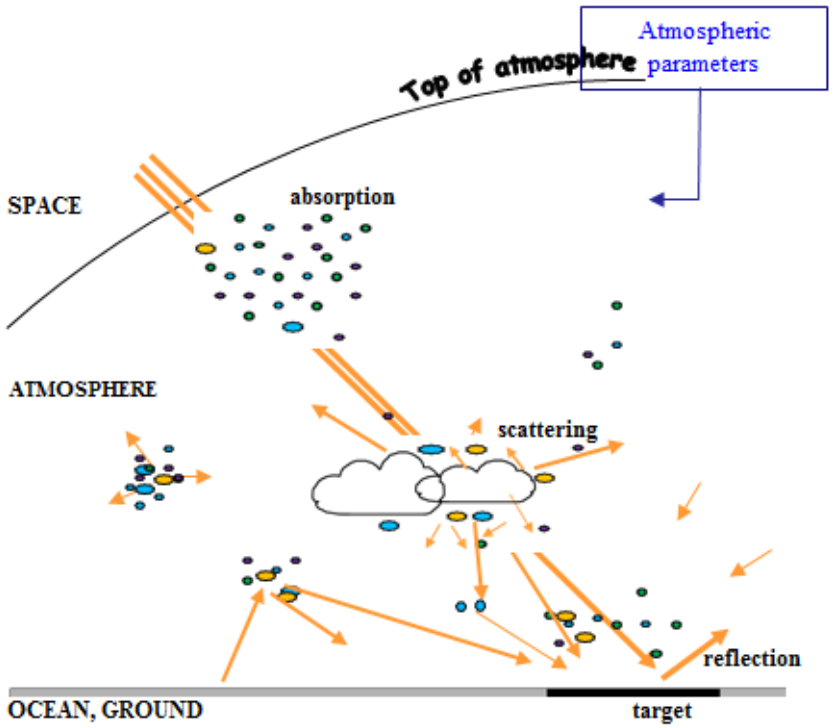


FIGURE 2.9 – Différences entre les irradiations ToA et le modèle ciel-clair ESRA. Le modèle ciel-clair ESRA prend en compte les différentes interactions qui ont lieu dans l'atmosphère.

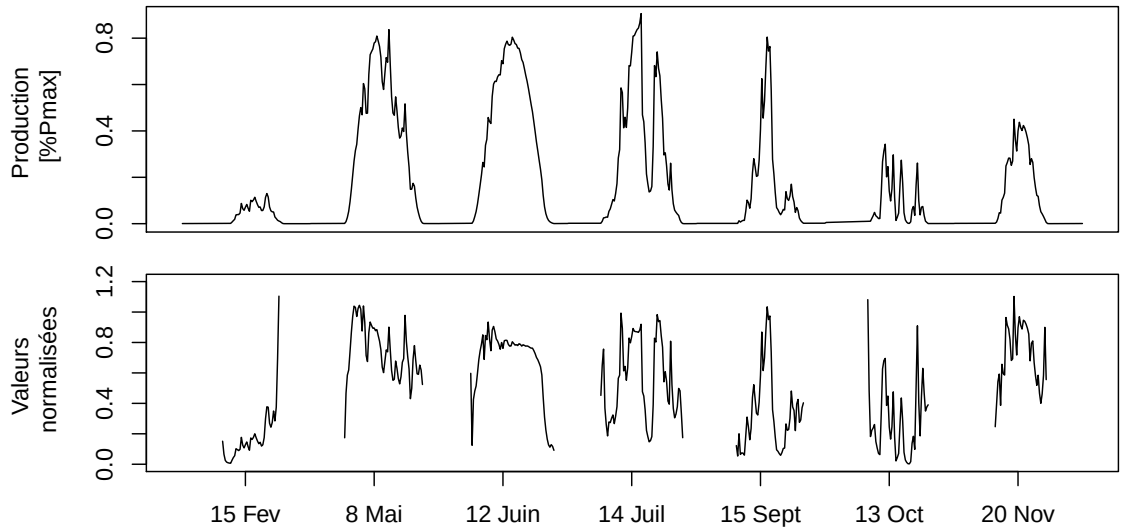


FIGURE 2.10 – Exemple de séries de production normalisées par la puissance maximale et de séries normalisées sur la base de l'irradiation ciel-clair ESRA pour différents jours de l'année 2014. La centrale est située dans le sud-est de la France.

tion ESRA pour deux périodes temporelles définies comme avant et après le midi solaire. Les deux classes de valeurs correspondent aux phases de croissance de la production (du

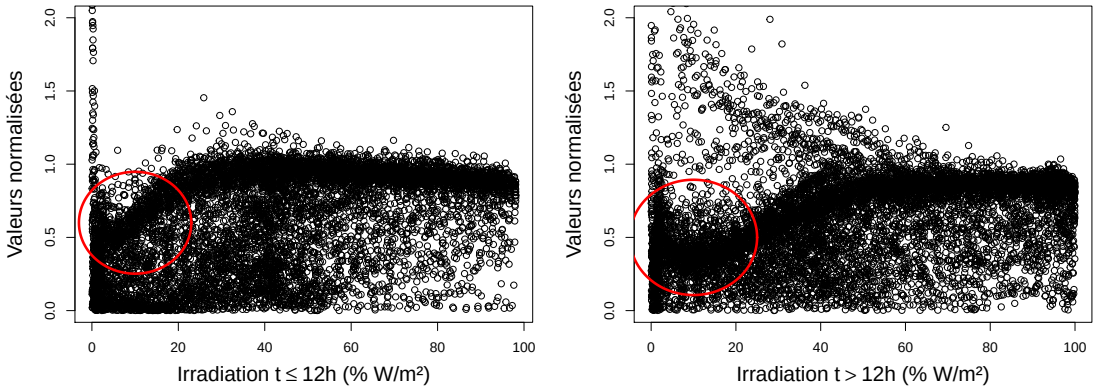


FIGURE 2.11 – Relation entre valeurs normalisées de production et irradiation ESRA selon le moment de la journée pour l'année 2014. La centrale est située dans le sud-est de la France.

lever du soleil au midi solaire) et de décroissance, c'est-à-dire du midi solaire au coucher du soleil. L'analyse de la figure montre (cercle rouge) l'inefficacité de la normalisation par l'irradiation ESRA pour les faibles valeurs de production.

Pour améliorer la qualité des séries normalisées et résoudre les problèmes spécifiques aux périodes de faible production comme les débuts et fins de journée, nous proposons une nouvelle relation entre la production réelle et P_t^{sim} en utilisant une fonction de normalisation f. Cette fonction sert à modéliser le lien entre les deux productions (réelle et simulée) dans le but de produire une nouvelle série u_t plus stationnaire définie pour les heures auxquelles P_t^{sim} est non nulle par :

$$u_t = P_t / f(P_t^{sim}). \tag{2.7}$$

Plusieurs formes ont été envisagées pour la fonction de normalisation f à savoir linéaire, linéaire par morceaux ou quadratique. Le choix de la fonction de normalisation appropriée a été effectué en utilisant un critère quantitatif basé sur l'évolution de l'écart-type journalier de la série u_t. En pratique, pour chaque type de fonction de normalisation f précédemment énuméré, nous construisons la série des écarts-types journaliers de u_t. La fonction f retenue est celle pour laquelle la série u_t présente une moyenne indépendante du temps et des écart-types journaliers de plus faible variabilité intra-journalière.

La fonction de normalisation que nous avons retenue pour corriger les défauts de la simple normalisation par l'irradiation est linéaire par morceaux et dépend du sens de l'évolution quotidienne de la production (soit une augmentation au début de la journée, soit une diminution après le midi solaire). L'idée est de corriger les faibles valeurs d'irradiation dans la normalisation par l'ajout d'une pénalisation. Cette pénalisation permet aussi de corriger le nombre non négligeable de valeurs de la série normalisée qui excèdent la valeur théorique de la borne maximale qui est 1. La fonction retenue s'écrit :

$$f(P_t^{sim}) = P_t^{sim} + f_a(P_t^{sim}) + f_b(P_t^{sim}) \tag{2.8}$$

avec f_a définie du lever du soleil au midi solaire et f_b du midi solaire au coucher du soleil. Les fonctions f_a et f_b participent à améliorer la stationnarisation particulièrement pour les faibles niveaux de production en début et fin de journée. Elles sont définies pour chaque jour par :

$$\left\{\begin{array}{rcl} f_a(0) & = & \alpha_a \\ f_a\left(\beta_a \frac{P_{max}^{sim}}{2}\right) & = & 0 \\ f_a(P_{max}^{sim}) & = & \gamma \end{array}\right. \quad \left\{\begin{array}{rcl} f_b\left(P_{max}^{sim}\right) & = & \gamma \\ f_b\left(\beta_b \frac{P_{max}^{sim}}{2}\right) & = & 0 \\ f_b(0) & = & \alpha_b \end{array}\right. \tag{2.9}$$

où P_{max}^{sim} représente la production maximale simulée pour la journée. La continuité en le midi solaire (ou maximum journalier d'irradiation sous hypothèses ciel-clair) est assurée par le coefficient γ. Les valeurs des coefficients $\alpha_a, \alpha_b, \beta_a, \beta_b, \gamma$ sont obtenues grâce à un processus d'optimisation qui vise à minimiser le critère d'écart-type. L'optimisation est faite sous les contraintes $\beta_{a,b} \in (0, 2)$. Les coefficients sont initialisés de manière aléatoire, une fenêtre d'un mois de valeurs de la série d'irradiation ESRA est choisie avant le jour d'intérêt. Les coefficients sont choisis sur la fenêtre de valeurs d'irradiation en optimisant l'écart-type de la série normalisée.

En pratique, on évalue pour plusieurs valeurs de coefficients, l'écart-type de la série normalisée et on retient les coefficients qui minimisent l'écart-type. Le critère d'optimisation des coefficients est donc directement lié aux propriétés de stationnarité de la série d'intérêt. La stationnarité de la forme normalisée de u_t a été évaluée en analysant son auto-corrélogramme et aussi par des tests de racine unitaire. La figure 2.12 présente les résultats de l'optimisation des coefficients à utiliser pour le processus de stationnarisation du jeu de données d_1. La figure montre que les coefficients les plus variables sont les α_a et α_b qui permettent de corriger les débuts et fins de journée.

La procédure de stationnarisation peut être résumée pour une centrale PV par les étapes ci-après :

1. Nettoyage des données aberrantes de la série de production PV
2. Simulation de la série d'irradiation ciel clair ESRA et de la série de puissance correspondante (équation (2.6))
3. Détermination des coefficients appropriés des fonctions (f_a, f_b) en utilisant un processus d'optimisation sur un intervalle glissant des valeurs d'irradiation simulées
4. Normalisation la série mesurée P_t^{meas} pour obtenir la série u_t.

Le graphique 2.13 présente pour les centrales du jeu données d_2 les corrélations entre couples de centrales en fonction de la distance avant et après stationnarisation. Il permet d'apprécier visuellement l'effet de la stationnarisation sur les corrélations entre couples de centrales ; on constate une baisse significative de ces corrélations et une répartition moins concentrée autour des fortes valeurs.

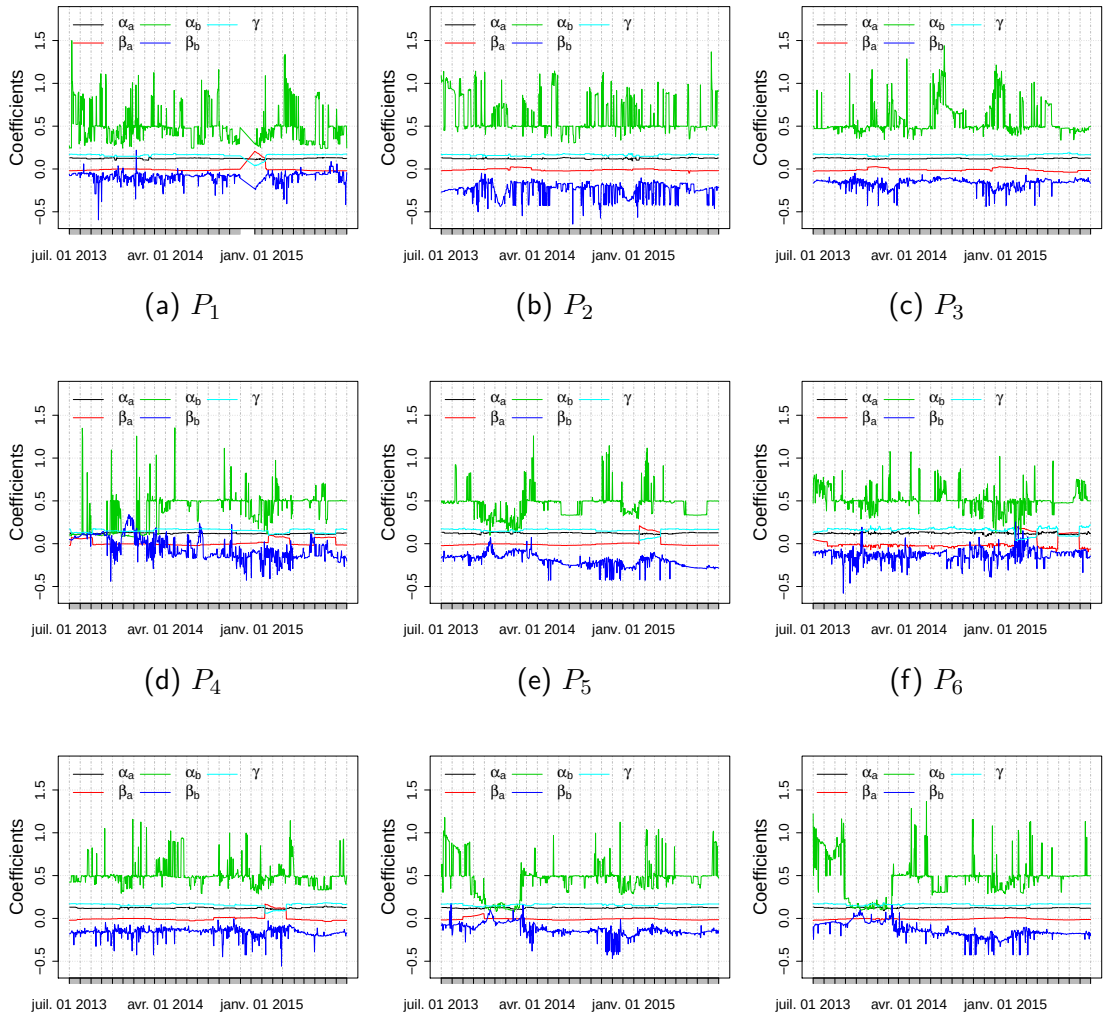
2.3.2 Étude des performances de la méthode de stationnarisation

Étude des autocorrélations

La méthode de stationnarisation proposée est comparée à la normalisation par la série d'irradiation ToA. Les figures 2.14a et 2.14b représentent les fonctions d'autocorrélations partielles (PACF) des séries stationnarisées selon les deux méthodes pour la centrale P_1 du jeu de données d_1. L'examen de ces autocorrélations montre que la méthode de stationnarisation proposée permet une meilleure correction (pour les lags élevés, la valeur de PACF est toujours contenue dans l'intervalle de confiance) des séries dans l'objectif d'appliquer un modèle autorégressif.

Tests de stationnarité

La notion de stationnarité faible, ou au second ordre se définie par l'invariance des moments d'ordre 1 et 2 au cours du temps. Il existe différentes sources de non-stationnarité.



(a) P_1 (b) P_2 (c) P_3

(d) P_4 (e) P_5 (f) P_6

(g) P_7 (h) P_8 (i) P_9

FIGURE 2.12 – Évolution journalière des coefficients des fonctions de stationnarisation pour chacune des centrales ($P_1 - P_9$) du jeu de données d_1.

La source de non stationnarité peut être déterministe (trend stationary) ou stochastique (difference stationary). En pratique, il est difficile de vérifier de manière rigoureuse l'hypothèse de stationnarité du second ordre. On utilise l'hypothèse de stationnarité intrinsèque qui est une version affaiblie de l'hypothèse de stationnarité du second ordre. Cette hypothèse intrinsèque stipule que les accroissements de la fonction aléatoire sont stationnaires au second ordre c-a-d l'espérance des accroissements est nulle et sa variance ne dépend que du vecteur séparant les deux points.

Pour évaluer l'efficacité de la méthode de stationnarisation, une première étude avec le test de Dickey-Fuller augmenté (ADF) [102, 103] a été réalisée. C'est un test de présence de racine unitaire ou test de stationnarité intrinsèque. En effet, si on considère un processus Y_t et $\phi(L)$ le polynôme de l'opérateur de retard L , si le processus $\phi(L)Y_t$ est stationnaire, si 1 est une racine du polynôme ϕ alors le processus Y_t sera non stationnaire. Cela explique pourquoi la plupart des tests de stationnarité sont des tests de détection de racine unitaire.

Les hypothèses du test de Dickey-Fuller augmenté pour un processus Y_t sont les suivantes :

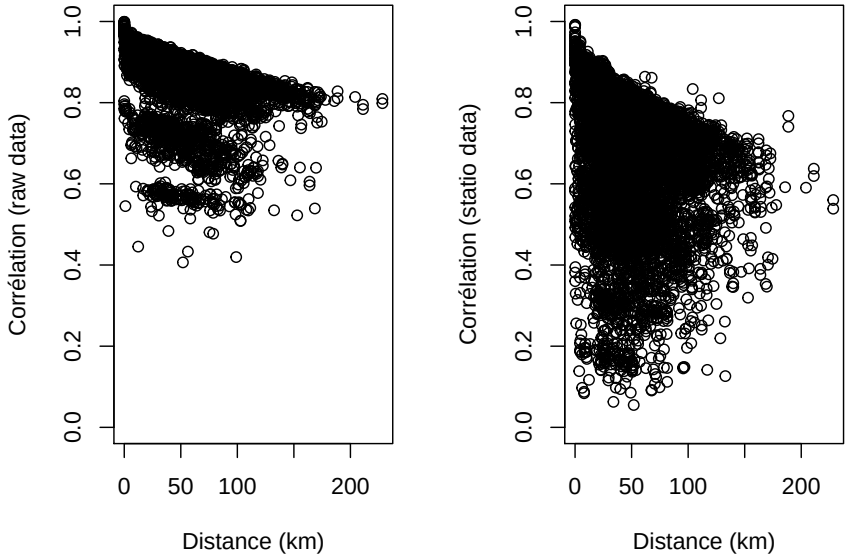
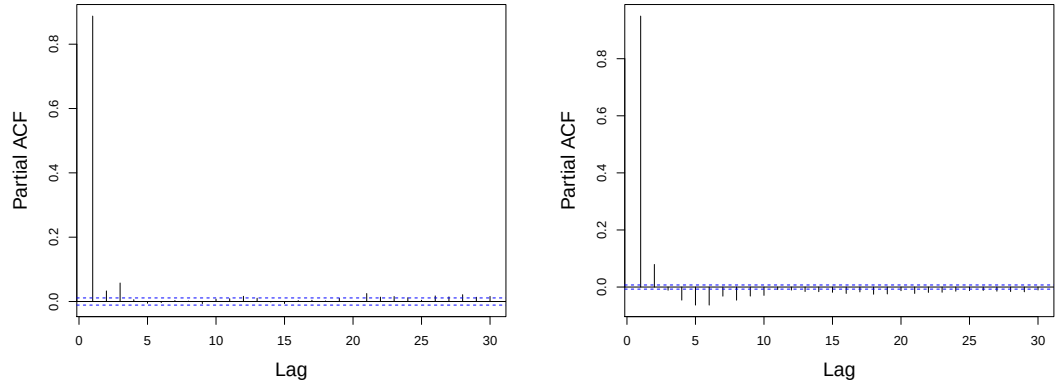


FIGURE 2.13 – Corrélation entre couples de centrales en fonction de la distance avant stationnarisation (à gauche) et après stationnarisation (à droite).



(a) PACF série normalisée par la méthode proposée

(b) PACF série normalisée par l'irradiation ToA

FIGURE 2.14 – Auto-corrélation partielle (PACF) des séries normalisées

(H0) : Il y a présence de racine unitaire donc le processus Y_t est non stationnaire.
(H1) : Il n'y a pas de racine unitaire donc le processus Y_t est stationnaire. Puisqu'on teste la stationnarité intrinsèque, on travaille sur la régression :

$$\Delta Y_t = d_t + \sum_{i=1}^{p} \gamma_i \Delta Y_{t-i} + \rho Y_{t-1} + u_t \tag{2.10}$$

avec d_t une fonction déterministe du temps et u_t les erreurs.

La statistique de Dickey-Fuller est alors définie comme :

$$\Delta D_\rho = \frac{D_\rho}{1 - \sum_{i=1}^{p} \gamma_i} \tag{2.11}$$

Tableau 2.2 – Valeur des statistiques de Dickey-Fuller pour les séries normalisées sur la base du ToA et suivant la méthode proposée pour les centrales $P_1 - P_4$

Centrales	Série normalisée par le ToA	Série normalisée par la méthode proposée
P_1	-2.00	-20.35
P_2	-2.35	-17.28
P_3	-1.14	-18.64
P_4	-2.78	-20.64

où D_ρ est la statistique de test.

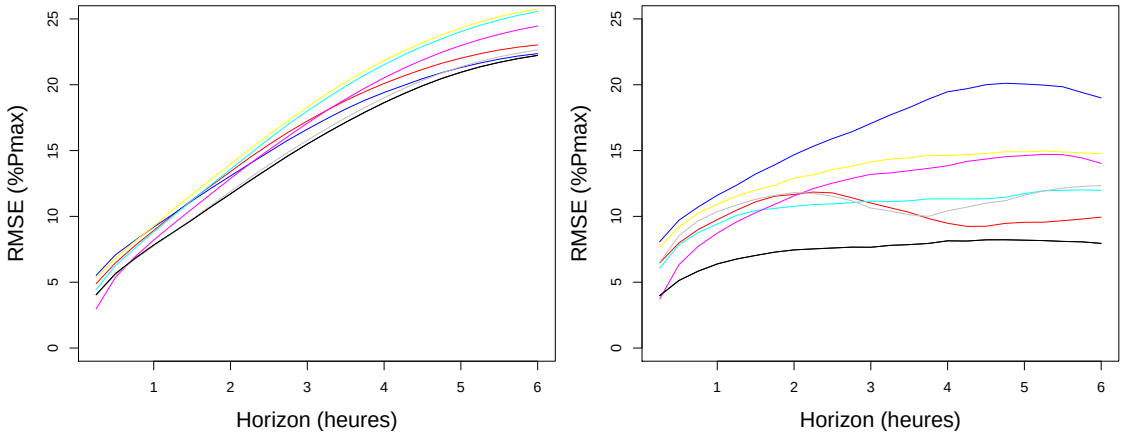
Dans le tableau 2.2, nous présentons les valeurs de statistique de Dickey-Fuller et de seuil pour les séries normalisées par l'irradiation ToA et celles stationnarisées par la méthode proposée précédemment pour quatre centrales du jeu de données d_2. On remarque que pour les séries normalisées suivant la méthode proposée, on rejette l'hypothèse de non stationnarité car les valeurs des statistique sont inférieures aux valeurs seuils à 10% et 5% qui sont respectivement -3.41 et -3.12. Le résultat est inverse pour la série normalisée par le ToA où l'hypothèse nulle de non stationnarité n'est pas rejetée. Le constat est le même pour la plupart des centrales du jeu de données. La méthode de stationnarisation proposée permet donc d'obtenir les conditions de stationnarité au sens du critère de Dickey-Fuller.

Performances des séries stationnarisées pour la prévision

L'objectif premier de la stationnarisation est d'améliorer les performances des méthodes de prévision. Au-delà des tests de stationnarité et de l'examen des autocorrélations, il convient donc d'évaluer l'efficacité de la méthode de stationnarisation proposée à travers l'examen des performances des séries stationnarisées.

Un modèle autorégressif (AR) a été utilisé pour prévoir la production PV pour les deux types de séries de chaque jeu de données à savoir les séries stationnarisées selon la procédure proposée et celles brutes (normalisées par la valeur maximale observée). Le critère du RMSE a été utilisé pour évaluer les performances. Pour l'ensemble des centrales du jeu de données d_1, l'amélioration moyenne du RMSE obtenue en utilisant les séries stationnarisées est de 7%. Les erreurs de prévision selon le critère RMSE sont représentées sur la figure 2.15 pour les séries stationnarisées ou non. Elles illustrent encore plus clairement l'apport de la stationnarisation dans le cadre de l'amélioration des performances de prévision. Les séries stationnarisées permettent donc d'obtenir de meilleures performances de prévision que la simple normalisation par les valeurs maximales. Une analyse plus précise de l'apport de la méthode de stationnarisation sur les performances en prévision pour les débuts et fins de journée a été effectuée. Les plages horaires de 6h à 9h et de 19h à 22h ont respectivement été utilisées pour calculer les critères d'évaluation. L'amélioration moyenne du MAE pour ces deux tranches horaires lorsqu'on utilise les séries stationnarisées plutôt que les données brutes est respectivement de 8% et 9%.

La même étude a été réalisée sur les centrales du jeu de données d_2. La figure 2.16 représente l'amélioration du RMSE obtenue en utilisant les séries stationnarisées. L'amélioration moyenne pour un horizon de 3 heures est de 10% et peut monter jusqu'à 15%. Cette réduction significative des erreurs de prévisions confirme l'efficacité de la méthode de stationnarisation et son intérêt à l'utiliser pour prétraiter les données avant utilisation dans les modèles de prévision.



(a) Série non stationnarisée (b) Série stationnarisée

FIGURE 2.15 – Erreurs de prévision RMSE d'un modèle AR avec des séries stationnarisées ou non pour le jeu de données d_1. Chaque ligne représente une centrale PV.

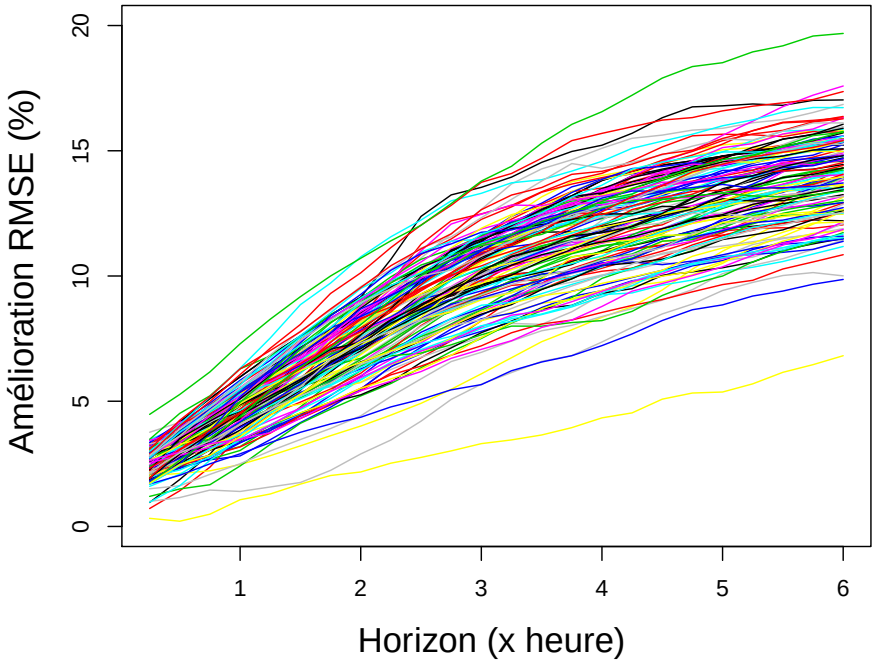


FIGURE 2.16 – Jeu de données d_2 : Amélioration du RMSE pour un modèle AR avec des séries stationnarisées par rapport à des séries non stationnarisées. Chaque ligne représente une centrale. Le pas de temps est de 15 min.

2.4 La corrélation spatio-temporelle

La prévision spatio-temporelle de la production PV au sens de l'utilisation des mesures de centrales voisines (comme un réseau de capteurs) repose sur l'existence de lien spatio-temporel entre ces centrales. Dans cette partie, nous justifions la création d'un modèle spatio-temporel par la mise en évidence de liens spatio-temporels entre des centrales spatialement distribuées. Pour cela, nous effectuons une analyse basée sur des indicateurs spécifiques aux liens spatiaux ainsi qu'une étude des corrélations des mesures de

production des différents sites.

2.4.1 Etude du lien spatial

La dépendance de la production PV au temps est une caractéristique du phénomène de production PV (voir figures 2.3 et 2.4). Il s'agit donc d'en analyser les occurrences spatiales. La production PV est donc considérée comme une variable régionalisée avec autant d'observations que de sites de production. Le lien spatial entre ces différentes observations est étudié avec différents critères incluant l'indice de Moran et les corrélogrammes.

Indice de Moran

La statistique la plus utilisée pour tester l'autocorrélation spatiale dans une série définie spatialement est celle de Moran [104]. Cette statistique s'écrit formellement de la façon suivante pour une variable régionalisée de production Y :

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_i (Y_i - \bar{Y})^2} \tag{2.12}$$

avec Y la production PV sur un réseau discret de N sites ; W une matrice carrée de poids positifs de dimension N tel que $w_{i,j}$ quantifie les influences du site j sur le site i. La matrice de poids que nous avons retenue ici est une matrice de distance. On peut retrouver dans la littérature des matrices de contiguïté mais elles ne sont pas adaptées au cas de la production PV qui est faite sur des sites et non des régions entières, ce qui rend difficile la définition de voisinage.

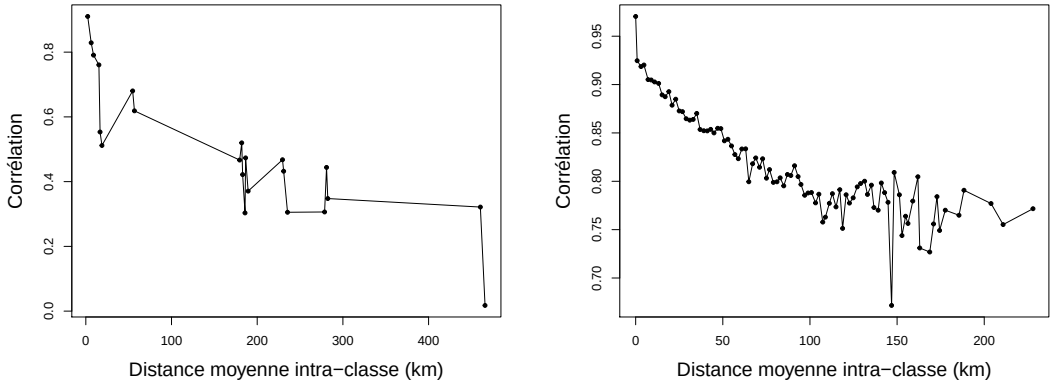
La statistique de Moran centrée et réduite suit asymptotiquement une loi normale d'espérance nulle et de variance unitaire et sert ainsi de base au test de l'autocorrélation spatiale dans une série. Lorsque I est proche de 1 en valeur absolue on a autocorrélation (positive si I l'est, négative sinon) ; une valeur proche de 0 en valeur absolue traduit l'indépendance des observations.

Le calcul de cette statistique sur les jeux de données d_1 et d_2 à un pas de temps horaire fournit des valeurs moyennes respectives de l'indice de Moran de 0.37 et 0.66. La valeur est plus faible pour le jeu de données d_1 car les centrales y sont plus distantes. Ces valeurs significatives traduisent la présence de corrélations spatiales entre les différentes centrales. La corrélation étant plus forte dans le cas du jeu de données plus dense.

Corrélogramme "modifié"

Les auto-corrélogrammes sont utilisés pour étudier l'existence de corrélation spatiale entre les mesures de production. Nous utilisons ici une version modifiée des méthodes classiques de calcul d'auto-corrélogramme [105]. Le principe de cette méthode peut être résumé par les étapes suivantes :

1. Répartir les corrélations entre les paires de centrales en classes en fonction des distances séparant les sites
2. Dans chaque classe, tester la significativité des corrélations par des tirages aléatoires de valeur de corrélation croisée de sorte à ce qu'une centrale soit utilisée une seule fois. Par exemple si la corrélation entre A et B est choisie, toutes les autres paires de combinaison intégrant A et B sont exclues de la base de tirage.
3. Répéter la procédure jusqu'à ce qu'il n'y ait aucune centrale inutilisée
4. Déterminer le nombre de coefficients de corrélation négatifs et positifs



(a) Auto-corrélogramme des centrales du jeu de données d_1 excepté la centrale P_3

(b) Auto-corrélogramme des centrales du jeu de données d_2

FIGURE 2.17 – Corrélogrammes spatiaux modifiés de la production PV pour les deux jeux de données

5. Après un nombre suffisant de tirage, la significativité peut être assurée si on a plus de valeurs positives que négatives.

Nous avons calculé pour les centrales des jeux de données d_1 et d_2 les corrélations intra-classes avec 1000 ré-échantillonnages. La figure 2.17 présente les valeurs de corrélations spatiales inter-classes obtenues pour les deux jeux de données. Les valeurs de corrélation obtenues décroissent avec la distance. De plus, ces valeurs sont supérieures à 0.65 dans le cas de d_2 où les centrales sont plus rapprochées et valent en moyenne 0.5 pour le jeu de données d_1. Ce résultat traduit la présence significative d'un lien spatial entre les productions des centrales. Dans toute la suite, nous utilisons les séries stationnarisées.

2.4.2 Calcul des corrélations spatio-temporelles

La présence de corrélation spatiale entre les mesures de production a été démontrée par l'analyse spatiale du phénomène de production. Pour compléter cette analyse, nous étudions les corrélations temporelles entre les séries de production des différentes installations. La figure 2.18 représente les fonctions de répartition empiriques des valeurs de corrélations croisées (cross-correlation) entre les séries temporellement retardées de production du jeu de données d_1. Pour un couple de centrales PV s_1, s_2 par exemple, la valeur de corrélation est obtenue en calculant la corrélation entre s_1 et les différentes séries affectées d'un retard $lag(s_2, k)$. Le retard maximal utilisé est l'horizon limite de prévision soit 6 heures dans ce cas. La même opération est répétée en intervertissant s_1 et s_2. La valeur retenue est la valeur maximale obtenue. Les fonctions de répartition sont représentées pour trois classes de distances de 0 à plus de 100 km avec des tailles de classes de 50 km. La première classe (moins de 50 km) présente les valeurs de corrélation les plus élevées. Sur l'ensemble du graphique, on observe que les valeurs de corrélation croisées sont plutôt élevées, concentrées dans l'intervalle $[0.4 - 0.8]$ et décroissantes avec la distance. L'effet de la course du soleil n'est plus présent dans les séries stationnarisées. Les valeurs de corrélations croisées observées traduisent donc la dépendance spatio-temporelle entre les séries de production. Ce transfert d'information essentiellement dû aux mouvements des nuages permet d'anticiper les perturbations météorologiques et d'améliorer la qualité des prévisions uniquement en utilisant les données de production.

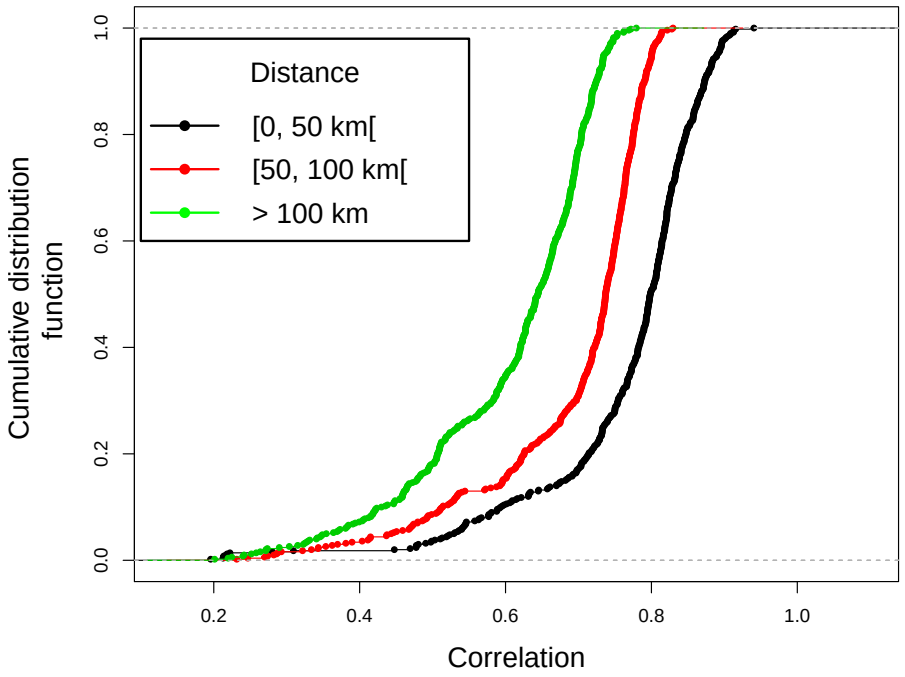


FIGURE 2.18 – Jeu de données d_2 : Fonction de répartition empirique des corrélations croisées entre les séries retardées de production. Les courbes vertes, rouges et noires correspondent respectivement aux trois classes de distance entre les centrales.

La Figure 2.19 qui présente le variogramme spatio-temporel pour les séries de production stationnarisées du jeu de données d_2 permet de compléter cette analyse. En effet, la portée est assez faible (de l'ordre de 10 km) et on n'observe pas d'effet pépite. De plus la forte montée des variogrammes traduit une forte variabilité spatiale entre les mesures de productions à courte distance.

Horizons et résolutions spatiales limites de la propagation des corrélations spatio-temporelles

La corrélation spatio-temporelle étant établie, nous allons à présent évaluer les limites temporelles et spatiales de la propagation de ces corrélations. Pour les corrélations temporelles, nous calculons le délai temporel t_{lim} à partir duquel la corrélation entre les séries de production retardées de chaque couple de centrales $(i, j), 1 \leq i, j \leq N$ est en valeur absolue inférieure ou égale à un seuil limite. Le tableau 2.3 présente les valeurs de t_{lim} pour un seuil de corrélation limite de 0.2 pour les centrales du jeu de données d_1. La valeur maximale est proche de 7 heures. La centrale la plus éloignée (P_3) présente les plus faibles valeurs de délai temporel car la propagation de corrélation est quasiment inexistante au vu de la distance. La figure 2.20 représente les valeurs de ce temps limite pour les centrales du jeu de données d_2. Le retard temporel limite dépasse les 7 heures pour ce cas d'étude. On peut conclure que la corrélation spatio-temporelle existant entre les centrales voisines est significative (> 0.20) pour des retards pouvant aller jusqu'à 7 heures. Au-delà de ces valeurs de délais temporels, il n'y a plus d'informations à exploiter de ces dépendances pour améliorer les prévisions. Ce résultat conforte les travaux de Bacher [26] sur les horizons limites d'utilisation des données historiques et justifie par la même occasion l'intérêt de la modélisation spatio-temporelle pour les horizons allant jusqu'à 6 heures.

Du point de vue spatial, le croisement des valeurs de corrélations entre les séries et les

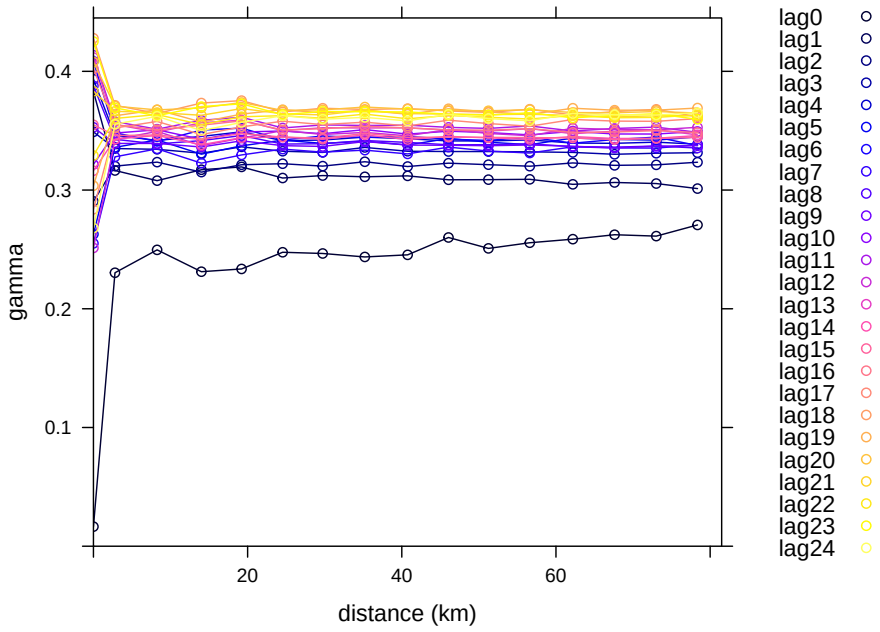


FIGURE 2.19 – Variogramme spatio-temporel des séries de production post stationnarisation

Tableau 2.3 – Jeu de données d_1 : Valeurs en heures de délai temporel à partir duquel la corrélation inter production est négligeable (seuil=0.2)

t_{lim} en heures	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
P_2	1.75							
P_3	0.75	0.25						
P_4	3.00	0.25	0.50					
P_5	5.25	4.75	0.50	0.75				
P_6	5.00	1.25	0.75	3.00	5.25			
P_7	5.25	4.00	0.50	1.75	6.75	5.25		
P_8	4.75	3.25	0.25	1.25	4.25	4.50	4.25	
P_9	4.75	4.75	0.25	1.25	5.25	4.75	5.25	4.75

distances entres sites montre qu'au-delà de 250 km les valeurs de corrélations sont très faibles. Cette distance constitue la limite au-delà de laquelle il n'y plus d'informations spatio-temporelles à exploiter.

2.5 Conclusion

La forte variabilité qui caractérise la production PV et ses différentes sources ont été présentées. La première source de variabilité est la saisonnalité. Nous avons montré dans ce chapitre la relation entre la saisonnalité et la production PV. Cette saisonnalité se traduit par une production négligeable la nuit, des cycles journaliers liés à la position du soleil et des cycles annuels dus à l'alternance des saisons. La variabilité de la production qui est due à la course du soleil a fait l'objet de nombreuses études et peut être presque complètement modélisée. Dans le cadre de cette thèse, nous nous intéressons à la modéli-

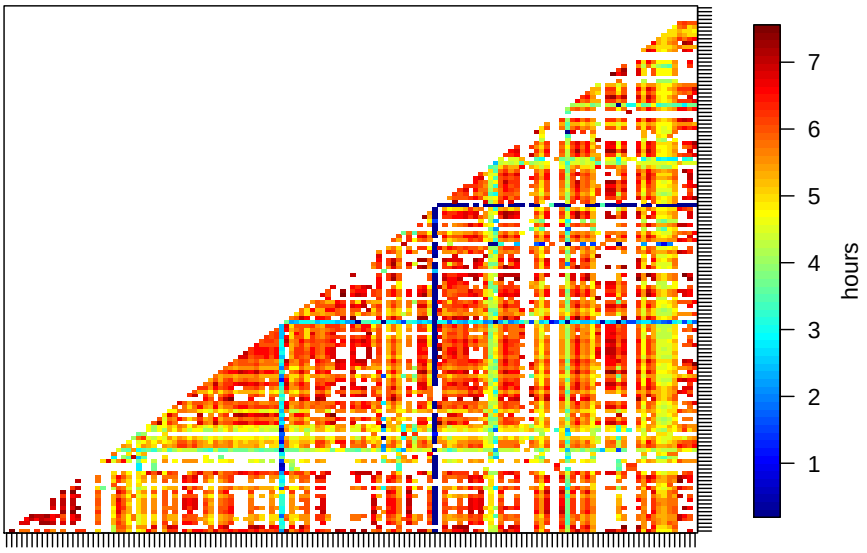


FIGURE 2.20 – Jeu de données d_2 : Valeurs en heures de délai temporel à partir duquel la corrélation inter production est négligeable (seuil=0.2). Chaque point d'axe représente une centrale PV.

sation spatio-temporelle de la production PV. Pour mettre en évidence cette corrélation et justifier l'intérêt de notre modélisation, il était nécessaire de s'affranchir des effets de la course du soleil sur les niveaux de production. En effet, un travail direct sur des séries non traitées aurait simplement été une mise en évidence de transferts de corrélation qui ne sont que la conséquence de la course du soleil qui serait à une position donnée sur certaines centrales avant d'autres. La méthode de stationnarisation que nous avons proposée dans ce chapitre nous permet d'exclure des séries de production l'effet de la course du soleil afin de nous consacrer aux autres sources de variabilité. Cette méthode de stationnarisation a montré de bonnes propriétés pour les tests de stationnarité mais a surtout donné de meilleures performances pour la prévision en comparaison avec les séries brutes. Les séries stationnarisées ont ensuite été utilisées pour mettre en évidence la corrélation spatiale et temporelle qui existe entre les centrales. Les valeurs de corrélation calculées sont significatives pour des horizons pouvant aller jusqu'à 6 heures. La mise en œuvre d'un modèle qui exploiterait ces informations serait donc prometteuse pour la réduction des erreurs de prévision. L'analyse de ces corrélations spatiales et temporelles a aussi mis en évidence les limites spatiales de la propagation de ces informations ; au-delà de 250 km de distance entre centrales, les corrélations entre séries de production sont très faibles. Dans la suite, nous allons proposer un modèle de prévision de la production PV qui exploite ces corrélations spatio-temporelles pour une meilleure prédictibilité de la production. La centrale P_3 qui est la plus distante des autres centrales du jeu de données d_1 ne sera pas utilisée dans toute la suite.

Chapitre 3

Modèle spatio-temporel déterministe pour la prévision photovoltaïque

3.1 Introduction

Ce chapitre est consacré à la prévision spatio-temporelle déterministe de la production PV. Les horizons de prévision envisagés vont de quelques minutes à plusieurs heures (6 heures). Les corrélations spatio-temporelles mises en évidence dans le chapitre précédent sont ici exploitées par le modèle de prévision déterministe proposé afin de réduire les erreurs de prévision par rapports aux méthodes "classiques" de la littérature. Ces méthodes "classiques" exploitent pour la plupart soit les dépendances temporelles soit celles spatiales. De plus pour la plupart des méthodes qui sont qualifiées de spatio-temporelles, la spatialité est souvent liée à la résolution spatiale des prévisions météorologiques utilisées.

Dans ce chapitre, nous présentons une approche qui exploite le lien spatial et temporel entre les données de production PV provenant d'installations photovoltaïques dispersées géographiquement pour prédire la puissance de sortie d'une centrale spécifique. Les prévisions météorologiques ne sont pas utilisées dans la première approche. Chaque centrale PV est utilisée comme un élément d'un ensemble de capteurs qui permettent de suivre les évolutions des perturbations météorologiques qui affectent la production. Cela différencie l'approche proposée de celles qui utilisent des données hors site provenant de stations météorologiques et de capteurs d'irradiation [95]. Les prévisions sont réalisées avec des données de production et non des données d'irradiation globale [92, 94] permettant ainsi de ne pas introduire un niveau d'incertitude supplémentaire lors du passage de l'irradiation à la production. Les performances du modèle proposé sont évaluées et comparées à d'autres méthodes de prévision. Une partie des résultats issus de ce travail a fait l'objet d'un article publié qui est présenté dans la section A de l'annexe.

Une seconde approche qui consiste à intégrer les prévisions météorologiques issues de modèles numériques au modèle de prévision spatio-temporel défini précédemment est aussi présentée dans ce chapitre. Cette intégration permet de tenir compte des phénomènes météorologiques pour des résolutions spatiales plus importantes. De plus avec les prévisions météorologiques, les horizons plus éloignés peuvent aussi être prédits et les performances pour les courts horizons stabilisés. Les performances de cette intégration de prévisions météorologiques sont évaluées.

La densité spatiale des centrales PV considérées dans les cas du monde réel peut être

variable. Pour cela, nous illustrons l'utilité de la méthodologie proposée avec les deux cas de test d_1 et d_2 qui présentent respectivement un nombre faible de centrales photovoltaïques réparties sur une grande zone géographique et un nombre élevé d'installations PV réparties sur une zone dense. Le problème de la dimensionnalité et la nécessité de l'introduction d'un processus de sélection de variables sont mis en évidence à travers le cas de test d_2. Nous présentons dans ce chapitre une méthode de sélection de variables pour pallier ce problème. Les données utilisées ici sont celles stationnarisées suivant la procédure présentée dans le chapitre précédent. Une transformation inverse est appliquée une fois la prévision faite pour retrouver des valeurs de production prédite. L'évaluation des erreurs est faite sur ces valeurs de production.

3.2 Les modèles de référence

3.2.1 Persistance et modèle autorégressif

Dans le but d'évaluer les avantages d'une approche spatio-temporelle pour la prévision PV, nous présentons des modèles de référence pour l'analyse comparative qui n'utilisent pas ces informations géographiquement distribuées. Plusieurs méthodes peuvent être utilisées pour prévoir la production PV telle que présenté dans l'état de l'art du chapitre 1. Le modèle de persistance est souvent utilisé comme référence dans la littérature pour comparer les performances des modèles avancés. En effet, il est simple à implémenter, et est basé uniquement sur les données mesurées et n'implique aucun processus de modélisation. Les résultats du modèle de persistance sont facilement reproductibles. De plus, dans les applications pratiques de la prévision PV, la persistance est souvent choisie comme un modèle de secours pour fournir des prévisions dans le cas où les modèles avancés échouent.

Nous définissons ici comme « persistance » un modèle qui considère que la production d'énergie d'une installation photovoltaïque au temps $t+h$ est la même que la production de cette centrale au même moment la veille. Cette approche ne tient pas compte des données hors site. Malgré sa popularité en tant que modèle de référence dans la littérature, sa performance globale est assez faible [2]. Pour tenir compte des différents facteurs qui affectent la production PV, on pourrait ajuster la persistance en fonction des valeurs observées durant le jour en cours. Toutefois, cela implique déjà une certaine manipulation de données, et différentes options peuvent être envisagées. Pour éviter d'obtenir des résultats trop optimistes à partir d'une méthode spatio-temporelle, il est également nécessaire d'utiliser un modèle de « référence avancé » qui présente des performances de pointe et une complexité raisonnable afin que les résultats puissent être facilement reproduits. Pour cela, nous considérons le modèle autorégressif (AR) défini comme suit :

$$\hat{P}^x_{t+h|t} = \hat{\beta}^0_h + \sum_{l=0}^{L} \hat{\beta}^l_h P^x_{t-l} \tag{3.1}$$

avec P^x_t la production de la centrale x au temps t et $\hat{P}^x_{t+h|t}$ la prévision de la production pour l'horizon h. L'ordre optimal L est très important pour la qualité de l'estimateur car il implique un compromis entre biais et variance. Cet ordre maximal est choisi ici par minimisation du critère de l'AIC (Akaike Information Criterion). En règle générale l'AIC se calcule par maximum de vraisemblance suivant la formule

$$AIC = -2log\tilde{L} + 2 * k$$

avec $\tilde{L}$ la vraisemblance maximisée et k le nombre de paramètres du modèle. Le modèle optimal est celui avec l'AIC le plus faible. Dans le cas du modèle AR, la vraisemblance peut

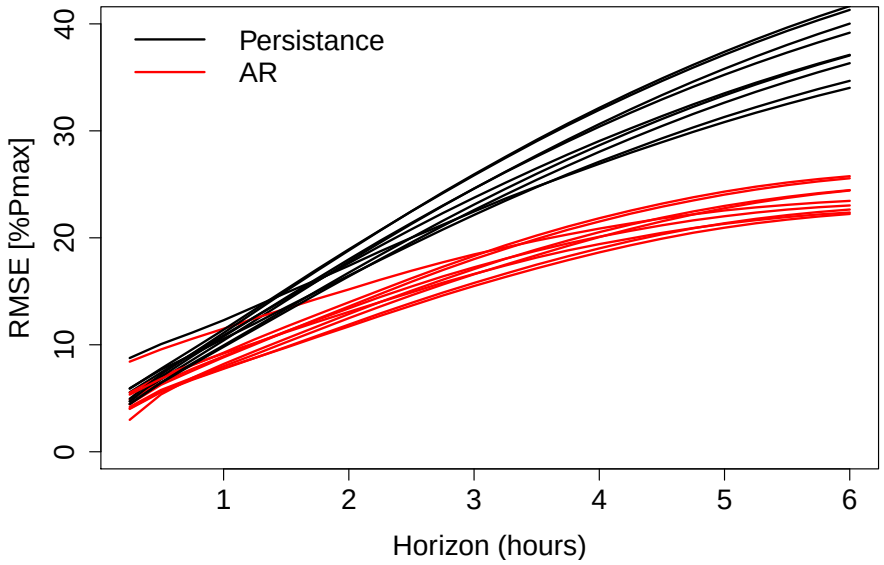


FIGURE 3.1 – Jeu de données d_1 : Comparaison des valeurs de RMSE normalisées des modèles AR (rouge) et persistance (noir). Le pas de temps est 15 min.

être remplacée lorsque n représente la taille des données par la variance des innovations σ_L^2 soit :

$$AIC = n(log\sigma_L^2 + 1) + 2 * (L + 1). \tag{3.2}$$

Les modèles de persistance et AR ont été appliqués au jeu de données d_1 avec des échantillons d'apprentissage et de test couvrant respectivement 15 mois et 5 mois. Les prévisions sont mises à jour au pas de temps de 15 minutes. La figure 3.1 présente l'erreur quadratique moyenne RMSE normalisée pour les modèles AR et persistance pour les centrales du jeu de données d_1 en fonction de l'horizon de prévision. La figure montre que le meilleur modèle est le modèle AR qui présente les niveaux de RMSE les plus bas. Les valeurs de biais et de MAE sont aussi plus basses pour le modèle AR que pour la persistance. Le modèle AR a donc été préféré au modèle de persistance. Ce modèle servira à évaluer la contribution de l'intégration d'informations supplémentaires relatives aux centrales avoisinantes.

3.2.2 Le modèle de forêts aléatoires

Le modèle AR n'est pas le seul modèle de référence que nous utilisons. Les forêts aléatoires sont aussi utilisées pour évaluer l'efficacité des modèles proposés. En effet, les forêts aléatoires font partie des méthodes déterministes les plus performantes que l'on retrouve dans la littérature pour prévoir la production PV [15]. Nous présentons ici le principe d'estimation de cette méthode et son application à la prévision. Dans cette partie la variable aléatoire Y représente la production PV, ses réalisations sont les mesures de production à chaque instant et X représente les variables explicatives.

Définition et propriétés

Une forêt aléatoire est l'agrégation d'une collection d'arbres aléatoires. Le nom forêt aléatoire vient du fait que les prédicteurs individuels sont des prédicteurs par arbre et de l'introduction de l'aléatoire dans le choix des variables de division et des échantillons "Out Of Bag". Les forêts aléatoires ont été développées par L. Breiman [106] et font partie

des méthodes d'ensemble. Le principe de ces méthodes est de construire une collection de prédicteurs et d'agréger ensuite l'ensemble de leurs prédictions. Dans le cadre d'une régression, si on dispose de q prédicteurs individuels qui fournissent chacun une prévision $\hat{Y}_t$, agréger leurs prédictions revient ici à faire une moyenne $\frac{1}{q}\sum_{t=1}^{q}\hat{Y}_t$.

Les forêts aléatoires sont caractérisées par :

— un nombre important de prédicteurs individuels (d'arbres) ;

— la création pour chaque arbre d'un échantillon destiné aux tests appelé échantillon « out-of-bag » ;

— le choix pour tous les arbres d'un paramètre correspondant à la taille du sous-échantillon de variables tiré aléatoirement à chaque nœud de chaque arbre ;

— une variable de division choisie pour chaque arbre parmi le sous-échantillon précédemment décrit ;

— l'absence d'élagage des arbres.

Prévision à l'aide de forêts aléatoires

Soit θ le vecteur de paramètres qui détermine la construction d'un arbre (les variables de division à chaque nœud par exemple), on notera $T(\theta)$ l'arbre correspondant. La prévision par les forêts aléatoires pour une nouvelle observation $X = x_0$ se fait suivant les étapes ci-après.

1. Construire K arbres $T(\theta_t), t = 1, \ldots, K$ comme décrit précédemment.
2. Faire passer x_0 dans l'arbre et conserver toutes les observations de la feuille terminale dans laquelle il tombe.
3. Calculer pour chaque arbre t des poids $w_i(x_0, \theta_t)$ tel que
$$w_i(x_0,\theta_t) = \begin{cases} \frac{1}{k_t} & \text{si} \quad \text{l'observation } x_i \text{ de l'échantillon d'apprentissage fait} \\ & \quad \text{partie des } k_t \text{ points du nœud terminal contenant } x_0 \\ 0 & \text{sinon} \end{cases}$$
4. La prévision pour un arbre t est alors
$$\hat{\mu}(x_0) = \sum_{i=1}^{n} w_i(x_0,\theta_t) Y_i.$$
5. On en déduit la prévision par les forêts aléatoires en moyennant sur tous les arbres :
$$\hat{\mu}(x_0) = \sum_{i=1}^{n} w_i(x_0) Y_i$$
avec $w_i(x_0) = K^{-1}\sum_{t=1}^{K} w_i(x_0,\theta_t)$.

On vient ainsi de construire une approximation de l'espérance conditionnelle $\mathbb{E}(Y|X)$ par une somme pondérée sur toutes les observations. Les pondérations, qui varient avec les covariables sont au centre de la démarche de prévision. Plusieurs études sont consacrées à la compréhension du lien entre ces pondérations et la distribution conditionnelle $(Y|X)$. Y. Lin & Y. Jeon [107] ont montré que les poids ont tendance à être d'autant plus importants que la distribution conditionnelle de $(Y|X = X_i)$ est similaire à celle de $(Y|X)$.

3.3 Les modèles proposés

3.3.1 Le modèle spatio-temporel

Le modèle spatio-temporel proposé ici permet de prévoir la production PV pour une centrale d'intérêt en utilisant les données des centrales voisines. Soit $\mathcal{X}$ l'ensemble des N centrales PV et Ls l'ordre temporel (lag) optimal. Le modèle est défini pour une centrale x par :

$$P_t^x = \beta^0 + \sum_{l=0}^{Ls} \sum_{y \in \mathcal{X}} \beta^{l,y} P_{t-l}^y \tag{3.3}$$

avec P^x la production de la centrale x. Pour un horizon h, les coefficients $\beta = (\beta^0, \beta_r)$ avec $\beta_r = (\beta^{l,y})_{0 \leq l \leq Ls,\, y \in \mathcal{X}}$ sont estimés en utilisant les moindres carrés. Cela revient à minimiser la somme carrée des résidus (RSS) :

$$RSS(\beta) = \|\mathbf{P^x} - \mathbf{X}\,\beta\|^2. \tag{3.4}$$

$\mathbf{X}$ est une matrice $N \times (Ls + 1)$ dont les lignes représentent les productions actuelles et retardées des centrales y_i

$$\mathbf{X} = \begin{pmatrix} 1 & P_t^{y_1} & \dots & P_{t-Ls}^{y_1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & P_t^{y_N} & \dots & P_{t-Ls}^{y_N} \end{pmatrix}. \tag{3.5}$$

La prévision au temps t pour l'horizon h de la production de la centrale x est donc définie par :

$$\hat{P}_{t+h|t}^x = \hat{\beta}_h^0 + \sum_{l=0}^{Ls} \sum_{y \in \mathcal{X}} \hat{\beta}_h^{l,y} P_{t-l}^y. \tag{3.6}$$

Dans le modèle ainsi défini, les séries de production de toutes les centrales voisines et les séries retardées respectives sont utilisées comme variables explicatives dans le modèle. L'estimation des coefficients de ce modèle reste très rapide même dans le cas d'un nombre important de centrales voisines parce que les coefficients sont linéaires. Le critère de l'AIC est utilisé pour réduire la complexité du modèle. Il représente ici un compromis entre le biais, diminuant avec le nombre de paramètres, et la parcimonie, volonté de décrire les données avec le plus petit nombre de paramètres possibles. La sélection est faite dans le sens descendant. Les séries de production et leurs décalages respectifs sont intégrés dans le modèle, puis supprimés un par un et l'AIC est recalculé à chaque fois. Le modèle avec le plus petit AIC est retenu. Dans la suite nous présentons une méthode plus optimale pour traiter la dimensionnalité.

3.3.2 Extension du modèle : Classification des situations météorologiques

Le principe.
Le modèle précédemment défini est uniquement basé sur les données historiques de la production. Les conditions météorologiques locales peuvent être une source d'information supplémentaire pour améliorer les prévisions avec un modèle spatio-temporel. En effet, ces variables décrivent les différents changements ou sources de variabilité qui peuvent affecter la production du site d'intérêt. Nous proposons de modifier le modèle spatio-temporel pour

prendre en compte ces conditions météorologiques locales. La méthode d'intégration que nous proposons n'est pas juste une addition d'une ou de plusieurs variables exogènes dans le modèle mais une intégration dans le processus d'estimation des coefficients. Avec les notations précédentes, le nouveau modèle fournit la prévision pour l'horizon h par :

$$\hat{P}^x_{t+h|t} = \hat{\beta}^0_h(Z) + \sum_{l=0}^{L} \sum_{y \in \mathcal{X}} \hat{\beta}^{l,y}_h(Z) P^y_{t-l}, \tag{3.7}$$

où Z est la variable représentant les données météorologiques. Les coefficients sont estimés par moindre carrés pondérés. L'estimateur des coefficients s'écrit :

$$\hat{\beta}^{l,y}_h(z) = Arg \min_{\beta} \sum_t \phi \left(\frac{Z_{t,y} - z}{\gamma} \right) \left(P^y_t - P^y_{t+h} \right)^2 \tag{3.8}$$

avec γ la moyenne de la variable aléatoire Z. Les poids choisis ici sont exponentiels :

$$\phi(x) = \exp\left(-||x||^2/2 \right). \tag{3.9}$$

Les coefficients sont estimés suivant les valeurs des variables météorologiques locales. Le modèle ainsi construit est nommé dans la suite, modèle spatio-temporel conditionné par rapport aux variables météorologiques.

Construction des classes de variables

La construction du modèle avec dépendance des coefficients aux conditions météorologiques locales nécessite la création de classes de la (ou les) variable(s) météorologique(s) retenue(s). Pour ce faire, la variable météorologique est décomposée en classes selon les valeurs prises et les coefficients sont estimés sur chaque classe (ou intervalle) de valeurs lorsqu'elle est quantitative. Dans le cas de variables catégorielles, les classes sont créées par regroupement des facteurs de la variable. Une analyse graphique de la relation entre les variables météorologiques locales et la production permet de définir des classes de valeurs pour lesquelles le comportement de la production est similaire. La construction des classes est plus simple dans le cas de variables catégorielles avec un faible nombre de facteurs. Ce modèle conditionné par la météorologie locale permet donc d'exploiter non seulement l'information spatio-temporelle portée par les centrales voisines mais aussi les conditions météorologiques locales du site pour lequel la prévision est faite.

3.3.3 Problème de dimensionnalité et de parcimonie : la sélection de variables

Les modèles spatio-temporels conditionnés ou non sont caractérisés par un nombre important de variables. Ces modèles posent donc un problème de dimensionnalité mais aussi de parcimonie (volonté de décrire les données avec le plus petit nombre de paramètres possibles). La solution proposée dans la partie 3.3.1 est basée sur l'AIC avec une sélection descendante. Cette solution impose une modélisation en plusieurs étapes à savoir la construction du modèle complet, la mise en œuvre de la sélection de variables AIC et après l'estimation du modèle optimal retenu. De plus la phase de sélection de variables est longue. Nous proposons ici une méthode pour résoudre à la fois le problème de dimensionnalité et de parcimonie qui soit automatique, intégrée au processus de prévision et avec de bonnes performances de prévision. Cette technique est basée sur la régression pénalisée Lasso (Least Absolute Selection and Shrinkage Operator) [108]. Cette régression

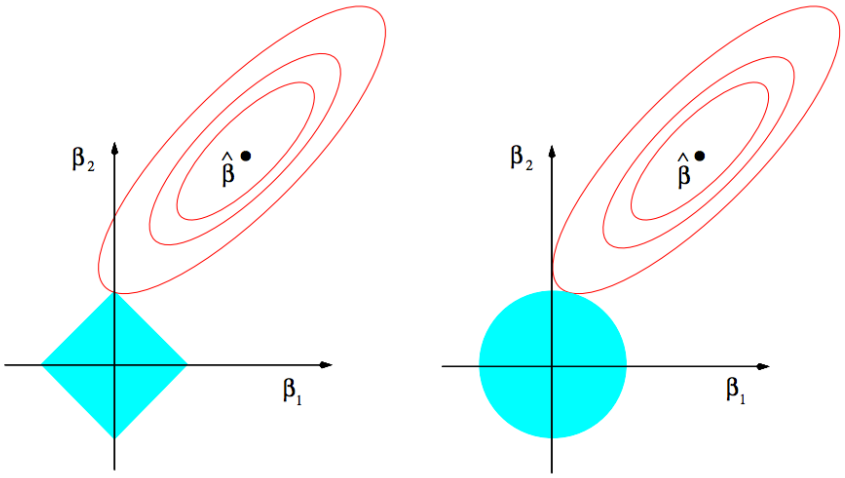


FIGURE 3.2 – Principe d'estimation du Lasso (gauche) et de la régression ridge (droite). Les zones en bleu sont les régions de contraintes $|\beta_1| + |\beta_2| \leq t$ et $|\beta_1^2| + |\beta_2^2| \leq t^2$. Les ellipses rouges représentent les contours de l'erreur des moindres carrés.

intègre une pénalité dans le problème de minimisation en appliquant une contrainte sur la somme des valeurs absolues des coefficients. L'estimateur est défini comme suit :

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} RSS(\beta) + \lambda \|\beta\|_1 \right\}. \tag{3.10}$$

Ce problème est équivalent à minimiser $RSS(\beta)$ sous une contrainte de la forme $\|\beta\|_1 \leq \hat{R}_\lambda$. Un certain biais est introduit dans l'estimateur, mais la variance est réduite. Pour des valeurs suffisamment élevées du paramètre de pénalisation λ, certains paramètres β sont annulés. L'estimateur Lasso est donc un estimateur parcimonieux. Le paramètre de régularisation λ est obtenu par validation croisée et le chemin des solutions de β est linéaire par morceaux, ce qui implique la possibilité d'ordonner les λ en suite de valeurs croissantes. La figure 3.2 représente le principe d'estimation du Lasso et du ridge (pénalisation L_2).

3.4 Evaluation des modèles

Les modèles proposés sont appliqués aux jeux de données d_1 et d_2 pour des horizons de 15 min à 6 heures avec un pas de temps de 15 minutes et avec une mise à jour des prévisions toutes les 15 minutes. Les échantillons d'apprentissage et de test représentent respectivement deux tiers et un tiers de la taille des données disponibles. Les prévisions sont comparées à celles du modèle de référence. Les modèles ont été développés à l'aide du logiciel R [109]. Les librairies utilisées sont *randomForest, glmnet*. Le temps de calcul pour le modèle spatio-temporel pour une centrale est d'environ 20 secondes par horizon sans conditionnement et 45 secondes avec sélection de variables Lasso.

Les critères d'évaluation

La première étape avant le calcul des prévisions de production PV est la définition des outils qui permettront d'évaluer la précision des prévisions fournies. Il existe plusieurs critères pour évaluer les prévisions de production PV qu'elles soient déterministes ou probabilistes [110, 111]. Nous présentons ici les plus courants qui sont utilisés pour l'évaluation des prévisions déterministes. Soit $\hat{P}_{t+h|t}$, la prévision à l'instant t pour l'horizon h. On

note $e_{t+h|t} = P_{t+h|t} - \hat{P}_{t+h|t}$, l'erreur de prévision à l'horizon h. Les critères d'évaluation de prévision déterministes les plus courants sont définis comme suit :

— l'erreur quadratique moyenne RMSE (Root Mean Square Error)

$$RMSE_h = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_{t+h|t}^2}$$

— l'erreur moyenne absolue MAE (Mean Absolute Error)

$$MAE_h = \frac{1}{n}|e_{t+h|t}|$$

— le biais BIAS

$$BIAS_h = \frac{1}{n}\sum_{t=1}^{n} e_{t+h|t}.$$

Ces critères sont utilisés par la suite pour évaluer les performances des modèles de prévision. L'objectif étant de minimiser l'ensemble des critères pour s'assurer d'avoir des prévisions de qualité.

3.4.1 Performances du modèle spatio-temporel

Analyse des erreurs

Les erreurs de prévision du modèle spatio-temporel défini dans la partie précédente selon l'équation (3.6) sont calculées. Les centrales des deux jeux de données sont utilisées. L'horizon de prévision est 6 heures avec un pas de temps de 15 min. Les densités des erreurs de prévision sont calculées à l'aide de l'estimation de la densité du noyau et sont présentées dans les figures 3.3 et 3.4 pour les centrales PV du jeu de données d_1 à différents horizons. La centrale P_3 n'est pas représentée car c'est la plus distante des autres centrales (tableau 2.1) et donc la corrélation spatio-temporelle est inexistante.

Pour les deux centrales PV, les distributions ne sont pas gaussiennes, car les modes et les moyennes sont significativement différents. Les moyennes sont proches de zéro et l'inclinaison est négative. Au fur et à mesure que l'horizon augmente, le mode de la distribution se déplace vers la gauche. La même analyse a été effectuée sur les autres centrales PV du jeu de données d_1 avec les mêmes conclusions.

Comparaison aux modèles de référence

Pour obtenir un aperçu plus complet de la performance du modèle proposé, nous le comparons aux modèles de forêts aléatoires (RF). Les modèles RF sont présentés dans la littérature [15] comme l'un des modèles les plus efficaces pour produire des prévisions précises de la production d'énergie photovoltaïque. Nous avons donc calculé les prévisions de production avec un modèle RF et comparé ses performances au modèle spatio-temporel. Les variables explicatives utilisées dans le modèle de forêts aléatoires sont les variables de temporalité et les séries de production retardées. Le tableau 3.1 présente les améliorations minimale, moyenne et maximale de RMSE sur les horizons de 6 heures du modèle spatio-temporel (ST) par rapport aux modèles AR et RF pour un échantillon de cinq centrales PV du jeu de données d_1. Le tableau montre une amélioration moyenne d'environ 10% pour le modèle ST par rapport au modèle AR et 6% par rapport au RF. L'amélioration par rapport aux modèles AR et RF peut atteindre respectivement 20% et 15%. Les valeurs d'amélioration sont assez similaires pour toutes les centrales PV.

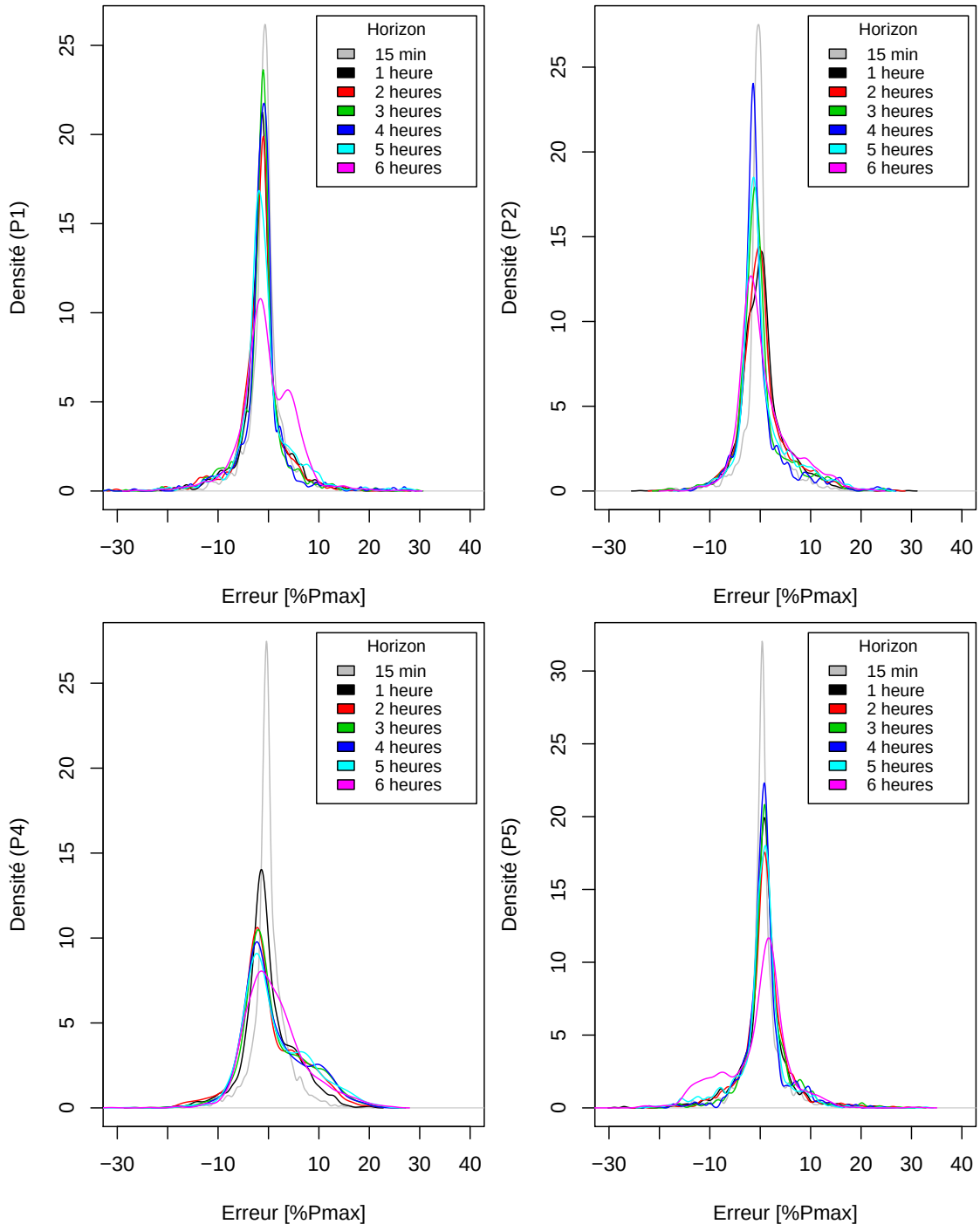


FIGURE 3.3 – Densités de probabilité pour différents horizons des erreurs de prévision pour le modèle spatio-temporel appliqué aux centrales PV ($P_1 - P_5$) du jeu de données d_1. La centrale P_3 qui est la plus éloignée n'est pas représentée.

Les performances du modèle de référence et du modèle spatio-temporel pour les critères de l'erreur absolue (MAE) et du biais sont présentés dans les figures 3.5 et 3.6 pour les centrales du jeu de données d_1. On remarque que comme pour le RMSE, le modèle spatio-temporel permet de réduire les erreurs absolues et le biais améliorant ainsi la qualité des prévisions fournies.

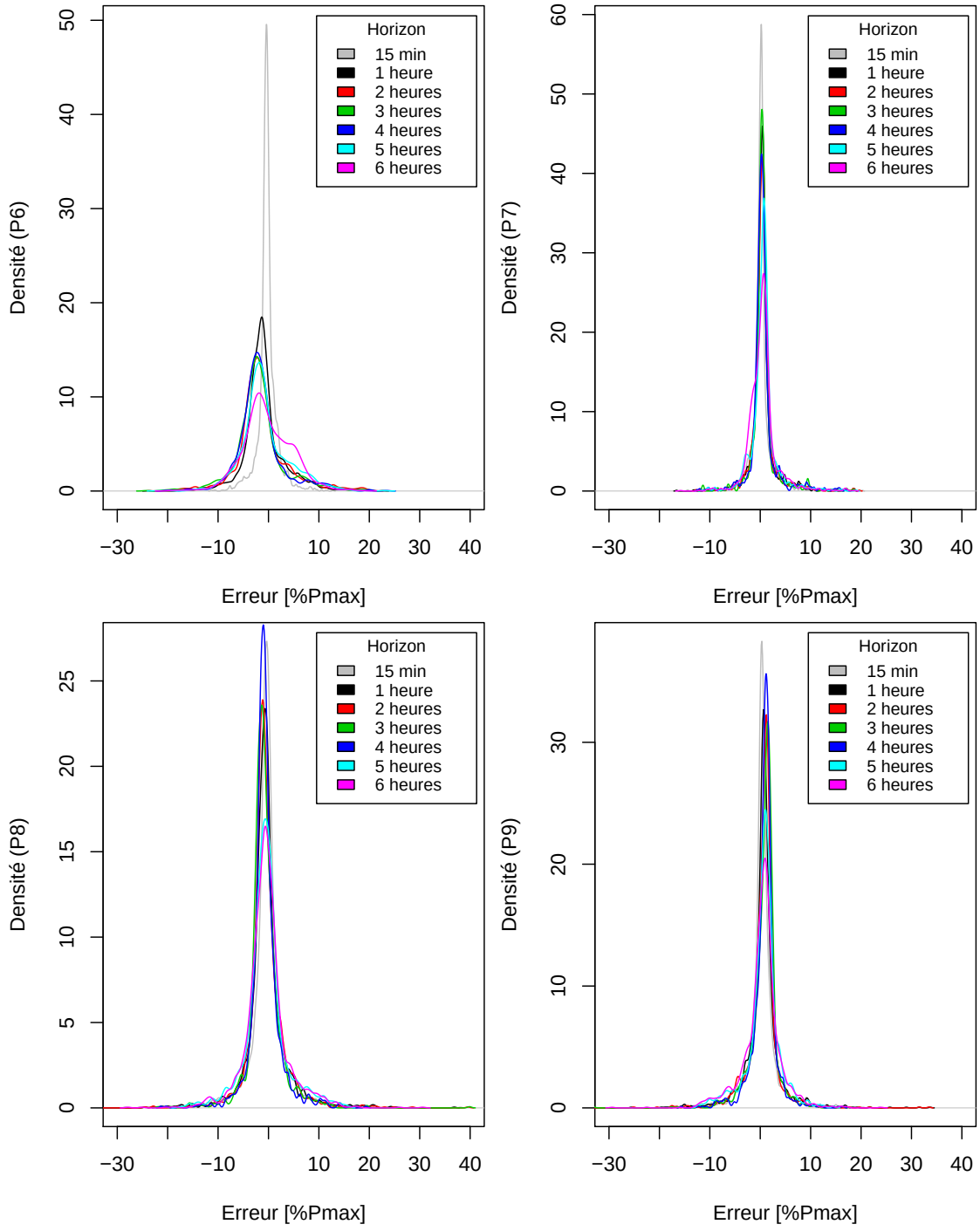


FIGURE 3.4 – Densités de probabilité pour différents horizons des erreurs de prévision pour le modèle spatio-temporel appliqué aux centrales PV ($P_6 - P_9$) du jeu de données d_1.

3.4.2 Performances des prévisions selon le niveau de couverture nuageuse

L'analyse de la performance du modèle spatio-temporel peut être faite en fonction du niveau de couverture du ciel. Les jours de l'ensemble de test peuvent donc être regroupés

Tableau 3.1 – Amélioration du RMSE du modèle spatio-temporel par rapport au modèle de référence AR et au modèle RF pour 5 centrales du jeu de données d_1.

Amélioration RMSE (%)		P_1	P_2	P_4	P_5	P_6
ST vs AR	min	0.4	3.02	0.61	-0.46	0.83
	moy	9.49	13.05	7.36	8.69	12.57
	max	16.81	19.27	12.5	15.71	20.13
ST vs RF	min	0.17	2.94	0.32	-0.72	2.14
	moy	6.52	10.27	4.5	5.03	7.84
	max	15.3	16.6	9.03	11.12	11.39

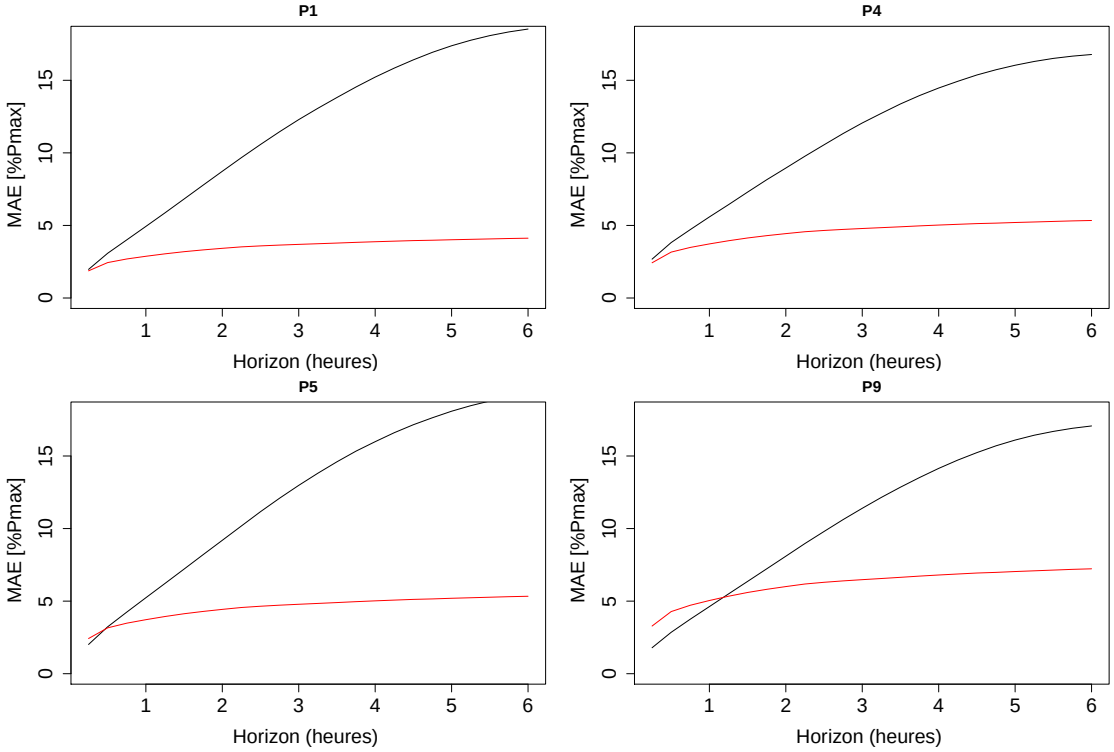


FIGURE 3.5 – Comparaison des MAE. Un graphique par centrale. La courbe en noir représente la performance du modèle de référence AR et celle en rouge celle du modèle spatio-temporel.

selon le niveau de couverture nuageuse. Nous définissons ainsi trois niveaux de couverture nuageuse : ciel clair (cs), moyennement nuageux (mc) et très nuageux (vc). Ces niveaux ont été calculés en utilisant un indice basé sur le rapport de la somme de la production journalière à la somme de l'irradiation simulée en utilisant le modèle ESRA. La figure 3.7 présente pour deux centrales PV de d_1, l'amélioration en termes de RMSE du modèle spatio-temporel par rapport au modèle de référence par type de jour.

Nous observons que pendant les deux premières heures, l'amélioration des jours nuageux dépasse celle des journées claires. Cette observation montre que le modèle spatio-temporel aide à capturer le mouvement des nuages. En effet, l'utilisation des centrales voisines permet d'anticiper les phénomènes qui impactent la production PV notamment le passage des nuages. Les graphiques montrent également que l'amélioration est plus

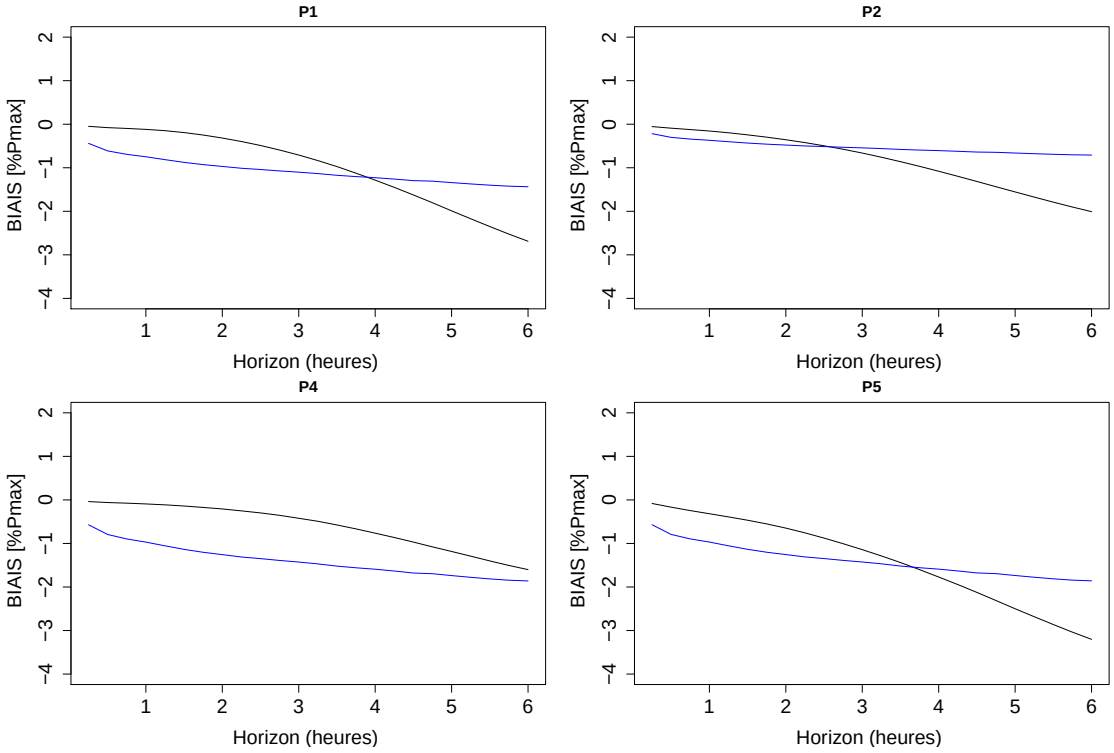


FIGURE 3.6 – Comparaison des BIAIS. Un graphique par centrale. La courbe en noir représente la performance du modèle de référence AR et celle en bleu celle du modèle spatio-temporel.

grande pour les jours de ciel clair pour les horizons plus longs et dépasse les 5%. L'analyse des résultats pour les autres centrales électriques a donné des conclusions similaires. La prévision spatio-temporelle permet d'améliorer les performances de prévision.

3.4.3 Effet du conditionnement par des variables météorologiques

Nous présentons ici les performances du modèle avec conditionnement (partie 3.3.2) par les variables météorologiques observées localement. Les variables retenues sont la vitesse du vent, la température et l'humidité relative. Ces variables sont connues pour leur influence sur la production PV. Des classes de ces variables sont donc constituées pour une introduction dans le processus d'estimation des coefficients du modèle spatio-temporel.

La vitesse du vent affecte le mouvement des nuages et donc la production PV. Toutefois nous ne considérons ici que la vitesse du vent de surface, qui est en général considérablement différente de la vitesse du vent dans les couches supérieures de l'atmosphère, le but n'est pas d'établir une relation explicite avec le mouvement des nuages mais d'évaluer l'impact de l'utilisation des mesures de vitesse de vent locales sur les performances du modèle spatio-temporel. Le modèle spatio-temporel conditionné par rapport à la vitesse du vent a été appliqué au jeu de données d_1. Le conditionnement se fait en autorisant une dépendance des coefficients du modèle spatio-temporel aux valeurs locales de vitesse de vent. La figure 3.8 présente l'amélioration des performances du modèle conditionné avec la vitesse du vent par rapport au modèle statistique pur. On observe une réduction du RMSE pour les deux premières heures de prévision. La valeur moyenne de cette amélioration est de 2% et la réduction la plus significative est notée pour la première heure de

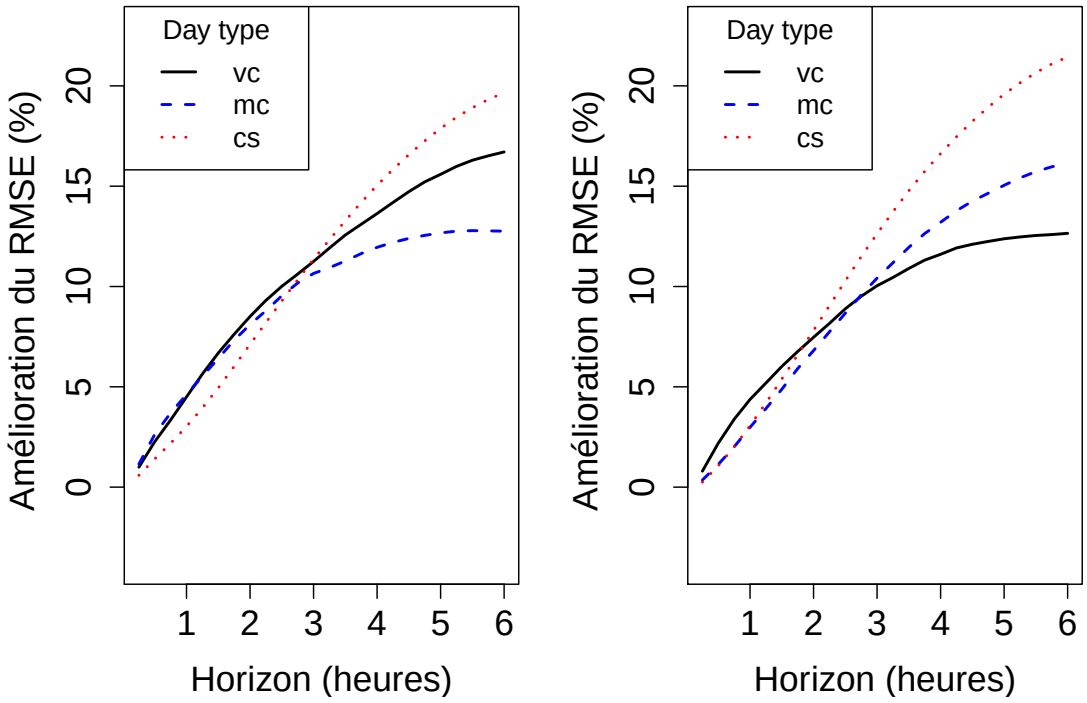


FIGURE 3.7 – Amélioration du RMSE du modèle spatio-temporel par rapport au modèle AR selon le type de jour pour deux centrales du jeu de données d_1
Les types de jours sont ciel clair (cs), moyennement nuageux (mc) et très nuageux (vc).

prévision. Au-delà de deux heures l'apport du modèle conditionné par la vitesse du vent ne montre aucune amélioration par rapport au modèle statistique pur.

Les mesures de température et d'humidité ont aussi été utilisées comme variables de conditionnement du modèle spatio-temporel. Il a été montré que les taux d'humidité importants et les températures trop élevées affectent les performances des cellules des panneaux PV [112, 113]. Il est donc intéressant de voir si l'utilisation de mesures locales de ces variables météorologiques permet d'améliorer les performances de prévision du modèle spatio-temporel proposé. La figure 3.9 présentent les améliorations en termes de RMSE en fonction des horizons du modèle conditionné par rapport au modèle non conditionné pour ces deux variables météorologiques. On remarque que l'apport de la température est très faible (moins de 1%) peu importe l'horizon de prévision. Quant à l'humidité relative, l'amélioration observée est de l'ordre de 2% en moyenne.

3.4.4 Les performances de la sélection de variables

Dans cette section, l'impact des différentes méthodes de sélection de variables est évalué. Nous considérons ici le jeu de données d_2 car il présente un nombre élevé de centrales électriques et pose donc le problème de la dimensionnalité. Le modèle spatio-temporel avec la procédure de sélection de variables basée sur l'AIC comme décrit dans la partie 3.3.1 a été évalué. L'extension du modèle avec une procédure de sélection de variables basée sur la régularisation Lasso (partie 3.3.3) a également été implémentée et évaluée sur le même ensemble de données d_2. La figure 3.10 représente la dispersion de la valeur moyenne sur tous les horizons de prévision du RMSE pour le modèle de référence et le modèle spatio-temporel résultant des deux procédures de sélection variable.

La figure montre que le modèle spatio-temporel réduit considérablement les erreurs

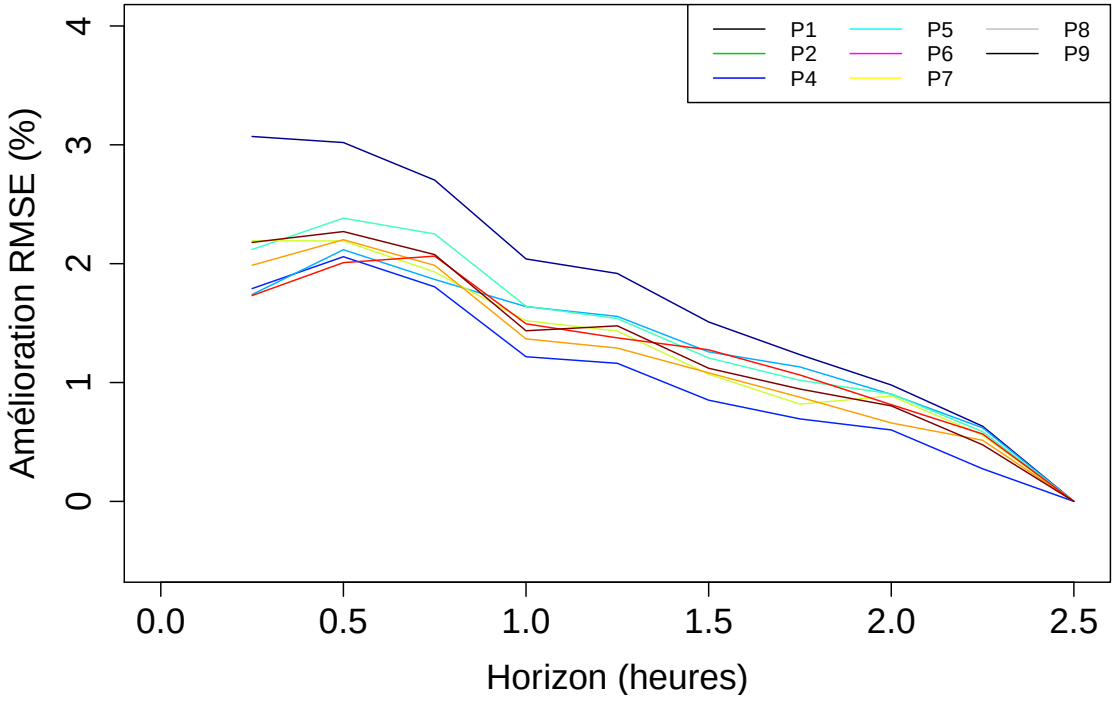


FIGURE 3.8 – Amélioration du RMSE entre le modèle spatio-temporel conditionné par la vitesse du vent et le modèle sans conditionnement pour le jeu de données d_1. Une ligne représente une centrale.

de prévision par rapport au modèle de référence (environ 28% en moyenne de réduction du critère d'erreur). En outre, la procédure de sélection de variables par le Lasso présente des erreurs de prévision inférieures à celles obtenues en fonction de l'AIC, montrant que la procédure Lasso est plus efficace (22% en moyenne de réduction du critère d'erreur).

La performance de la procédure de sélection Lasso peut également être analysée par le niveau de réduction du problème de dimensionnalité. Pour chacune des centrales PV de l'ensemble de données d_2, la figure 3.11 représente le nombre de centrales PV voisines (parmi les autres 135) retenues par la sélection Lasso. Dans 75% des cas, le nombre de centrales voisines sélectionnées est inférieur à 30, tandis que le nombre maximal utilisé est de 57. Ces chiffres montrent que la procédure de sélection de Lasso permet une bonne réduction de la dimension du problème. Les résultats en termes de performance de prévision soulignent l'intérêt de l'utilisation des centrales voisines dans le but d'améliorer la qualité des prévisions de production photovoltaïque.

3.5 Intégration des prévisions météorologiques

Le modèle spatio-temporel précédemment défini ne prend en compte que les données de production des différentes centrales et dans le cas du conditionnement, crée une dépendance entre les coefficients et les conditions météorologiques locales. Les fréquences de mise à jour importantes des prévisions météorologiques NWP, les durées de calcul importantes, la forte sensibilité aux conditions initiales de même que les limites en termes de résolution spatiales et temporelles sont autant d'éléments qui ont justifié la prise en compte des seules données de production dans le cadre de la proposition d'un modèle spatio-temporel pour de très courts horizons (< 6h). Il est toutefois intéressant d'évaluer

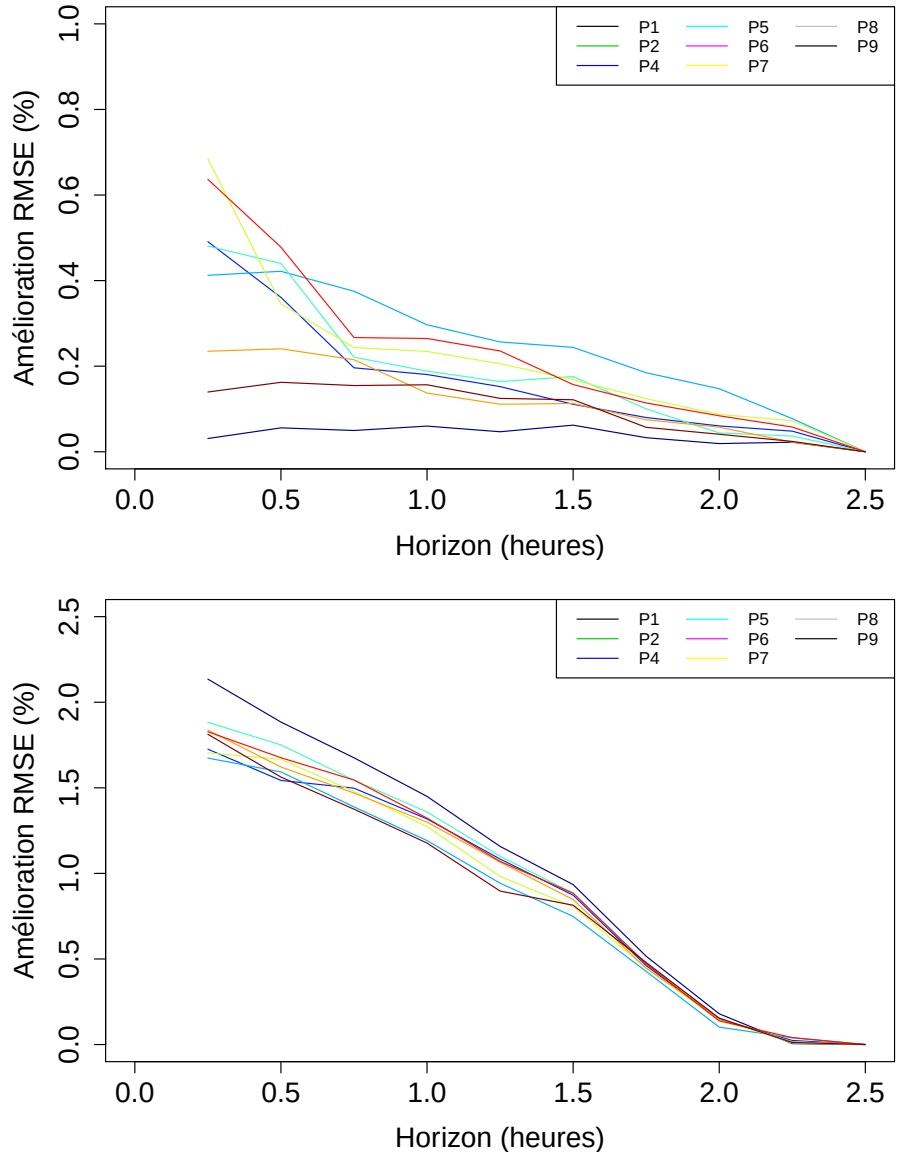


FIGURE 3.9 – Amélioration du RMSE entre les modèles spatio-temporels conditionnés par la température (haut) et l'humidité relative (bas) et le modèle sans conditionnement pour le jeu de données d_1. Une ligne représente une centrale.

l'apport que pourrait avoir l'utilisation de ces prévisions météorologiques sur les performances de prévision tout en restant dans une dynamique spatio-temporelle. De plus, cela permet de proposer un modèle qui prenne en compte la diversité des sources de données sans en exclure d'office une catégorie. Aussi, l'évaluation du modèle spatio-temporel avec données NWP intégrées permet de quantifier en fonction des horizons de prévision l'apport de ces données NWP. En effet dans la littérature, les horizons à partir duquel les prévisions numériques sont apporteuses d'informations varient entre 4 et 6 heures en fonction du type de modèle de référence. L'intégration de ces données pourrait donc répondre à la question de savoir à partir de quel horizon ces données apportent de l'information pour améliorer les performances dans le cas spécifique du modèle spatio-temporel. Nous présentons dans la suite les données de prévisions issues de modèles numériques utilisées et les variables retenues. La procédure d'intégration de ces données au modèle spatio-

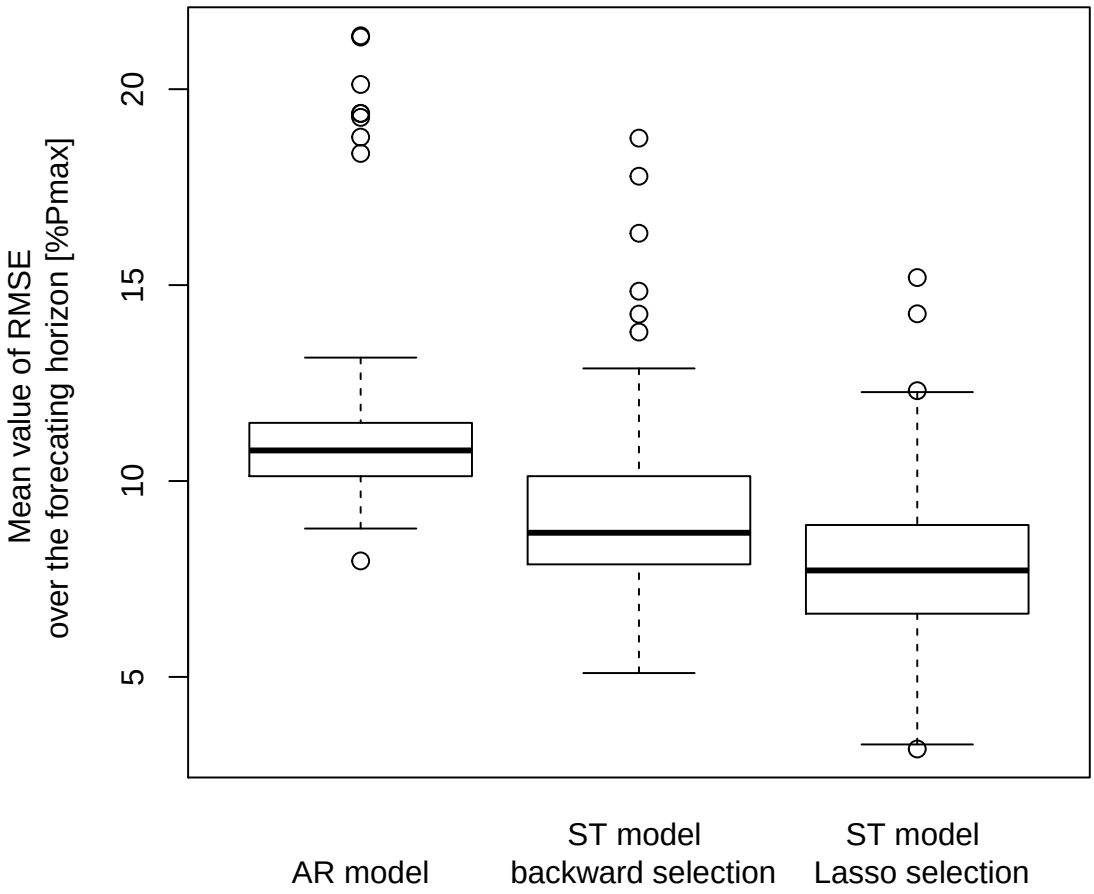


FIGURE 3.10 – Jeu de données d_2 : Répartition des valeurs moyennes (sur les 6 heures d'horizon) du RMSE pour les modèles de référence, spatio-temporel avec sélection par AIC et spatio-temporel avec la sélection Lasso.

temporel est ensuite détaillée et enfin, une analyse de l'apport de ces variables sur la qualité des prévisions en fonction de l'horizon est présentée.

3.5.1 Le modèle NWP utilisé et les variables retenues

Les données issues de modèles numériques de prévision de type NWP proviennent du modèle global de ECMWF (European Centre for Medium-Range Weather Forecasts). Ce centre fournit plusieurs produits de prévision. Nous utilisons un produit de prévision moyen-terme à forte résolution appelé HRES. La spécificité de HRES est que son état initial est la meilleure estimation des conditions météorologiques et qu'il utilise les meilleurs modèles physiques disponibles [1]. HRES fournit une description très détaillée des conditions climatiques futures et est le meilleur modèle de prévision sur 10 jours en comparaison à d'autres modèles. Ce modèle ne peut toutefois pas fournir une estimation de l'incertitude associée à la prévision [114].

Les résolutions du produit HRES [115] :

— une résolution de $0.1^\circ \times 0.1^\circ$ de latitude/longitude (ou tout multiple)

1. Plus de précisions sur la construction du modèle HRES et sa configuration peuvent être obtenues sur le site de ECMWF.

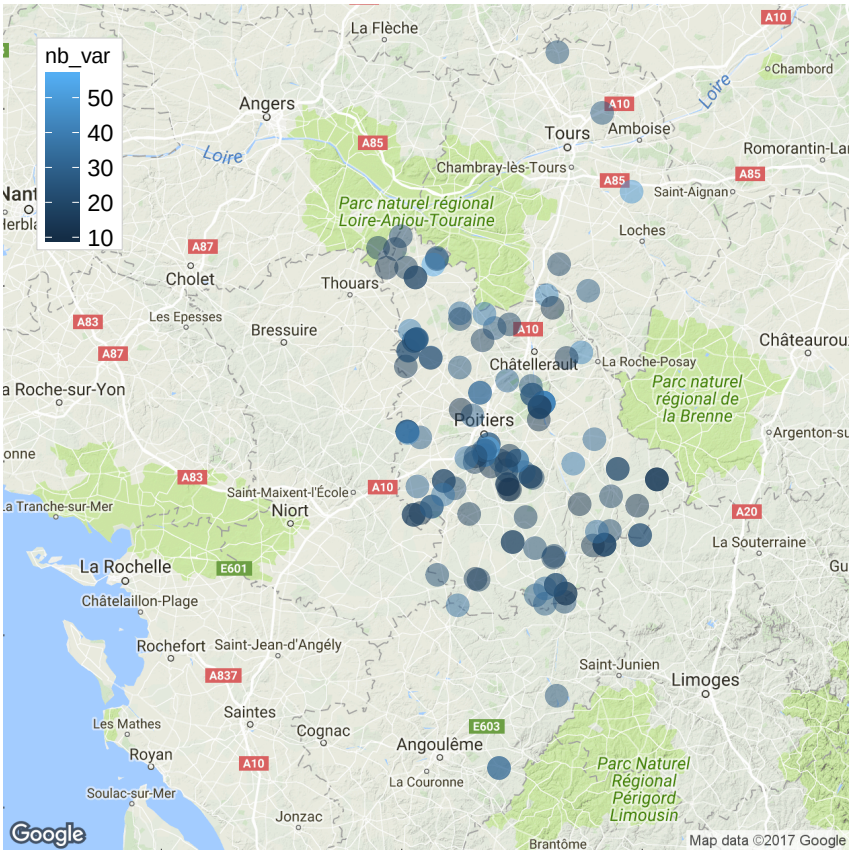


FIGURE 3.11 – Jeu de données d_2 : Carte des centrales avec le nombre de centrales voisines choisies par la sélection Lasso.

Tableau 3.2 – Description des variables météorologiques (NWP) retenues pour intégration dans le modèle spatio-temporel.

ID	Description
TCLW	Total Column Liquid Water
TCIM	Total Column Ice Water
SP	Surface Pressure
RH	Relative Humidity at 1000 mbar
10U/10V	10m U,V wind component
2T	2m Temperature
SSRD	Surface Solar Radiation Downward
STRD	Surface Thermal Radiation Downward
TSR	Top net Solar Radiation
TP	Total Precipiation

— un modèle de grille (octohédral) O1280 grid

— des composantes spectrales (TCO1279) pour les strates aériennes les plus hautes.

Ce modèle de prévision fournit un grand nombre de données en sortie. La résolution temporelle utilisée ici est 1 heure. Dans le cadre de la prévision spatio-temporelle déterministe de la production PV, les variables que nous avons retenues sont celles les plus couramment utilisées dans la littérature [53, 61]. Le tableau 3.2 présente une description de ces variables fournies par HRES.

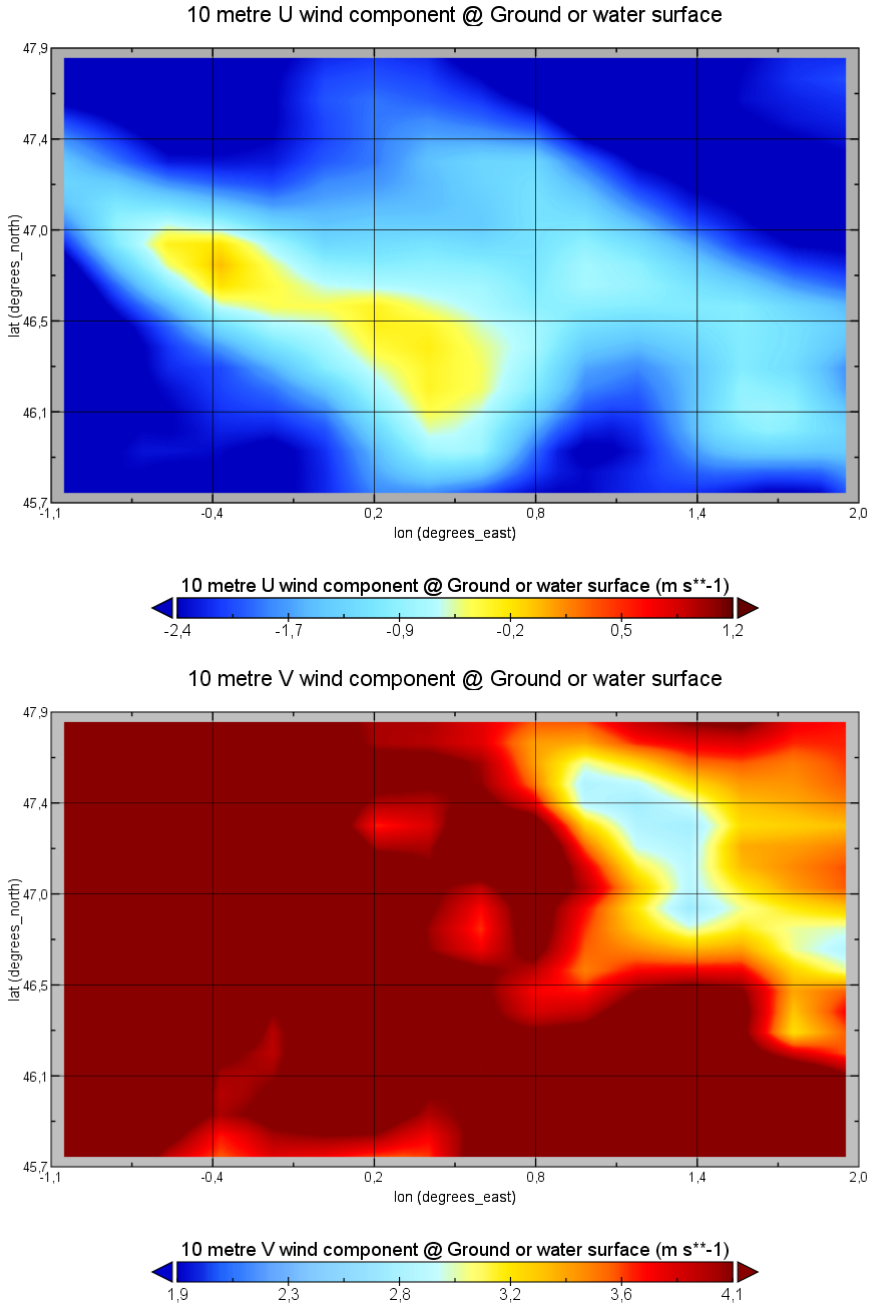


FIGURE 3.12 – Représentation 2D (lat/long) des composantes U (haut) et V (bas) de la vitesse du vent à 10m prévue par le modèle HRES de ECMWF. Run du 11/11/2014 à minuit pour la prévision du 11/11/2014 à 10h

La figure 3.12 présente des valeurs de prévision NWP pour les variables 10U/V sur la grille définie autour des centrales du jeu de données d_2. Les valeurs représentées sont celles du run du 11/11/2014 à minuit et l'instant prévu est le même jour à 10h. Il s'agit donc de proposer une méthode qui puisse extraire les informations de ces prévisions pour les intégrer au modèle spatio-temporel.

3.5.2 Méthode d'intégration des variables météorologiques au modèle spatio-temporel

Nous proposons la procédure suivante pour l'intégration des prévisions issues de modèles numériques dans le modèle spatio-temporel pour chaque centrale pour laquelle la prévision est faite :

1. choisir les points de grille les plus proches de la centrale ; en pratique nous avons choisi les 50 points de grille les plus proches
2. interpoler à une résolution horaire qui couvre les horizons de 1 heure à 12 heures
3. intégrer les séries de prévisions météorologiques NWP dans le modèle spatio-temporel comme variable exogène

$$P^x_{t+h|t} = \beta^0_h + \sum_{l=0}^{Ls} \sum_{y \in \mathcal{X}} \beta^{l,y}_h P^y_{t-l} + \boxed{\sum_{k=1}^{p} \gamma_k NWP^k_t} \tag{3.11}$$

4. intégrer la pénalisation Lasso au processus d'estimation des coefficients pour s'assurer de la parcimonie du modèle et sélectionner automatiquement les variables de production adéquates ; il faut donc résoudre le problème

$$\underset{\beta,\gamma}{\operatorname{argmin}} \left\{ \frac{1}{2} RSS(\beta, \gamma) + \lambda \|\beta\|_1 \right\}. \tag{3.12}$$

Le modèle ainsi défini est implémenté pour calculer les prévisions de production PV. L'horizon de prévision est 12 heures. Le modèle spatio-temporel avec prévisions météorologiques NWP est implémenté est nommé M_1 dans la suite. Le modèle qui n'inclut pas ces prévisions météorologiques est nommé M_2 dans la suite. Ces deux modèles sont implémentés et comparés pour les centrales du jeu de donnée d_2. Le tableau 3.3 présente les valeurs normalisées en fonction des horizons de prévision pour les critères du RMSE, du MAE et du BIAIS pour une centrale du jeu de données. Pour les 6 premières heures de prévision, les données NWP n'apportent pas d'amélioration significative en termes de performances de prévision. Par contre pour les horizons au-delà de 6 heures le modèle avec prévisions NWP présente de meilleures performances pour l'ensemble des critères présentés par rapport au modèle basé uniquement sur les mesures de production. L'analyse est la même pour les autres centrales du jeu de données pour lesquelles on observe aussi une forte augmentation des erreurs de prévision au-delà de l'horizon 6 heures et de meilleures performances avec le modèle intégrant les données NWP. Les différences de RMSE et MAE peuvent être de l'ordre de 10%.

3.6 Conclusion

Dans ce chapitre nous avons proposé un modèle basé sur les corrélations spatio-temporelles entre différents sites de production pour améliorer les prévisions à court-terme ($\leq$ 6 heures) de la production photovoltaïque. Le modèle spatio-temporel prend en compte les dépendances temporelles entre les mesures de production mais aussi les liens spatiaux. Le modèle de prévision que nous avons proposé ici se base sur des séries de production stationnarisées afin de se concentrer sur les informations non liées à la course du soleil. Ce modèle a été mis en œuvre et comparé à des modèles de prévisions qui n'exploitent pas ces informations spatio-temporelles. Les critères d'évaluation les plus courants dans la littérature ont été utilisés pour les comparaisons de performance. Pour l'ensemble des critères

Tableau 3.3 – Comparaison des performances des modèles avec ou sans prévisions NWP pour une centrale du jeu de données d_2 pour les critères de RMSE, MAE et BIAIS normalisés par la puissance maximale observée pour différents horizons.

Le modèle M_1 est celui purement statistique et M_2 celui qui intègre les données NWP.

Critères ($\%P_{max}$)	Horizons					
	15 min	1h	3h	6h	9h	12h
RMSE (M_1)	6.2	8.2	8.6	10.3	16.8	22.6
RMSE (M_2)	6.2	8.5	9	10.6	12.3	13.4
MAE (M_1)	7.1	8.3	10.2	13.5	18.7	21.5
MAE (M_2)	7.2	8	10.5	11.8	13.4	15
BIAIS (M_1)	0.01	0.01	0.01	-0.38	-0.64	-1.43
BIAIS (M_2)	0.01	0.01	0.2	0.3	0.12	0.23

retenus, le modèle spatio-temporel présente de meilleures performances que les modèles qui n'exploitent pas ces corrélations spatio-temporelles. L'utilisation de ces informations a donc permis de réduire les erreurs de prévision.

Des pistes d'amélioration du modèle spatio-temporel ont aussi été proposées afin d'en améliorer les performances. La première amélioration consiste à utiliser en complément des mesures de production sur site et des sites voisins, des mesures locales (proches du site d'intérêt) de variables météorologiques pertinentes dans le cadre de la prévision de production PV. Ces informations sont intégrées dans le modèle spatio-temporel initial non pas comme des variables explicatives mais en créant une dépendance entre les valeurs des coefficients du modèle et celles prises par la variable météorologique locale. Les informations météorologiques locales sont donc directement prises en compte dans le processus d'estimation des coefficients du modèle. Cette modification a permis d'améliorer les performances du modèle spatio-temporel sur les deux premières heures de l'horizon de prévision pour les variables météorologiques de vitesse du vent et d'humidité relative.

Dans les cas présentant un nombre important de sites de production tel que le deuxième jeu de données utilisé dans cette thèse, nous avons été confronté à des problèmes de surapprentissage, de dimension des modèles et de parcimonie. Ces problèmes sont en effet dus au fait que le modèle spatio-temporel considère non seulement les séries de production de sites voisins mais aussi les séries retardées respectives. Nous avons proposé une solution à ce problème qui est automatique, directement intégrée au modèle spatio-temporel et qui produit de bonnes prévisions de la production PV. Cette solution est basée sur la régularisation Lasso. Cette régularisation permet non seulement de réduire la dimension du modèle en sélectionnant les variables adéquates pour la prévision mais elles contribuent aussi à la parcimonie et à l'amélioration des performances du modèle de prévision. En effet, nous avons montré que la mise en place de cette méthode de sélection de variables participe à la réduction des erreurs de prévision au sens de différents critères.

Enfin, nous avons proposé une comparaison entre le modèle spatio-temporel purement statistique avec sélection de variables intégrés et un modèle spatio-temporel qui intègre les prévisions météorologiques de type NWP. Ces deux approches sont très différentes notamment par le type de données utilisées et par l'approche statistique utilisée. L'idée principale est de déterminer à partir de quels horizons, les données issues de modèles de prévision numériques permettent d'obtenir de meilleures performances que la seule utilisation des données de mesure. L'horizon de prévision considéré est 12h. Nous montrons que pour des horizons inférieurs à 6 heures, le modèle basé uniquement sur les mesures de production est plus rapide, et ses performances en prévision sont très bonnes. Sur ce même

horizon l'intégration des données issues de modèles NWP n'apporte pas d'amélioration significative face au modèle de données de mesures pures. Pour les horizons plus longs (au-delà de 6 heures), les performances du modèle spatio-temporel avec les seules données de mesure se dégradent très significativement. Cette dégradation est corrigée par le modèle intégrant les données NWP.

Chapitre 4

Modèles spatio-temporels probabilistes pour la prévision de production PV

4.1 Introduction

Les modèles spatio-temporels définis dans le chapitre précédent fournissent des prévisions déterministes qui ne permettent pas d'évaluer les incertitudes liées aux prévisions. Les prévisions probabilistes répondent à ce besoin. En effet, elles permettent non seulement de fournir plus d'informations sur la distribution future de la production PV mais aussi d'évaluer les incertitudes liées à la prévision. Cette quantification des incertitudes sur les prévisions est très utile pour les différents acteurs du domaine notamment pour la prise de décision et l'évaluation des risques. Il existe dans la littérature très peu de modèles qui exploitent les relations spatio-temporelles entre mesures de production pour la prévision probabiliste à court-terme de la production PV. On retrouve un modèle basé sur une combinaison entre le modèle autorégressif vectoriel et la méthode du gradient boosting [99], une autre approche est non paramétrique [57] et une dernière sur les champs aléatoires gaussiens [116].

Dans ce chapitre, nous proposons une méthodologie de prévision probabiliste à court terme qui exploite l'information disponible à partir d'installations photovoltaïques à grande échelle. L'objectif est d'améliorer la prévisibilité pour les horizons compris entre 0 et 6 heures. Les prévisions attendues sont donc les distributions futures de la production PV. Contrairement aux méthodes proposées dans la littérature pour ces horizons, notre approche ne nécessite pas d'informations provenant de caméras ou d'images satellites. Les prévisions météorologiques de type NWP sont utilisées comme informations complémentaires dans les modèles proposés pour construire les densités de probabilité les plus proches possibles des réalisations. L'utilisation de ces prévisions permet aussi d'avoir un meilleur couplage entre les prévisions spatio-temporelles à très court terme et les prévisions classiques faites pour la journée à venir (J+1). Le modèle probabiliste proposé est évalué grâce à un modèle de référence de même nature, c.à.d. probabiliste. Les performances des modèles sont évaluées avec les critères appropriés des évaluations probabilistes sur le jeu de données d_2 qui comporte un grand nombre d'installations PV. Les problèmes de dimensionnalité et de parcimonie des modèles dus à la diversité des sources de données disponibles (données historiques de chaque centrale, prévisions NWP) sont traités par deux méthodes de sélection de variables. Une partie des résultats présentés dans ce

chapitre ont fait l'objet du deuxième article soumis et présenté dans l'annexe A.

4.2 Modèles probabilistes pour la prévision spatio-temporelle de la production PV

Dans cette partie nous présentons le modèle de référence choisi. C'est un modèle de prévision probabiliste très courant dans la littérature de la prévision de la production PV : l'estimation par noyau de la densité (Kernel Density Estimation, KDE) [57, 117, 26]. Nous présentons ensuite les modifications proposées pour affiner la sélection des variables en entrée du modèle. Le modèle de référence KDE n'exploite que les corrélations temporelles pour fournir des prévisions probabilistes de la production PV. Nous présentons donc dans la suite la nouvelle méthode de prévision probabiliste qui exploite les corrélations spatio-temporelles pour fournir des prévisions probabilistes de la production PV. Cette méthode est basée sur la régression quantile et sur la régularisation Lasso. Son principe, les conditions de sa mise en œuvre et les modifications proposées pour améliorer son efficacité sont présentées.

4.2.1 L'estimation par noyau (KDE)

La méthode d'estimation par noyau KDE est une méthode d'estimation non paramétrique de la densité [118, 119, 120]. Les KDE permettent une meilleure réduction des erreurs d'estimation en comparaison aux méthodes paramétriques car elles n'admettent pas d'hypothèses sur la distribution sous-jacente au phénomène estimé. Le problème de minimisation du KDE consiste à fournir une estimation de la densité de probabilité f d'une variable aléatoire X. L'estimateur des noyaux multidimensionnel (taille n) s'écrit :

$$\hat{f}(x) = \frac{1}{N|\mathbf{H}|} \sum_{i=1}^{N} K\left(\mathbf{H^{-1}}(x - x_i)\right) \tag{4.1}$$

où x représente le point d'évaluation de l'estimateur, $x_i, i = 1 \dots N$ les données. $\mathbf{H}$ est une matrice $n \times n$ appelée matrice bandwidth ou de "lissage", $|\mathbf{H}|$ étant son déterminant. K est le noyau choisi.

Le noyau K et la matrice de lissage $\mathbf{H}$ sont les deux paramètres à déterminer pour mettre en œuvre un modèle de KDE. L'impact du type de noyau sur la qualité de l'estimation est faible mais les noyaux gaussiens nécessitent d'importantes capacités de calcul [121]. Nous avons donc choisi des noyaux d'Epanechnikov dont la fonction s'écrit :

$$K(u) = \frac{3}{4\sqrt{5}}(1 - u^2/5) \qquad u \in [-1, 1]. \tag{4.2}$$

Le passage à la version multivariée donne le produit $K(u) = \prod_{j=1}^{n} K(u_j)$.

La matrice de lissage $\mathbf{H}$ est le paramètre le plus important de l'estimation KDE car elle a une grande influence sur la qualité de l'estimation [120]. Il existe plusieurs méthodes pour choisir la matrice de lissage optimale [122, 123]. Nous utilisons ici la méthode de validation croisée [124]. De plus, puisque les données de production PV sont positives et bornées, la méthode d'estimation a été modifiée pour prendre en compte cette particularité comme proposé dans la littérature [120, 121].

Soit $Y \in \mathbb{R}^p$ la variable aléatoire dont les réalisations sont la production PV d'une centrale, $X \in \mathbb{R}^p$ les variables explicatives. La prévision de la production consiste à calculer la densité de probabilité de la variable conditionnée $Y_{t+k}|X_t$ où t est l'instant auquel la prévision est faite et k l'horizon de prévision. Cette densité s'obtient par :

$$f_{Y_{t+k}|X_t} = \frac{f_{Y_{t+k},X_t}}{f_{X_t}}. \tag{4.3}$$

La densité estimée se retrouve donc par :

$$\hat{f}_{Y_{t+k}|X_t} = \frac{1}{|\mathbf{H}|} \sum_{i=1}^{N} w(x, x_i) K(\mathbf{H^{-1}}(y - y_i)) \tag{4.4}$$

avec

$$w(x, xi) = \frac{K\left(\mathbf{H^{-1}}(\mathbf{x} - \mathbf{x_i})\right)}{\sum_{j=1}^{N} K\left(\mathbf{H^{-1}}(\mathbf{x} - \mathbf{x_j})\right)}.$$

4.2.2 La régression quantile

Lorsqu'on s'intéresse à un échantillon de réalisations d'une variable aléatoire réelle, la moyenne n'est pas toujours une statistique satisfaisante à elle seule. Une autre statistique intéressante est le quantile d'ordre α qui est défini par

$$q_\alpha(Y) = \mathbb{F}^{-1}(\alpha) = \inf\{y, \mathbb{F}(y) \geq \alpha\}, \tag{4.5}$$

lorsque Y a pour fonction de répartition $\mathbb{F}$. La régression quantile [55] est basée sur le fait que le quantile d'ordre α est solution du problème de minimisation suivant :

$$q_\alpha(Y) = \arg\min_g \ \mathbb{E}[\rho_\alpha(Y - g(X))] \tag{4.6}$$

avec $\rho_\alpha(u) = u\ (\alpha - \mathbf{1}_{\{u<0\}})$ une fonction de perte L_1 appelée "pinball loss". En étendant cette approche au quantile conditionnel d'ordre α, on retrouve le problème d'estimation :

$$q_\alpha(Y|X) = \arg\min_g \ \mathbb{E}[\rho_\alpha(Y - g(X))|X = x]. \tag{4.7}$$

La régression quantile permet d'évaluer comment les quantiles conditionnels de la variable d'intérêt se déforment en fonction des régresseurs X. De plus, avec la fonction pinball loss, seul le signe des écarts importe pour la minimisation. Cela se traduit par une pénalisation plus faible des très grands écarts donc une meilleure robustesse aux valeurs extrêmes ou aberrantes [125]. Dans le cas simple où la relation entre Y et X est linéaire ($Y = X'\beta + \varepsilon, \beta \in \mathbb{R}^p$) on construit l'estimateur :

$$\hat{\beta}_\alpha = \arg\min_\beta \ \sum_{i=1}^{n} \rho_\alpha(Y_i - X_i'\beta). \tag{4.8}$$

L'équation (4.8) n'a pas de solution explicite, le programme doit donc être résolu numériquement. Un problème important dans cette résolution numérique est que la fonction ρ_α n'est ni dérivable en 0, ni strictement convexe. Les algorithmes standards tels que celui de Newton Raphson ne peuvent être utilisés ici. On reformule donc l'équation comme un programme d'optimisation linéaire :

$$\underset{(\beta,u,v)\in\mathbb{R}^p\times\mathbb{R}_+^{2n}}{min} \alpha\mathbf{1}'u + (1-\alpha)\mathbf{1}'v \quad s.c. \ \ X'\beta + u - v - Y' = 0 \tag{4.9}$$

avec $\mathbf{1}$ un vecteur composé de 1 et de taille n. Les problèmes linéaires de ce type sont résolus par la méthode du simplexe dans le cas de petits échantillons ou des méthodes de point intérieur dans le cas de grands échantillons. Ces méthodes ainsi que quelques propriétés asymptotiques de l'estimateur sont présentées dans l'annexe B.

4.2.3 Extension des modèles : cas de données de grande dimension

La prévision spatio-temporelle pour les cas avec un nombre important de centrales photovoltaïques pose le problème du choix des variables d'entrée appropriées pour les modèles impliqués. Les entrées possibles incluent la série de production du site pour lequel la prévision est faite, les mesures des sites voisins et leurs séries temporellement retardées respectives ainsi que les données météorologiques NWP. De plus, les centrales photovoltaïques peuvent être réparties sur une large surface de plusieurs kilomètres carrés, ce qui requiert de choisir adéquatement les points de grille pouvant être pris en considération. Toutes ces informations représentent une quantité importante de variables d'entrée potentielles pour les modèles de prévision et soulèvent la question de leur dimensionnalité et aussi de la parcimonie. Nous proposons ici deux méthodes de sélection de variables adaptées à chacun des deux modèles précédemment définis.

Le critère d'information mutuelle

La première méthode de sélection de variables est basée sur le critère d'information mutuelle [126]. Cette technique a été utilisée avec succès dans [127] pour sélectionner des variables pour la prévision de l'énergie éolienne. La sélection de variables suivant le critère de l'information mutuelle est utilisée ici pour déterminer les variables à utiliser pour le modèle KDE. L'information mutuelle est une mesure de "distance" (Kullbac-Leiber) entre deux fonctions de densité de probabilité. Soit X la variable aléatoire représentant la météorologie, Y celle de la production PV, f_X et f_Y leurs fonctions de densité de probabilité respective. L'information mutuelle permet d'évaluer la distance entre la fonction de densité de probabilité de la loi jointe $f_{X,Y}$ (entre la production et la variable météorologique), et celle de la loi jointe $f_X.f_Y$. L'égalité de ces deux densités signifie que X et Y sont indépendants. De plus, une valeur élevée de l'information mutuelle traduit une forte dépendance entre X et Y. Cette évaluation de distance est faite pour chacune des variables météorologiques. Les variables météorologiques sont ensuite classées en fonction de la valeur de l'information mutuelle calculée. L'information mutuelle moyenne pour les variables Y et X s'obtient par la formule :

$$I(X,Y) = \int f_{X,Y} log\left(\frac{f_{X,Y}}{f_X.f_Y}\right). \tag{4.10}$$

L'information mutuelle entre X et Y est bornée par le minimum des entropies de X et Y [126]. Cette valeur seuil est utilisée pour normaliser les valeurs du critère d'information mutuel. Dans la suite nous utiliserons la valeur normalisée du critère d'information mutuelle. Ces valeurs sont calculées pour les prévisions NWP afin de choisir les plus adéquates pour la mise en œuvre des KDE.

La pénalisation en régression quantile : Lasso

Le Lasso (Least absolute shrinkage and selection operator) est la même pénalisation [108] que celle définie dans la partie 3.3.3 mais qui, cette fois, est intégrée à un modèle de régression quantile (dont le problème d'estimation est L1 et non quadratique). L'estimateur du quantile de niveau α s'écrit :

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{\mathbb{E}[\rho_\alpha(Y - X'\beta)] + \lambda\|\beta\|_1\right\} \tag{4.11}$$

avec $\rho_\alpha(u) = u\ (\alpha - \mathbf{1}_{\{u<0\}})$ la fonction "pinball loss". Pour les grandes valeurs de λ, certains coefficients sont annulés permettant ainsi un choix de variables.

Le problème (4.11) peut être reformulé en :

$$\underset{\beta}{\operatorname{argmin}} \{\mathbb{E}[\rho_\alpha(Y - X'\beta)]\} \quad \text{st} \quad \|\beta\|_1 \leq \hat{R}_\lambda \tag{4.12}$$

Cette modification de la régression quantile classique proposée ici est essentielle dans le paradigme spatio-temporel en raison du nombre élevé des potentielles variables explicatives. Le nouveau modèle de régression quantile qui incorpore cette procédure de sélection de variable Lasso est nommé QR-Lasso dans toute la suite.

4.3 Évaluation des modèles

La performance du modèle proposé est évaluée pour des horizons de 6 heures avec un pas de temps de 15 minutes. Les prévisions sont mises à jour toutes les 15 minutes. Cette performance est comparée à celle du modèle de référence. Les données NWP ont été obtenues grâce au centre européen pour les prévisions météorologiques à moyen terme (ECMWF). Les modèles ont été développés à l'aide du logiciel R [128]. Les librairies utilisées sont *ks, rq*. Les prévisions probabilistes (de distribution) sont caractérisées par leurs quantiles. Le dernier tiers de l'ensemble de données (environ 5 mois) a été utilisé comme ensemble de test.

Critères d'évaluation des prévisions probabilistes
L'évaluation des prévisions probabilistes nécessite des critères spécifiques qui prennent en compte les différents points de la densité et pas seulement la moyenne. Les prévisions fournies par les modèles QR-Lasso et KDE seront évaluées par des critères d'évaluation liés aux quantiles. Ces critères d'évaluation des quantiles des distributions prévues sont très utilisés dans la littérature. Soit $\hat{q}^\alpha_{t+h}$ le quantile conditionnel de niveau α estimé à l'instant t pour l'horizon h, les principaux critères d'évaluation utilisés sont :

— la fiabilité d'un quantile estimé ; on la définit comme la proportion excédante constatée en prévision pour le quantile

$$rel^\alpha_h = \alpha - \frac{1}{n}\sum_{t=1}^{n} m_{p_{t+h} \leq \hat{q}^\alpha_{t+h}}$$

où $m_{p_{t+h} \leq \hat{q}^\alpha_{t+h}}$ est la proportion de données constatées en dessous du quantile estimé

— la statistique de Kolmogorov-Smirnov KS (voir figure 4.1) qui est la plus grande différence en valeur absolue entre la proportion théorique d'excédance et celle observée pour la prévision

$$ks = \max_\alpha \ |\alpha - \frac{1}{n}\sum_{t=1}^{n} m_{p_{t+h} \leq \hat{q}^\alpha_{t+h}}|$$

— le MAEP, Mean Absolute Excess Probability (voir figure 4.1)

$$MAEP = \int_0^1 |\hat{q}_\alpha - q_\alpha| d_\alpha$$

qui est l'aire entre la courbe de la proportion d'excédance théorique et celle obtenue en prévision

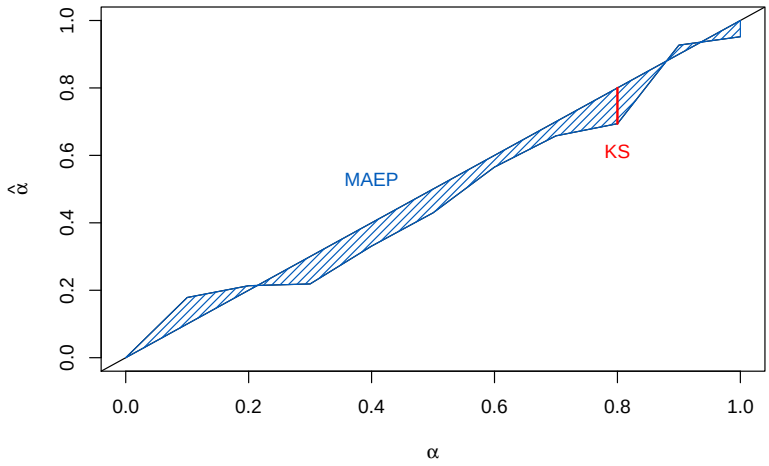


FIGURE 4.1 – Les critères de KS et de MAEP

— la finesse des intervalles de prévision (sharpness) ; elle peut se mesurer par la largeur moyenne de l'écart inter-quantile de niveau α d'une distribution

$$shap_h^{\alpha} = \frac{1}{n}\sum_{t=1}^{n}(\hat{q}_{t+h|t}^{1-\frac{\alpha}{2}} - \hat{q}_{t+h|t}^{\frac{\alpha}{2}})$$

— le score pinball loss qui est la moyenne de la fonction pinball loss évaluée pour tous les quantiles.

4.3.1 Évaluation des méthodes de sélection de variables

Information mutuelle

La procédure de sélection de variable permet de choisir les variables pertinentes parmi de nombreuses entrées, mesures et prévisions issues de modèles NWP disponibles dans les cas où l'on dispose d'un grand nombre d'installations photovoltaïques. Dans le tableau 4.1, nous présentons les valeurs du critère d'information mutuelle normalisées entre la production PV de quatre centrales photovoltaïques et les variables NWP. L'information mutuelle n'a pas été calculée pour la variable de rayonnement solaire à la surface ou sur plan incliné puisque le rayonnement solaire est à la base du processus de production PV. La variable de rayonnement solaire à la surface est d'office intégrée au modèle. Le tableau montre que les deux variables les plus importantes sont la température et l'humidité relative, suivies par la direction du vent. L'importance du niveau de précipitation est la plus faible.

Le modèle KDE de référence sera alors multivarié avec les quatre variables exogènes suivantes :

— le rayonnement solaire net au sol (TSR) ;

— la température à 2m du sol (2T) ;

— l'humidité relative (RH) ;

— la direction du vent (10U/V).

Tableau 4.1 – Valeurs normalisées du critère d'information mutuelle entre productions et prévisions NWP pour 4 centrales du jeu de données.

Variables NWP	Mutual information (%)			
	P_1	P_2	P_3	P_4
Température (2T)	72.38	70.74	80.79	78.22
Humidité relative (RH)	70.56	74.56	74.83	71.27
Direction du vent	25.11	21.88	26.11	21.86
Vitesse du vent	6.18	5.78	6.57	5.46
Précipitation	0.18	0.48	0.26	0.27

L'efficacité de la prévision par la méthode du KDE avec sélection de variables en amont par le critère de l'information mutuelle est aussi évaluée. Pour cela, la qualité des quantiles prévus est examinée à la fois visuellement à la fois au sens des critères d'évaluation probabiliste de prévision. La figure 4.2 présente les productions prédites avec le modèle KDE multivarié pour les centrales P_1 à P_3 pour On remarque que les intervalles inter-quantiles sont larges traduisant une sur-estimation des densités pour les niveaux de production les plus élevés. De plus cette taille d'intervalles de confiance ne permet pas de juger efficacement de la fiabilité des estimations.

QR-Lasso

Les variables météorologiques sélectionnées par le critère de l'information mutuelle sont utilisées comme entrées dans le modèle QR-Lasso en complément des données de production. Pour chaque installation PV $Pi, i = 1, \ldots, 136$, le nombre total de variables d'entrée est de $814 = 4$ (prévisions NWP) $+ 135 \times 6$ (nombre de délais). La figure 4.3 présente le nombre de coefficients non nuls liés aux variables de production PV sélectionnées par le modèle QR-Lasso pour chaque décile. Le nombre maximum de variables sélectionnées est de 320, montrant que le processus de sélection a une bonne efficacité. De plus, on note une réduction du nombre de variables sélectionnées pour les quantiles extrêmes, ce qui est concordant avec le faible nombre d'observations qui caractérise ces quantiles. La performance du processus de sélection des variables est également évaluée en examinant la position des centrales électriques sélectionnées par rapport à l'installation photovoltaïque pour laquelle la prévision est faite. La figure 4.4 présente les positions des centrales PV sélectionnées lors de la prévision de la médiane de la production future pour 3 centrales PV situées respectivement à l'ouest, au centre et à l'est de la région couverte par le jeu de données. L'horizon considéré est 15 min. Le nombre d'installations photovoltaïques sélectionnées est de 33 pour la centrale à l'ouest, 43 pour celle au centre et 39 pour celle à l'est (parmi 135 initiales) et la centrale la plus lointaine choisie est à 89 km de distance. Cette importante réduction du nombre de variables montre l'efficacité de la méthode de sélection de variables quant à la réduction de la dimension du problème.

4.3.2 Fiabilité et niveau de précision des quantiles estimés

La figure 4.5 présente pour les déciles (intervalles prédictifs), la fiabilité (reliability) ou l'écart observé par rapport à la couverture nominale pour les horizons 3h. La fiabilité parfaite signifie qu'il n'y a pas d'écart et est représentée par la ligne horizontale noire ($y = 0$) sur la figure. Deux écarts sont tracés pour le modèle KDE représentant deux matrices de lissage différentes $H1, H2$. Ces deux matrices de lissage sont respectivement les matrices optimales des méthodes "unbiased cross validation" et "smooth cross-validation"

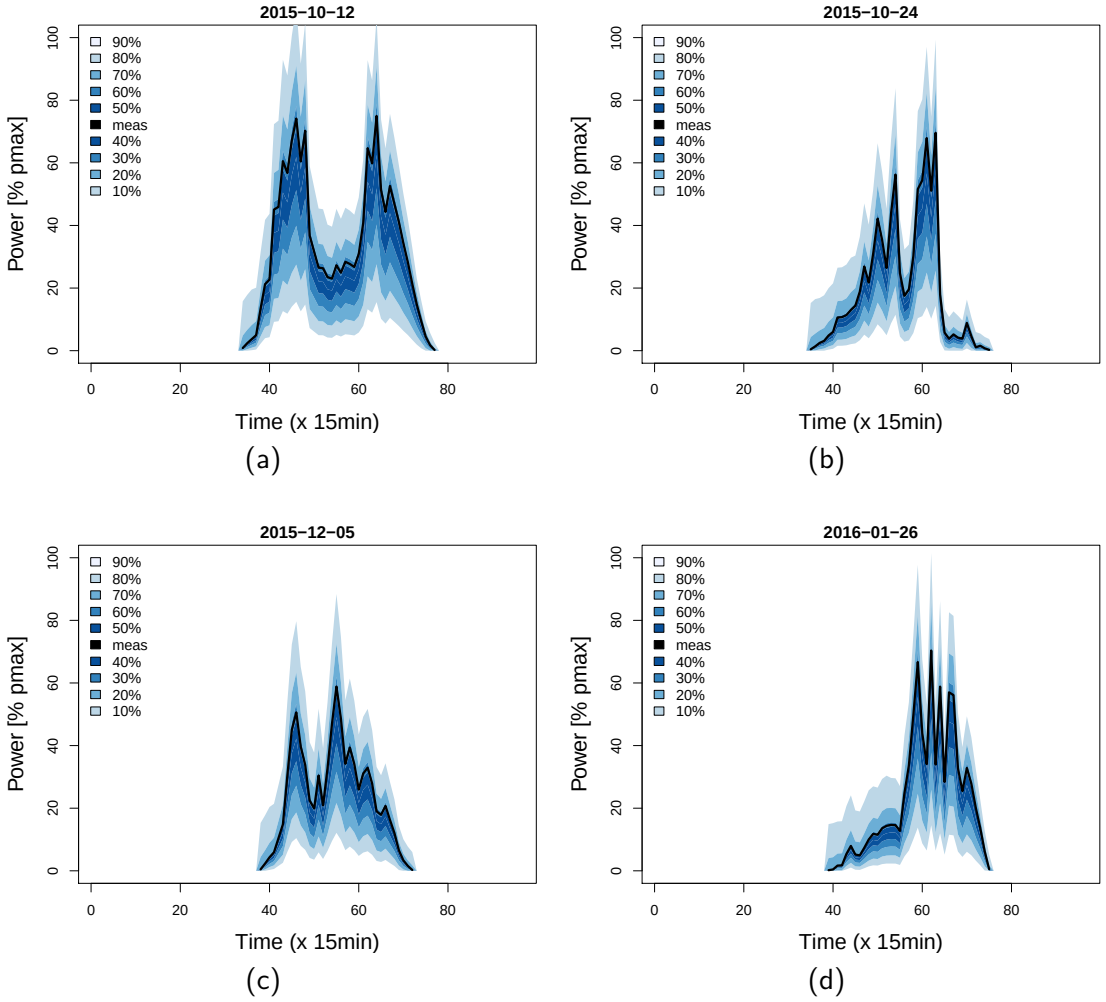


FIGURE 4.2 – Prévisions de production faites avec la méthode du KDE pour la centrale P_3

(voir [124] pour plus de détails sur les fonctions et le processus de minimisation). La figure montre que selon les valeurs des matrices de lissage, les densités prédictives peuvent être surestimées (ligne rouge pour KDE / H1) ou sous-estimées (ligne verte pour KDE / H2). Les valeurs de l'écart pour le modèle QR-Lasso ne dépassent pas 5% et sont inférieures à celles du modèle KDE. De plus, les valeurs de fiabilité du modèle QR-Lasso sont légèrement inférieures à la fiabilité du modèle de machine learning proposé dans [57].

Les valeurs de précision des quantiles (sharpness) sont présentées sur la figure 4.6 pour les mêmes intervalles de prévision que celui de la fiabilité. La figure montre que la sharpness augmente avec le quantile nominal (taux de couverture). Les valeurs vont de 3% à 58% et sont inférieures à celles observées pour le modèle KDE pour les quantiles inférieurs à q^{60}. Pour les quantiles supérieurs, le modèle QR-Lasso présente des valeurs de sharpness supérieures à celles du modèle KDE.

4.3.3 Puissances prédites

Les critères de fiabilité et de précision ne sont pas suffisants pour évaluer les performances du modèle QR-Lasso. La figure 4.7 représente les puissances prédites comme un

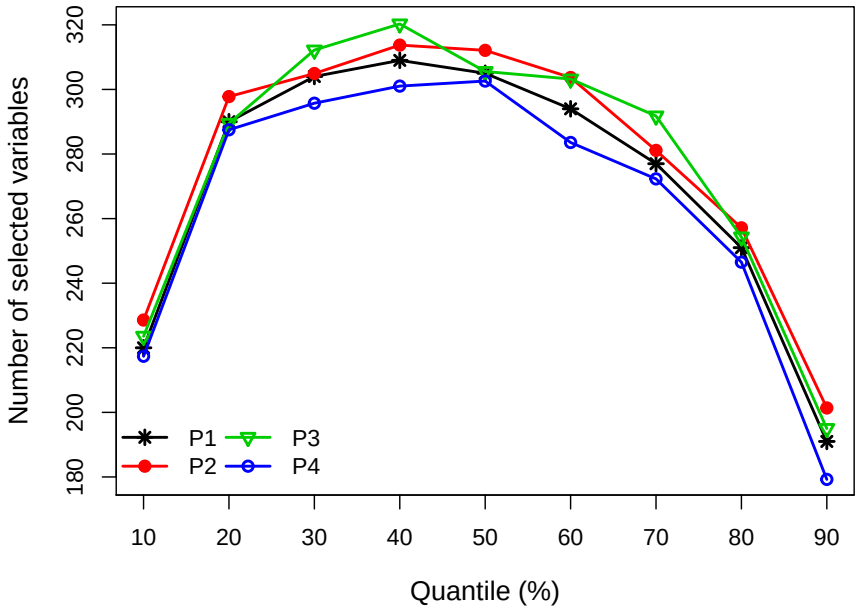


FIGURE 4.3 – Nombre de coefficients non nuls retenus hors variables météorologiques par le modèle QR-Lasso par décile. Une ligne par centrale pour quatre centrales sur l'échantillon de validation. L'horizon est 6 heures.

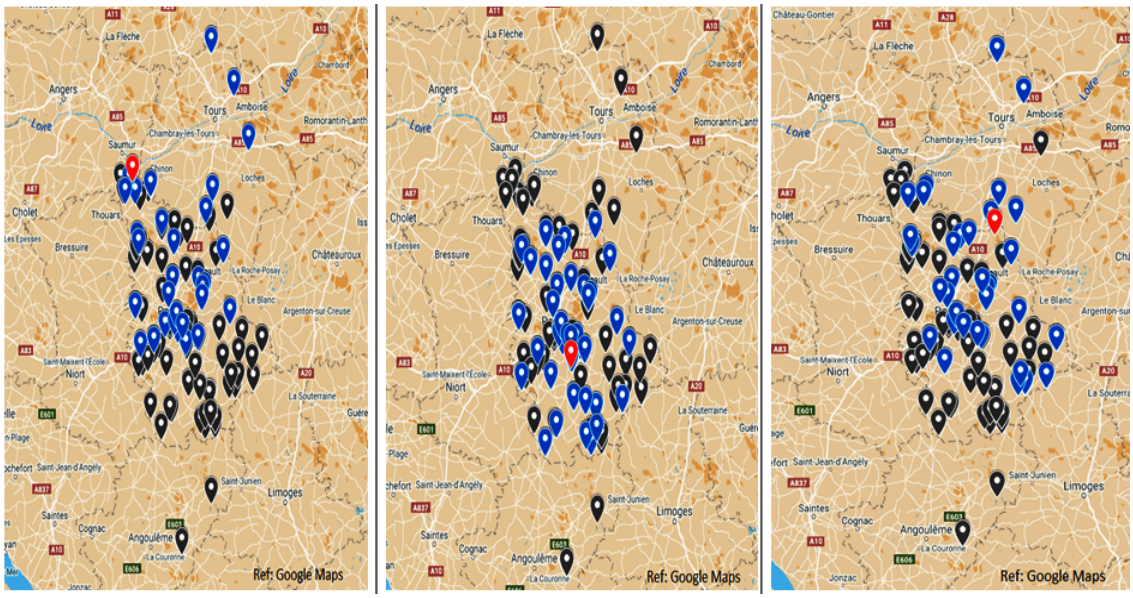


FIGURE 4.4 – Les centrales sélectionnées par le modèle QR-Lasso pour la prévision de la médiane pour trois centrales d'intérêt (en rouge). Les centrales en bleu sont celles sélectionnées parmi les autres (en gris). L'horizon de prévision considéré est 15 min.

ensemble d'intervalles de prévision pour six jours de l'ensemble de tests pour la centrale P_1. Le modèle QR-Lasso anticipe assez bien les variations de production. La longueur des intervalles inter-quantile est faible par rapport à ce qui peut être observé dans la littérature [57]. La plupart des variations pour les jours caractérisés par une forte variabilité de la production sont assez bien prédites dans l'ensemble.

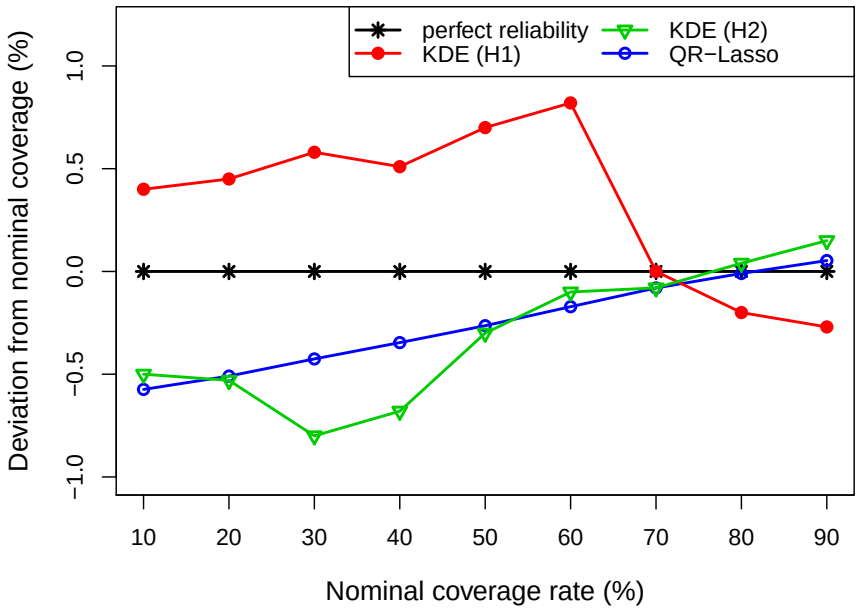


FIGURE 4.5 – Fiabilité des quantiles estimés pour 3 heures d'horizon. Pour le modèle KDE, deux matrices de lissage ont été utilisées (courbes rouges et vertes) pour montrer l'impact de la matrice de lissage sur la fiabilité. La courbe bleue représente le modèle QR-Lasso

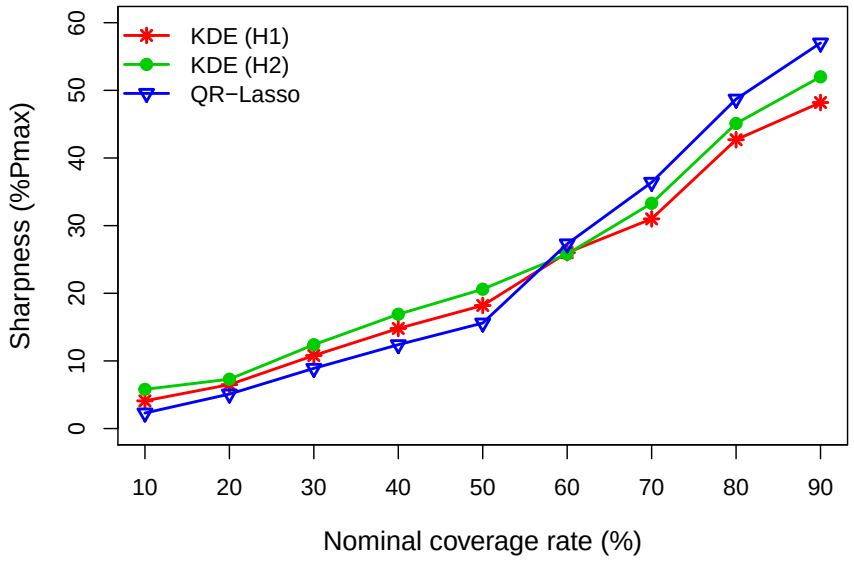


FIGURE 4.6 – Sharpness des densités prévues pour des horizons 3h. Les courbes rouges et vertes représentent respectivement les modèles KDE avec deux cas de matrice de lissage. La courbe bleue représente le modèle QR-Lasso.

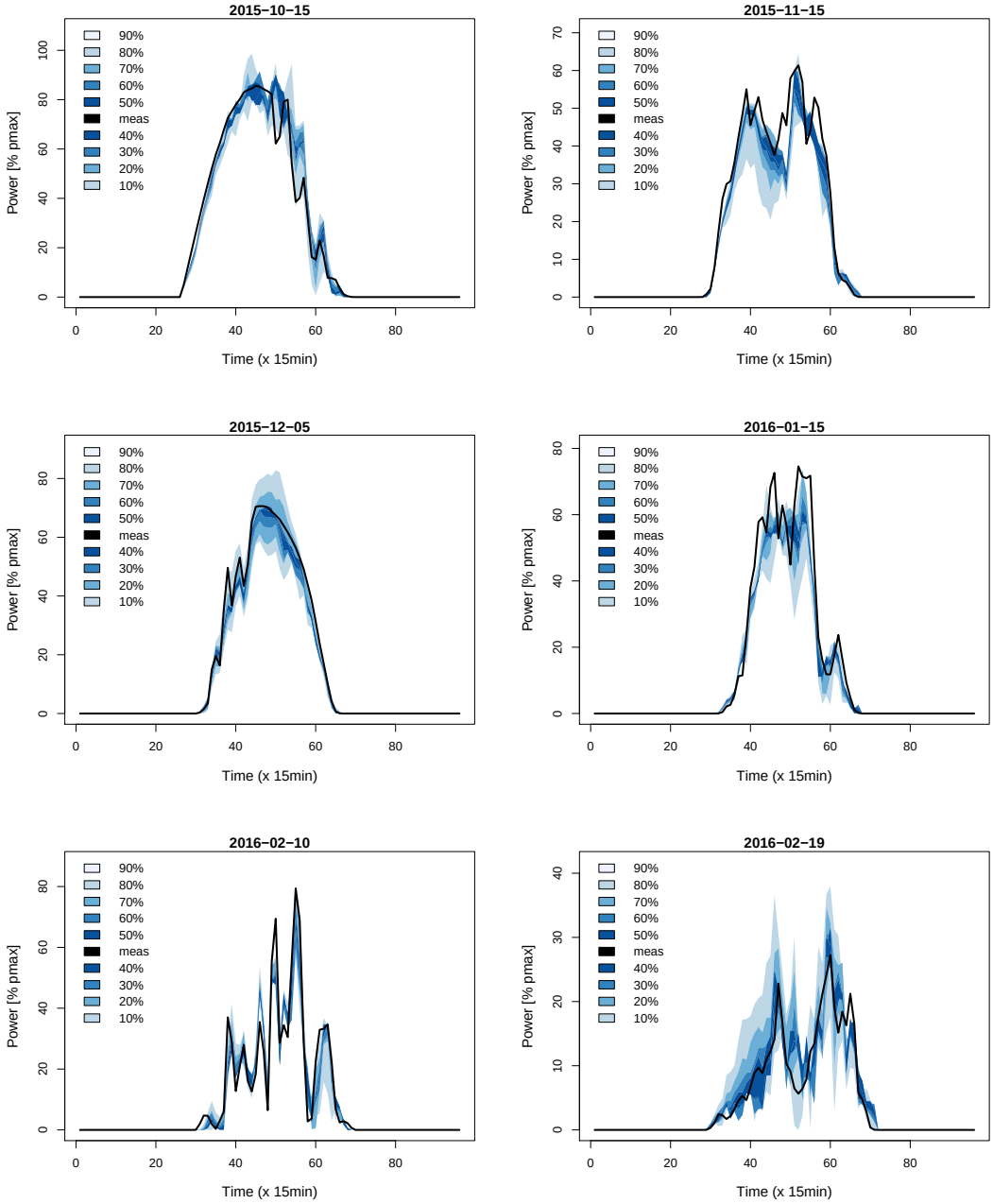


FIGURE 4.7 – Exemples de quantiles prédits par le modèle QR-Lasso pour 6 jours avec des conditions météorologiques différentes. Les quantiles sont représentés de 10% à 90% avec un pas de 10%. Le pas de temps est 15 min. L'horizon est 6 heures.

Score CRPS

Ces bonnes performances du modèle QR-Lasso sont confirmées par l'analyse des scores de probabilité (CRPS) [129] de nos modèles. Le CRPS évalue la distribution en synthétisant les critères de fiabilité et de précision. Le CRPS est défini pour une fonction de distribution cumulative F et son observation y par :

$$crps(F, y) = \int_{-\infty}^{\infty} \left(F(x) - H(x - y)\right)^2 dx \tag{4.13}$$

Tableau 4.2 – CRPS du modèle spatio-temporel QR-Lasso et du modèle de référence KDE pour cinq centrales du jeu de données.

Power plant	CRPS (% Pmax)	
	KDE	QR-Lasso
P_1	7.2	5.2
P_2	8.32	4.14
P_3	7.65	5.23
P_4	8.67	5.14
P_5	8.32	4.38

avec $H(x)$ la fonction Heaviside dont la valeur est de 0 pour les x négatifs et 1 pour les valeurs positives de x. Le CRPS peut être interprété comme une mesure de la distance entre la fonction de distribution cumulée observée et celle prévue. Les valeurs moyennes de CRPS sur 6 heures d'horizon de prévision sont présentées dans le tableau 4.2 pour le modèle KDE de référence avec la matrice $H1$ (le meilleur des deux modèles KDE) et le modèle QR-Lasso pour cinq centrales du jeu de données. Les valeurs CRPS sont plus faibles pour le modèle QR-Lasso confirmant l'amélioration de la performance de prévision par rapport au KDE. De plus, ces valeurs du CRPS sont inférieures à celles de l'ensemble des modèles présentés dans la revue bibliographique [16].

4.4 Conclusion

Dans ce chapitre, nous avons proposé un modèle spatio-temporel probabiliste dans un double objectif. Le premier est de fournir des prévisions de production PV à court terme (inférieure à 6 heures) avec des niveaux d'erreurs associés afin de donner plus d'informations sur la distribution future de la production. Le second objectif est de proposer un modèle de prévision probabiliste qui puisse exploiter les corrélations spatio-temporelles mises en évidence dans le chapitre 1. Pour répondre à ces deux objectifs, nous avons dans un premier temps proposé un modèle de prévision probabiliste de référence. Le modèle retenu est le modèle KDE qui fournit des estimations de densité par de la régression à noyau. Le modèle que nous avons proposé pour la prévision spatio-temporelle de la production PV est dénommé QR-Lasso. C'est un modèle basé à la fois sur la régression quantile classique et sur la régression pénalisée de type Lasso. L'une des particularités de ce modèle est que l'estimation des coefficients n'est pas quadratique mais L_1 comme la norme de pénalisation. Ce modèle permet de gérer directement sans phase de prétraitement la grande quantité de données générées par l'utilisation des séries de centrales voisines et leurs valeurs retardées mais aussi les prévisions issues de modèles de prévision de type NWP. En effet, la pénalisation Lasso permet de sélectionner automatiquement les données les plus utiles pour la prévision PV. La recherche de parcimonie dans les modèles est aussi valable pour le modèle KDE, c'est pourquoi nous avons proposé l'utilisation du critère d'information mutuel pour déterminer les variables les plus intéressantes pour la prévision par le KDE.

Les critères d'évaluation des modèles de prévisions probabilistes ont été présentés. Ces critères ont ensuite été utilisés pour évaluer le modèle QR-Lasso proposé et comparer ses performances au modèle KDE de référence mais aussi à certains modèles de la littérature. Un jeu de données réelles avec un nombre élevé d'installations photovoltaïques a été utilisé pour montrer tout le potentiel de l'approche spatio-temporelle probabiliste. Le modèle QR-Lasso proposé présente une amélioration significative de la performance par rapport

au modèle KDE de référence pour l'ensemble des critères évalués.

Chapitre 5

Intégration des images satellites pour la prévision PV

5.1 Introduction

Les méthodes de prévision spatio-temporelles présentées dans les chapitres précédents sont basées sur les corrélations spatio-temporelles entre des sites de production géographiquement dispersés. Dans ce chapitre, nous nous intéressons à une autre source de données adaptées à la prévision à court-terme de la production PV : les images satellites. Ces images présentent des résolutions spatiales et temporelles plus fines que celles des modèles NWP et sont mises à jour beaucoup plus rapidement que ces derniers. Les images satellites sont utiles pour la prévision de production PV car elles permettent de prévoir les évolutions de la couverture nuageuse. Rappelons que la couverture nuageuse est la principale source de variabilité intra-journalière de la production. Les méthodes de traitement d'images sont utilisées pour extraire des images les informations relatives aux mouvements des nuages, notamment leurs vitesse et direction. Il existe plusieurs satellites en orbite autour de la Terre qui sont utilisés pour différentes applications dont la fourniture des images utilisées pour prévoir les mouvements des nuages. La figure 5.1 présente quelques satellites en orbite autour de la Terre utilisés pour l'observation des phénomènes météorologiques. L'objectif de ce chapitre est de proposer une méthode d'intégration des informations contenues dans les images satellites dans le modèle spatio-temporel défini dans le chapitre 3. Cette intégration vise à utiliser d'une source de données avec des temps de mise à jour rapide et des résolutions spatiales plus importantes. En effet, les images satellites peuvent couvrir de grandes régions avec des résolutions fines (environ un tiers de la Terre et une taille de pixel approximative de 5 km) et les délais d'acquisition des images sont très courts. Il existe aussi d'importantes archives d'images satellites exploitables.

La première partie de ce chapitre est consacrée à une revue de la littérature des méthodes de prévision à court-terme à partir d'images satellites et à la présentation des images satellites que nous avons utilisées dans le cadre de nos travaux. La deuxième partie décrit la démarche d'intégration des informations des images satellites dans un modèle spatio-temporel qui exploite les mesures de centrales voisines. La nature spatio-temporelle du modèle ainsi proposé est double en ce sens où elle provient à la fois de l'information distribuée de la production, mais aussi des images satellites. Le modèle intégrant les images satellites est évalué et comparé au modèle sans images satellites. Dans toute cette partie, seul le jeu de données d_2 a été utilisé. La dernière partie de ce chapitre est consacrée à une étude comparative des performances du modèle spatio-temporel en fonction des différentes sources de données. Nous proposons une approche incrémentale dans le sens

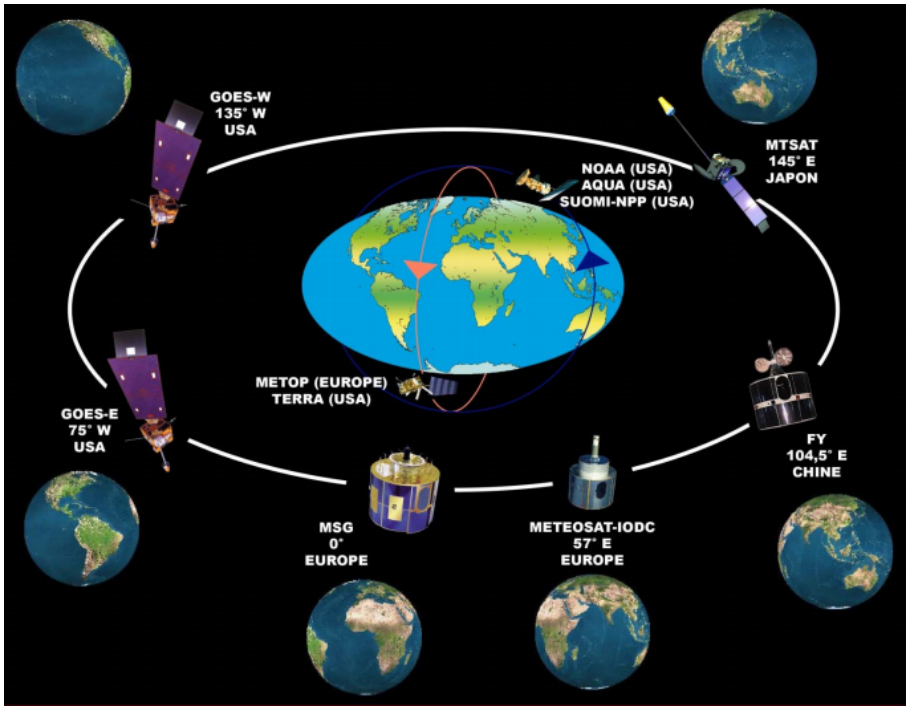


FIGURE 5.1 – Exemples de satellites météorologiques géostationnaires. Source : Météo France

où nous partons d'un modèle temporel, et nous évaluons l'apport respectif des mesures de centrales voisines, des prévisions NWP et des images satellites sur les performances de prévision. Cette étude permet de répondre à la question du choix des sources de données à privilégier en fonction des horizons de prévision envisagés. Cette question est cruciale dans un contexte caractérisé par la multiplication des mesures et des sources de données liées à la maîtrise de la production PV.

5.2 Revue de littérature et présentation des données satellites

5.2.1 État de l'art

Les images de satellites géostationnaires peuvent être utilisées pour estimer l'irradiation au sol. Il existe dans la littérature différentes méthodes pour réaliser cette estimation. La principale différence entre ces méthodes est le procédé utilisé pour caractériser les interactions entre le rayonnement solaire et l'atmosphère. On retrouve dans [130, 131] une revue des premiers procédés utilisés, classifiés selon qu'ils soient physiques ou statistiques. Les différentes évolutions dans la caractérisation des phénomènes atmosphériques et les avancées technologiques dans le domaine de l'imagerie satellitaire ont conduit à des méthodes de plus en plus performantes pour passer des images satellites à des données d'irradiation [132, 133, 134, 21, 20]. Les données satellites peuvent être couplées à des mesures d'irradiation au sol pour améliorer la qualité des estimations fournies ; c'est la méthode de site-adaptation qui permet de rapprocher les estimations des mesures réelles sur site [83, 135, 136].

Les données d'irradiation au sol produites soit par la seule exploitation des images satellites, soit par le couplage aux mesures au sol sont utilisées pour fournir des prévisions

d'irradiation pour des horizons allant de 0 (*nowcasting*) à 6 heures, en se basant sur des méthodes de prévision physiques, statistiques ou hybrides présentées dans les revues de littérature [137, 11]. On retrouve entre autres, les cloud motion vector (CMV) qui déterminent les vitesses et directions des nuages par l'analyse des images satellites [138, 139, 140, 84, 141, 19] afin de fournir de meilleures prévisions de l'irradiation. Les réseaux de neurones artificiels [142, 143, 144], les SVM [145, 146] et l'estimation bayésienne [147] sont aussi utilisés dans le cadre de la prévision de l'irradiation à partir d'images satellites. On retrouve aussi des méthodes spatio-temporelles [88] et des méthodes qui combinent à la fois images satellites, prévisions NWP et mesures au sol [148, 87].

5.2.2 Présentation des données satellites

Les données satellites utilisées dans le cadre de ces travaux sont issues de la base de données Helioclim [149] créée à partir des images des satellites MFG d'EUMETSAT (European Organization for the Exploitation of Meteorological Satellites). La version Helioclim-3 de cette base de données est l'une des plus performantes et présente de meilleures résolutions temporelles (15 minutes) et spatiale (3 km au nadir ; le nadir étant l'ensemble des points à surface de la Terre qui se trouvent directement en dessous de la trajectoire du satellite). Cette base de données repose sur la méthode de conversion d'images satellites en estimation de radiation au sol Héliosat-2 [134] dont nous ne présentons pas ici les caractéristiques. La station de réception METEOSAT de TRANSVALOR permet un traitement en temps réel des images. La base de données comporte les données d'irradiation sur plan horizontal, incliné et normal. Les pas de temps peuvent être de 15 min, horaires, journaliers ou mensuels. La couverture géographique correspondante va de -66°à 66°en latitude et longitude. Les données peuvent être post-traitées selon les étapes suivantes :

— chaque pixel est interpolé dans le temps à 15 min ;
— tous les pixels pour lesquels l'élévation du soleil est en dessous de 2 degrés sont interpolés ;
— les altitudes, biais et indice de ciel-clair sont corrigés.

Plus de détails sur la base de données et le traitement des images sont fournis par le site du service SODA (Solar radiation data) [1] et dans les références [150, 151, 152, 153].

5.3 Méthode d'intégration des données d'images satellites

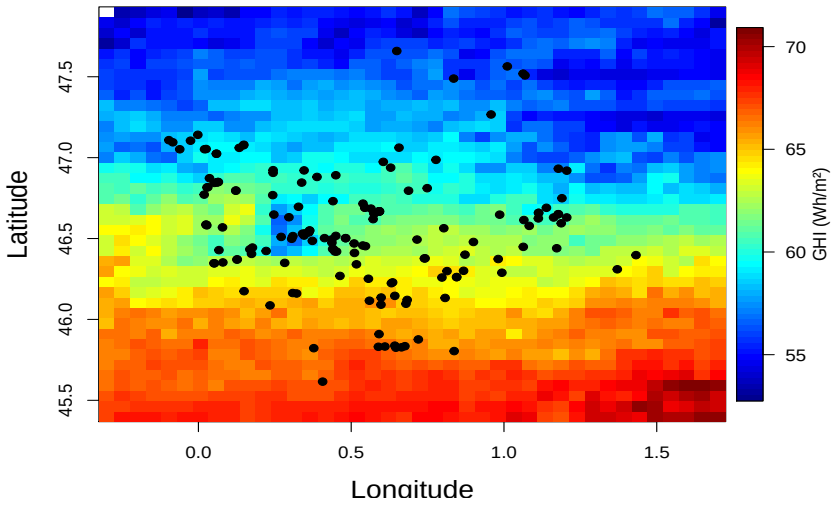
Nous présentons ici la procédure appliquée pour intégrer les informations des images satellites dans le modèle spatio-temporel. Nous avons utilisé ici des images satellites sous formes de cartes qui recouvrent la zone géographique des centrales du jeu de données d_2. La première étape de la procédure est de définir les points de cartes qui sont intéressants dans le cadre de la prévision PV et le traitement à apporter à ces points. Nous présentons ensuite le modèle statistique de prévision qui intègre les données satellites. Enfin, nous détaillons les résultats de l'évaluation des performances des prévisions issues de ce modèle et de sa comparaison aux modèles sans images satellites.

5.3.1 Détection des points d'intérêts des images

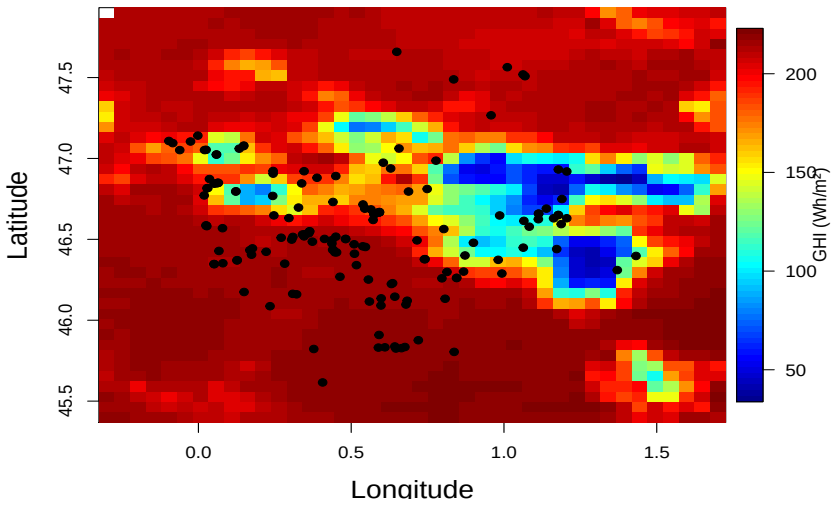
Sur une carte issue des données satellites, chaque point (pixel) correspond à une série temporelle à un pas de temps de 15 min des valeurs d'irradiation globale sur plan horizontal (GHI, Global Horizontal Irradiation) sur une période donnée. La figure 5.2 représente

1. www.soda-pro.com

la carte recouvrant les centrales du jeu de données d_2 pour deux dates de l'année 2015. Cette image comporte 3115 pixels. On remarque que les niveaux d'irradiation les plus élevés sont observés pour le mois de juillet avec des valeurs qui approchent du double de celles observées en janvier. Cela traduit l'effet de la saisonnalité sur le niveau d'irradiation reçu à la surface de la terre. Nous disposons aussi des cartes qui fournissent les séries d'irradiation ToA et d'index de ciel-clair.



(a) 01/01/2015 à 12h00 UTC



(b) 01/07/2015 à 12h00 UTC

FIGURE 5.2 – GHI, provenant d'une carte satellite recouvrant les sites de production du jeu de données d_2. Les centrales PV sont représentées par les points noirs (voir correspondance avec le graphique2.2).

Nous procédons pour chaque centrale PV à la détection des points de l'image satellite qui sont intéressants dans le cadre de la prévision spatio-temporelle. Cette analyse vise un double objectif, le premier est de déterminer la sous-partie de l'image dont les pixels (donc les séries d'irradiation) sont les plus liés à la production du site et de quantifier ce lien. Le second objectif est de déterminer l'intérêt de l'utilisation des images satellites.

Pour chacune des $s = 1, \ldots, n$ installations PV, l'identification des points de l'image intéressants pour la prévision se fait en plusieurs étapes. La première est de choisir les

pixels dans un rayon de 50 km autour du site d'intérêt. La figure 5.3 présente pour trois centrales du jeu de données, les zones d'intérêt retenues sur l'image pour le 01/01/2015 à 12h00 UTC.

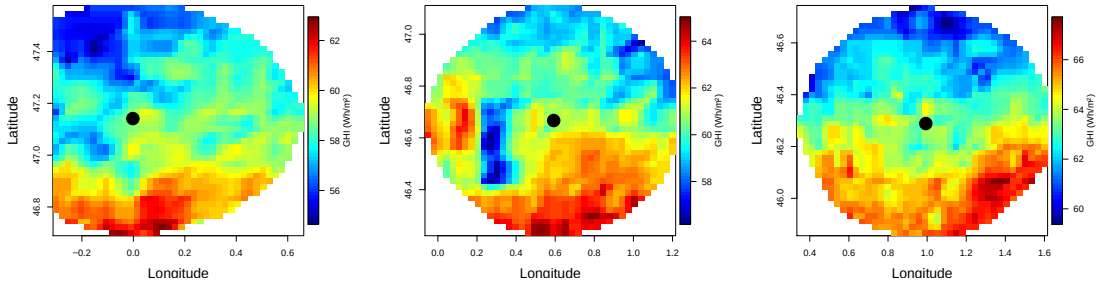


FIGURE 5.3 – Zones d'intérêt de l'image satellite retenues autour de trois sites du jeu de données d_2. L'image correspond au 01/01/2015 à 12h00 UTC. Les centrales PV sont représentées par les points noirs.

Nous avons fait ici un choix de taille de bloc fixe indépendant de l'horizon de prévision mais il existe des méthodes qui choisissent des tailles de bloc évolutives en fonction de l'horizon, notamment pour les applications de détection de déplacement de structures d'une image à une autre [154]. La deuxième étape consiste à transformer les séries d'irradiation GHI en série de production. Ensuite, nous évaluons le lien entre la mesure de production sur le site et les données issues de l'estimation satellite. Pour cela, nous utilisons un critère bi-varié d'association spatiale proposé par Wartenberg [155] inspiré de l'indice de Moran. Notons $X_{i,j}(t)$ l'estimation de la production fournie par la carte satellite au point (i, j) pour l'instant t, $Y_s(t)$ la mesure de production enregistrée sur le site s à l'instant t et τ un retard temporel. Le coefficient d'association spatiale inspiré de celui de Wartenberg s'écrit :

$$I_{(i,j),s}(\tau) = \frac{\sum_t \left(X_{i,j}(t)) - \bar{X}\right)\left(Y_s(t-\tau) - \bar{Y}_s\right)}{\sqrt{\sum_t \left(X_{i,j}(t) - \bar{X}\right)^2}\sqrt{\sum_t \left(Y_s(t-\tau) - \bar{Y}_s\right)^2}}. \tag{5.1}$$

Le coefficient ainsi défini permet de répondre aux problématiques de mise en évidence de lien à la fois spatial et temporel. Pour un décalage temporel nul de la série de mesure (τ=0), le coefficient d'association permet d'évaluer la corrélation entre les points de grille retenus et la mesure. La figure 5.4 présente pour trois centrales du jeu de données les valeurs de corrélations obtenues entre la production et les estimations pour les pixels de l'image satellite. Rappelons que pour chaque point de grille, nous disposons d'une série temporelle d'estimation de GHI et que les corrélations ont été calculées entre les productions équivalentes de ces estimations et la mesure sur site. Les corrélations les plus importantes sont observées pour les points les plus proches des centrales avec des valeurs de corrélations qui restent toutefois élevées sur l'ensemble de la zone d'intérêt.

Le calcul du coefficient d'association avec des valeurs de retards temporels non nuls ($\tau > 0$) permet d'évaluer l'intérêt de l'utilisation des images satellites pour les horizons envisagés. Dans notre cas, les retards temporels considérés sont liés aux horizons de prévision envisagés, c'est-à-dire 6 heures. Nous avons appliqué des décalages temporels de 1 heure à 6 heures aux séries de production PV. Les coefficients d'association entre ces séries et les estimations de production pour les pixels de l'image satellite sont calculés. Ils permettent de déterminer les zones d'intérêt des images satellites pour les prévisions

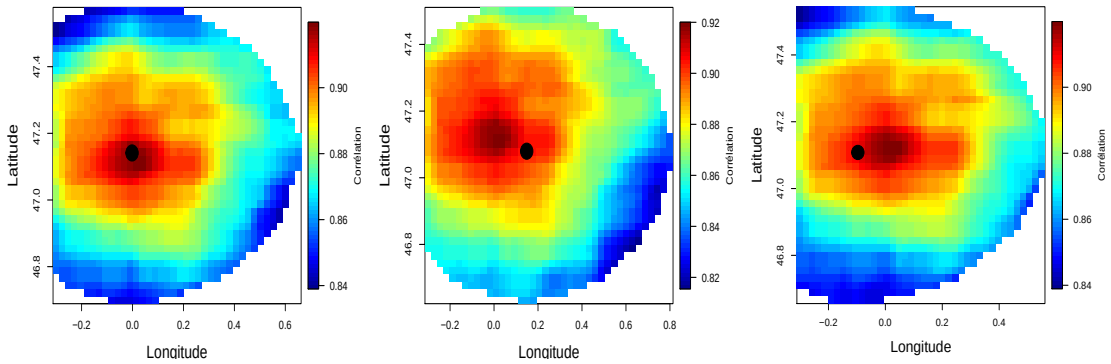


FIGURE 5.4 – Corrélation entre mesure sur site et estimation d'images satellites pour trois centrales PV. Les corrélations sont calculées pour le mois de janvier 2015 ; les séries sont au pas de temps de 15 min. Les centrales PV sont représentées par les points noirs.

pour des horizons correspondant au décalage appliqué. La figure 5.5 présente pour une centrale à l'ouest de la région couverte par le jeu de données, les valeurs du coefficient d'association pour différents décalages temporels. On remarque que pour des décalages temporels (ou horizons) faibles, la zone d'intérêt qui correspond aux plus fortes valeurs du coefficient d'association reste proche de la centrale d'intérêt. Cette zone s'éloigne progressivement lorsque les valeurs de décalage temporel augmentent. On peut expliquer ce déplacement par l'advection des nuages. De plus, les valeurs de coefficient d'association diminuent avec le décalage temporel et la zone d'intérêt se déplace vers le nord-est quand le décalage augmente. Les figures 5.6 et 5.7 présentent pour deux centrales respectivement situées au centre et à l'est de la région couverte par l'ensemble des centrales du jeu de données, les valeurs du coefficient d'association pour différents décalages temporels. On observe aussi le déplacement de la zone d'intérêt pour la centrale située la plus à l'est. Les valeurs de coefficients d'association observées sont aussi élevées, décroissantes avec l'horizon et indiquent que l'exploitation des images satellites pour la prévision permet la prise en compte d'informations supplémentaires.

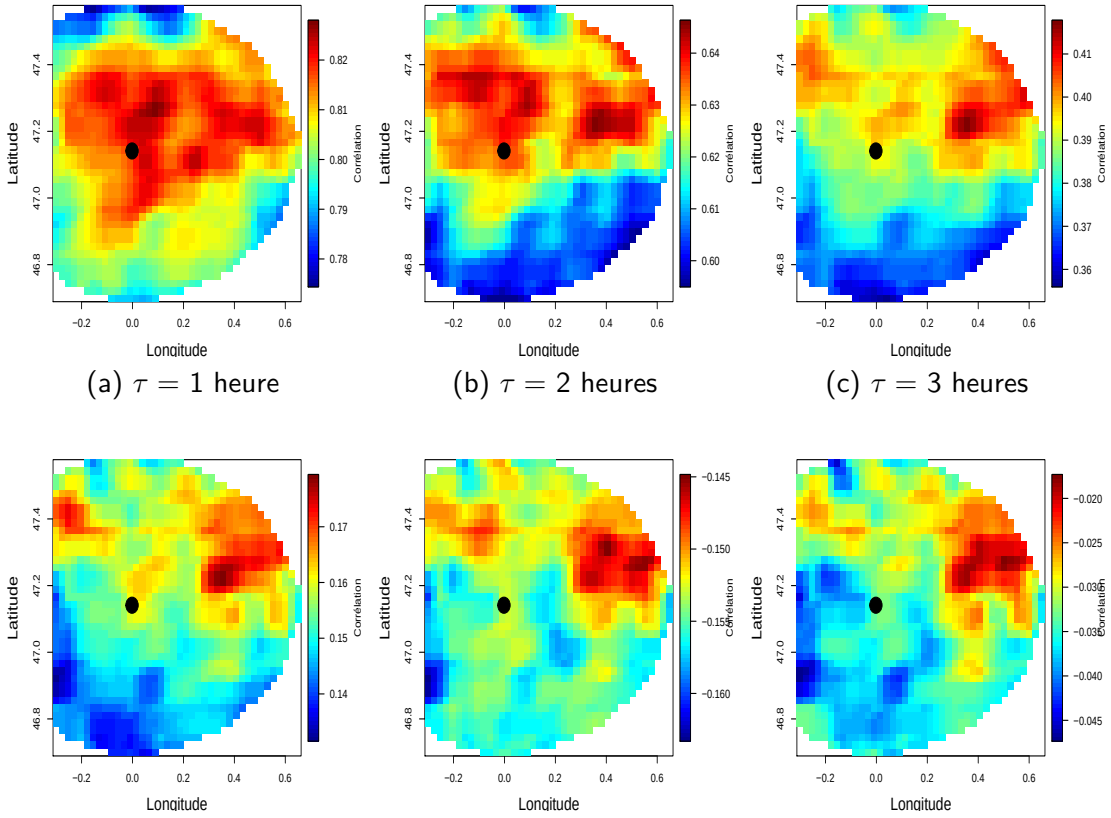
5.3.2 Le modèle de prévision

Le modèle spatio-temporel déterministe avec sélection de variable Lasso a été utilisé pour intégrer les données des images satellites au modèle de prévision. Rappelons que ce modèle est défini par

$$\begin{aligned} P^x_{t+h|t} &= \beta^0_h + \sum_{l=0}^{Ls} \sum_{y \in \mathcal{X}} \beta^{l,y}_h P^y_{t-l} \\ \text{s.c} \quad & \underset{\beta,\gamma}{\operatorname{argmin}} \left\{ \frac{1}{2} RSS(\beta,\gamma) + \lambda \|\beta\|_1 \right\} \end{aligned} \tag{5.2}$$

où $\mathcal{X}$ représente l'ensemble des centrales voisines.

La méthode d'intégration des données des images satellites à ce modèle que nous proposons est l'ajout de ces informations comme variables exogènes dans le modèle de prévision. Cette intégration sous forme de variables exogènes nécessite de choisir quels sont les pixels à intégrer au modèle pour la prévision de la production PV. En effet, la figure 5.3 représente les pixels d'intérêts autour de quelques centrales du jeu de données. La zone de



(a) $\tau = 1$ heure (b) $\tau = 2$ heures (c) $\tau = 3$ heures

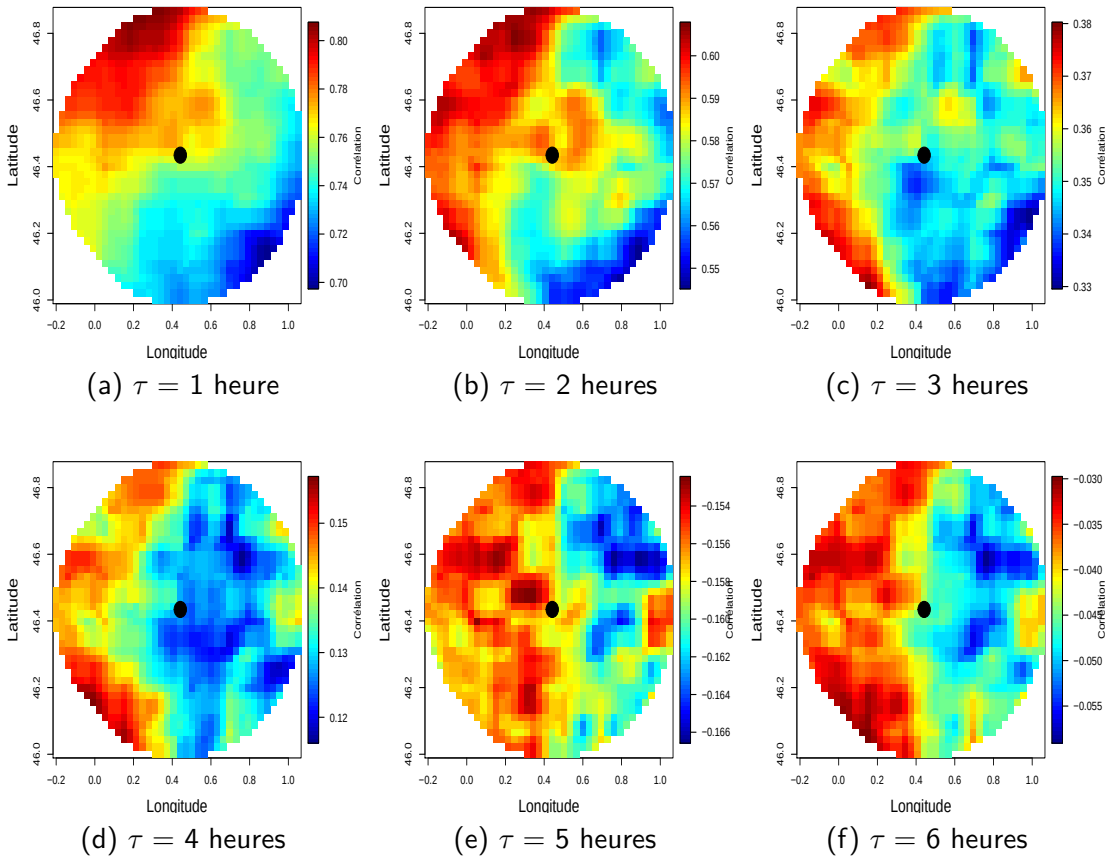
(d) $\tau = 4$ heures (e) $\tau = 5$ heures (f) $\tau = 6$ heures

FIGURE 5.5 – Valeurs du coefficient d'association entre mesure sur site avec des décalages temporels et les estimations d'images satellites pour une centrale à l'ouest de la région couverte. La centrale PV est représentée par le point noir, τ représente le décalage temporel.

50 km définie autour de ces centrales représente un nombre important de pixels susceptible de poser un problème de dimension pour le modèle. Nous proposons donc une intégration des pixels voisins de la position de la centrale qui se fait suivant le même principe que les mesures de centrales voisines, c'est à dire en appliquant la sélection de variables Lasso. Ce choix de méthode d'intégration permet d'éviter des pertes d'informations qui pourraient subvenir dans les cas de choix arbitraires de pixels, ou d'utilisation de la moyenne des estimations pour les pixels autour de la centrale d'intérêt. Le modèle final obtenu est le suivant :

$$\begin{aligned} P^x_{t+h|t} &= \beta^0_h + \sum_{l=0}^{Ls} \sum_{y \in \mathcal{X}} \beta^{l,y}_h P^y_{t-l} + \sum_{k=1}^{p} \sum_{l=0}^{Ls'} \gamma_l Psat^k_{t-l} \\ \text{s.c} \quad & \underset{\beta,\gamma}{\operatorname{argmin}} \left\{ \frac{1}{2} RSS(\beta,\gamma) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\gamma\|_1 \right\}. \end{aligned} \tag{5.3}$$

avec $Psat_t$ les données satellites, Ls' le lag maximal appliqué aux pixels. Les termes de pénalités λ_1 et λ_2 sont associés respectivement aux données de production et aux entrées liées aux images satellites.



(a) $\tau = 1$ heure (b) $\tau = 2$ heures (c) $\tau = 3$ heures

(d) $\tau = 4$ heures (e) $\tau = 5$ heures (f) $\tau = 6$ heures

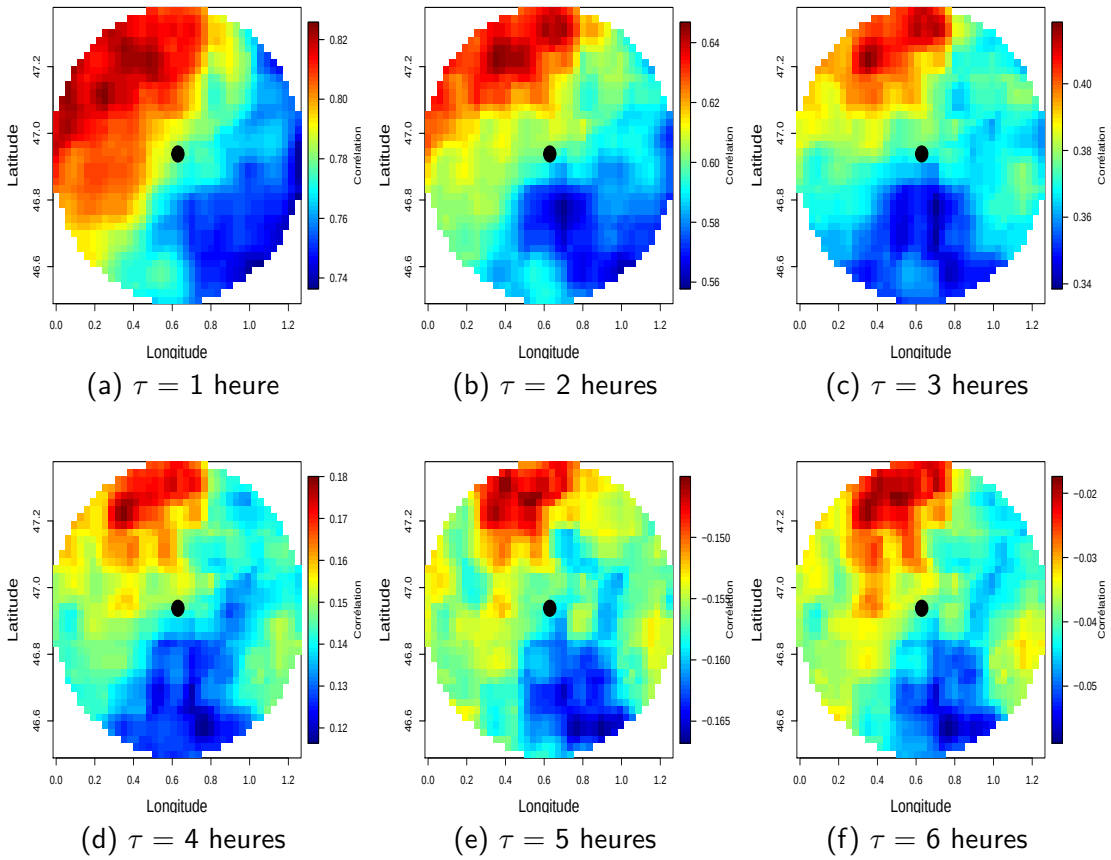
FIGURE 5.6 – Valeurs du coefficient d'association entre mesure sur site avec des décalages temporels et estimation d'images satellites pour une centrale au centre le région couverte. La centrale PV est représentée par le point noir, τ représente le décalage temporel.

5.3.3 Évaluation des performances du modèle

Les performances du modèle spatio-temporel avec images satellites sont comparées avec les performances du modèle spatio-temporel tel que défini par l'équation (5.2). L'horizon maximal de prévision considéré est 6 heures. La première analyse que nous avons menée est d'examiner les variables sélectionnées afin de déterminer la source d'information privilégiée entre images satellites et données de centrales voisines. Nous évaluons ensuite l'apport de l'intégration des images satellites sur les performances de prévision grâce aux critères de RMSE, de MAE et de biais. Les termes de pénalisation sont obtenus par validation croisée [124].

Analyse des variables sélectionnées

La région d'intérêt de l'image satellite retenue autour de chaque centrale PV est de 50 km. Cette région correspond approximativement à l'ensemble des 400 pixels de l'image satellite, les plus proches de la centrale PV. Ces 400 pixels sont intégrés dans le modèle de prévision spatio-temporel. Le nombre initial de variables d'entrée dans le modèle spatio-temporel avec images satellites pour une centrale donnée est donc de 2015 ; ce qui correspond aux pixels avec leurs retards respectifs (400 * 3 (3 heures)) et aux séries de production des centrales voisines avec leur retards respectifs (136 * 6 (nombre de lags) -1). Le tableau 5.1 présente pour une centrale le nombre de variables sélectionnées en



(a) $\tau = 1$ heure (b) $\tau = 2$ heures (c) $\tau = 3$ heures

(d) $\tau = 4$ heures (e) $\tau = 5$ heures (f) $\tau = 6$ heures

FIGURE 5.7 – Valeurs du coefficient d'association entre mesure sur site avec des décalages temporels et estimation d'images satellites pour une centrale située à l'est de la région couverte. La centrale PV est représentée par le point noir, τ représente le décalage temporel.

fonction de l'horizon. Les nombres de pixels et de centrales PV différentes (sans les séries retardées) sélectionnés sont aussi présentés dans le tableau. Le faible nombre de variables sélectionnées montre que la procédure de sélection de variables est efficace pour la réduction de la dimension du problème. De plus, on constate que les variables sélectionnées sont majoritairement des variables liées à la production des sites voisins, suivies des pixels d'images satellites. On pourrait s'attendre à ce que le nombre de pixels choisis augmente avec l'horizon de prévision mais ce n'est pas le cas. Cela peut s'expliquer par le fait que les sites voisins limitrophes concentrent l'essentiel de l'information spatio-temporelle. La sélection de pixels supplémentaires n'apporte donc pas de valeur ajoutée à la modélisation.

Tableau 5.1 – Nombre de pixels et de centrales sélectionnés par horizon pour une centrale PV

Horizons	Nombre initial de variables = 1351		
	Nbre de pixels sélectionnés	Nbre de centrales PV sélectionnées	Nbre total de variables sélectionnées (lag et pixels inclus)
15 min	4	28	67
1 h	4	22	64
3 h	7	25	56

Apport des images satellites sur les performances de prévision

Les tableaux 5.2 et 5.3 présentent pour deux centrales les critères du RMSE et du MAE en fonction de l'horizon. L'analyse de ces valeurs de critères montre que le modèle avec les images satellites présente des valeurs plus faibles pour l'ensemble des horizons. On constate une amélioration moyenne du MAE de 10% et du RMSE de 3% due à l'ajout des images satellites. L'analyse est la même pour les autres centrales du jeu de données.

Tableau 5.2 – Comparaison des RMSE des modèles spatio-temporel intégrant les images satellites et spatio-temporel sans images satellites pour deux centrales du jeu de données.

Les valeurs sont normalisées par la puissance maximale observée. "ST" le modèle spatio-temporel sans images satellites et "ST + SAT" celui avec les images satellites.

Horizons	P_1			P_2		
	RMSE (%P_{max})		Amélioration	RMSE (%P_{max})		Amélioration
(heure)	ST	ST + SAT	(%)	ST	ST + SAT	(%)
0.25	6.12	6.01	1.80	6.24	6.13	1.76
0.50	7.80	7.65	1.92	8.08	7.93	1.86
0.75	8.86	8.70	1.81	9.13	8.97	1.75
1.00	9.92	9.74	1.81	10.15	9.97	1.77
2.00	13.13	12.89	1.83	13.61	13.36	1.84
3.00	15.53	15.23	1.93	16.03	15.72	1.93
4.00	16.97	16.64	1.94	17.48	17.15	1.89
5.00	17.84	17.46	2.13	18.38	17.99	2.12
6.00	18.46	17.90	3.03	18.99	18.41	3.05

Tableau 5.3 – Comparaison des MAE des modèles spatio-temporel intégrant les images satellites et spatio-temporel sans images satellites pour deux centrales du jeu de données.

Les valeurs sont normalisées par la puissance maximale observée. "ST" le modèle spatio-temporel sans images satellites et "ST + SAT" celui avec les images satellites.

Horizons	P_1			P_2		
	MAE (%P_{max})		Amélioration	MAE (%P_{max})		Amélioration
(heure)	ST	ST + SAT	(%)	ST	ST + SAT	(%)
0.25	2.62	2.53	3.44	2.40	2.31	3.75
0.50	3.75	3.55	5.33	3.44	3.25	5.52
0.75	4.52	4.22	6.64	4.22	3.93	6.87
1.00	5.23	4.82	7.84	4.95	4.56	7.88
2.00	8.07	7.11	11.90	7.57	6.67	11.89
3.00	10.21	8.66	15.18	9.60	8.14	15.21
4.00	11.52	9.52	17.36	10.92	9.02	17.40
5.00	12.30	9.99	18.78	11.66	9.47	18.78
6.00	12.59	10.14	19.46	12.00	9.66	19.50

5.4 Analyse comparative des performances des modèles proposés

Dans cette partie, nous proposons une analyse comparative des performances des différents modèles proposés dans le cadre de cette thèse. La figure 5.8 présente une schématisation des différents modèles implémentés.

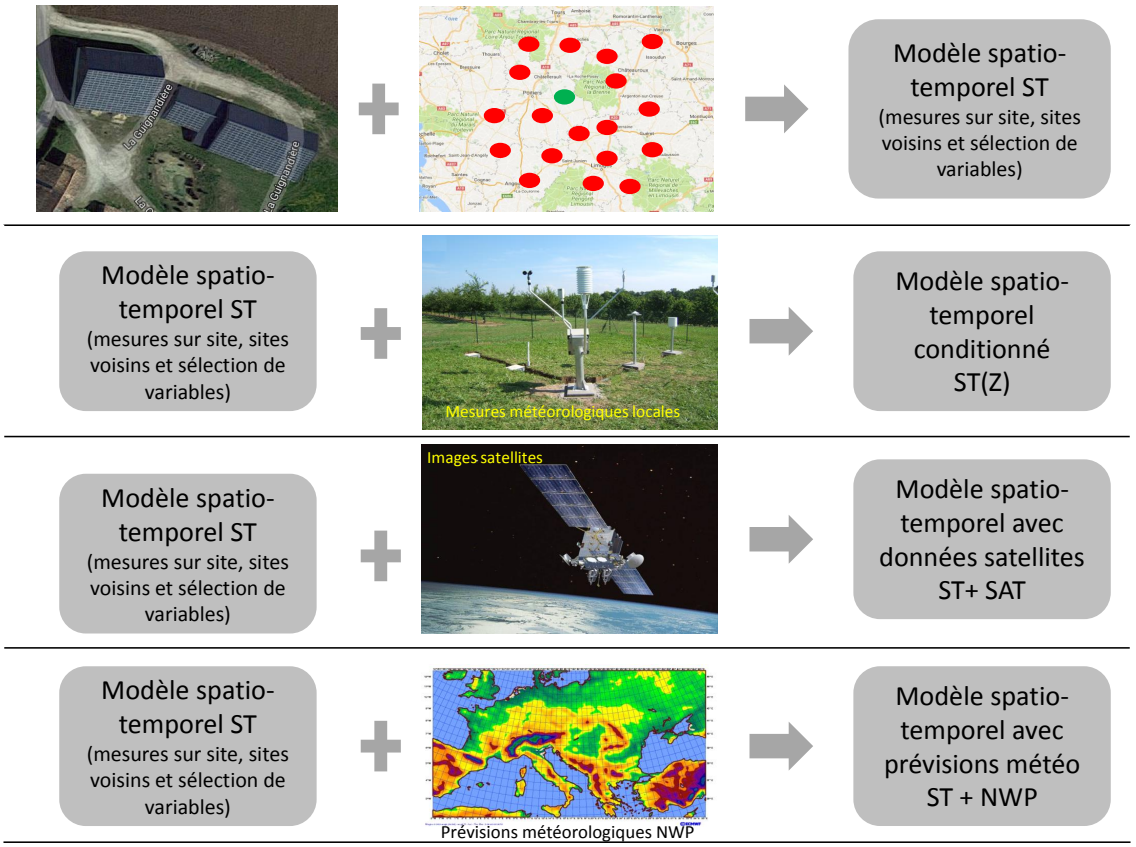
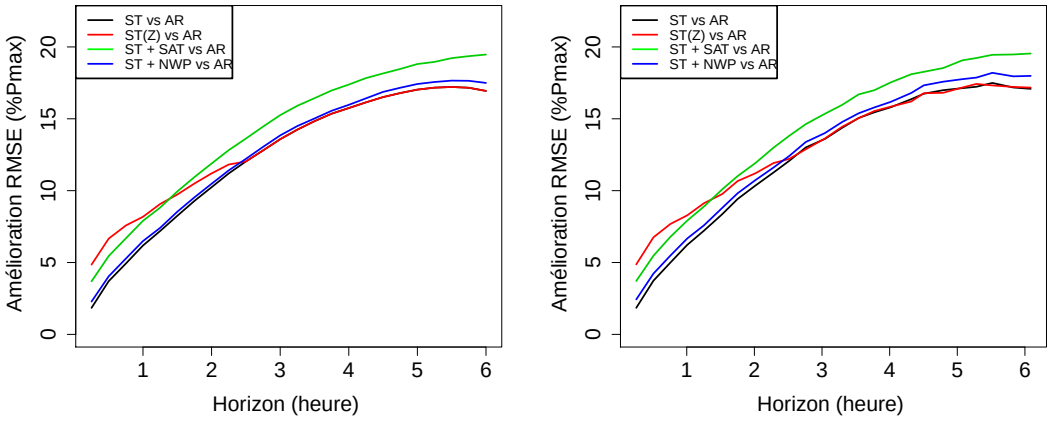


FIGURE 5.8 – Présentation des différents modèles comparés dans le but de déterminer l'apport de chaque source de données sur la qualité des prévisions.

Les modèles comparés

Nous proposons ici une approche incrémentale qui vise à quantifier l'apport de chaque source de données en termes de performance de prévision. Le modèle de référence est le modèle autorégressif AR (équation (3.1)) qui est un modèle exploitant uniquement les dépendances temporelles. La première comparaison est faite entre ce modèle de référence et le modèle spatio-temporel. La première source de données dont l'impact est évalué est donc l'ensemble des mesures des centrales voisines de la centrale d'intérêt. Le modèle ainsi défini est nommé modèle spatio-temporel ST. Dans ce modèle, la procédure de sélection de variables proposée au chapitre 3 est intégrée, nous assurant ainsi le traitement des problèmes de parcimonie et de dimension. Les données météorologiques locales (autour du site d'intérêt) constituent la deuxième source de données dont nous avons évalué l'impact par rapport au modèle de référence. Ce modèle est nommé modèle spatio-temporel conditionné ST(Z). Les deux autres sources de données dont l'impact sur la qualité des prévisions a été évalué sont les images satellites et les prévisions NWP. Ces deux sources de données ont permis de construire deux modèles respectivement nommés modèle ST +



(a) P_1 (b) P_2

FIGURE 5.9 – Comparaison des performances des différents modèles avancés par rapport au modèle AR en fonction de l'horizon pour deux centrales du jeu de données. Le pas de temps est 15 min.

SAT pour les images satellites et modèle ST + NWP pour les prévisions NWP. L'horizon de prévision envisagé est 6 heures et les performances de ces modèles sont évaluées par rapport au modèle de référence pour différents critères d'évaluation. Le jeu de données d_2 a été utilisé avec 10 mois pour l'apprentissage et 5 mois pour le testing. La variable météorologique locale utilisée ici est la vitesse du vent. Pour chaque centrale PV, les mesures de vitesse de vent sont obtenues grâce à une station de mesure à proximité.

Résultats de l'analyse comparative

La figure 5.9 présente les améliorations en terme de RMSE des différents modèles spatio-temporels par rapport au modèle de référence pour deux centrales du jeu de données. Le modèle spatio-temporel permet une amélioration moyenne de RMSE de 10% pour 3 heures. Cette amélioration peut atteindre les 20% en fonction des centrales. Ce résultat est concordant avec ceux présentés au chapitre 3. L'utilisation des mesures de vitesse de vent locales grâce à des stations météorologiques à proximité des centrales permet d'améliorer de 2% en moyenne les performances de prévision pour les deux premières heures de prévision. Au-delà de 3 heures, ces mesures ne contribuent pas à améliorer davantage les performances de prévision par rapport au modèle spatio-temporel de base. Le modèle spatio-temporel qui intègre les prévisions NWP ne présente pas d'amélioration significative par rapport au modèle spatio-temporel initial sur l'ensemble des 6 heures de prévision. Il faut toutefois noter une légère amélioration des performances de ce modèle pour des valeurs d'horizon à partir de 5 heures. L'intégration des images satellites permet de réduire davantage les erreurs de prévision. En effet, on constate sur la figure une amélioration du RMSE de l'ordre de 3% en moyenne du modèle avec images satellites intégrées par rapport au modèle spatio-temporel simple.

5.5 Conclusion

Dans ce chapitre, nous avons proposé un modèle spatio-temporel qui exploite non seulement les informations spatio-temporelles transmises par les mesures de production des sites voisins, mais aussi les images satellites. Ces dernières étant caractérisées par des

résolutions plus fines et des vitesses des mises à jour plus fréquentes que les prévisions NWP, elles constituent une source de données très intéressante pour la prévision PV. Nous avons présenté une procédure de sélection des pixels autour des centrales pour lesquelles la prévision de production PV est envisagée. Cette procédure permet de passer d'images recouvrant l'ensemble des centrales considérées à une image plus fine qui se concentre autour de la centrale d'intérêt. La relation entre les points d'intérêts de la zone retenue autour des centrales et la production du site a permis de voir que les pixels les plus proches de l'images sont les plus corrélés à la production. Les informations des images satellites ont donc été intégrées au modèle spatio-temporel avec la méthode de sélection de variables définie au chapitre 3. Cette intégration a été faite sous forme de variables exogènes supplémentaires en adaptant aussi la régularisation aux pixels de l'image. Le modèle spatio-temporel proposé procède donc au choix optimal des variables de production mais aussi des pixels de l'image satellite. Les variables sélectionnées par le modèle sont en majorité des variables de production complétées de quelques pixels retenus. Les pixels retenus permettent d'apporter des informations supplémentaires au modèle entrainant une amélioration des performances de prévision.

Nous avons dans la deuxième partie de ce chapitre, quantifié l'apport de chacune des différentes sources d'informations à savoir les images satellites, les mesures de centrales voisines, les prévisions NWP et les mesures météorologiques locales sur les performances de prévision en comparaison avec un modèle de référence exclusivement temporel. Les horizons de prévisions envisagés sont de 6 heures. La plus grande source d'amélioration provient de l'utilisation des mesures de centrales. Les images satellites permettent de réduire davantage les erreurs de prévision lorsqu'elles sont associées aux modèles spatio-temporels. L'effet des prévisions NWP est très faible sur les premiers horizons en opposition avec celui des mesures météorologiques locales. Les prévisions NWP permettent cependant de corriger les mauvaises performances du modèle spatio-temporel pour les horizons supérieurs à 12 heures confirmant ainsi l'importance de la météorologie pour ces horizons de prévision.

Chapitre 6

Conclusions et perspectives

6.1 Conclusions générales

La prévision à court-terme de la production photovoltaïque est nécessaire pour la maitrise de la variabilité due à l'intermittence de la production, le maintien de l'équilibre entre l'offre et la demande en électricité et la gestion de la réserve. Elle est aussi très utile pour les producteurs aussi bien pour la participation aux marchés de l'électricité que la planification des opérations de maintenance et la gestion de la production. Il existe dans la littérature de nombreuses méthodes de prévision de la production PV pour différents horizons. Cette thèse porte sur la prévision à court-terme ($\leq$ 6 h) de la production photovoltaïque basée sur l'exploitation des corrélations spatio-temporelles entre les différents sites de production. L'objectif principal de cette thèse était d'améliorer la prévision de la production d'un site donné en utilisant les sites de production voisins comme des sources supplémentaires d'informations.

La première partie de ce travail a été de démontrer l'existence de corrélations spatio-temporelles entre les séries de mesures de sites de production géographiquement distribués. Pour cela, une étude de la variabilité de la production PV a été menée dans la première partie du chapitre 2. Les deux principales origines de la variabilité de la production PV intra-journalière sont la course du soleil et les perturbations météorologiques. La mise en évidence de corrélations spatio-temporelles entre les séries de production de différents sites passe par l'extraction de la variabilité due à la course du soleil. En effet, une analyse directe de corrélations entre les séries brutes de différentes centrales ne serait qu'une mise en évidence d'une propagation est-ouest due à la course du soleil. De plus, les méthodes de traitement des séries de production PV proposées dans la littérature présentent des limites sur les faibles productions et ne sont pas bien documentées. Nous avons donc proposé dans la deuxième partie du chapitre 2, une nouvelle méthode de stationnarisation qui a pour but d'extraire la variabilité due au mouvement du soleil des séries de production PV. Cette méthode de stationnarisation que nous avons proposé s'inscrit dans la continuité de celle de la normalisation des séries d'irradiation par des modèles qui font l'hypothèse de ciel-clair. Toutefois la méthode que nous avons proposée ici contrairement à d'autres, ne nécessite pas une modélisation spécifique des phénomènes qui ont lieu dans l'atmosphère. La méthode de stationnarisation proposée a été présentée dans le premier article publié qui est fourni en annexe. Les performances de la méthode de stationnarisation ont été évaluées au sens des critères statistiques de stationnarité mais aussi des performances en prévision des nouvelles séries. Les séries stationnarisées ainsi construites ont été ensuite utilisées pour calculer les corrélations spatiales et temporelles entre les séries de production PV de différents sites. Cette étude de corrélation basée sur les séries stationnarisées a mis

en évidence des corrélations significatives entre les productions de différents sites pour des horizons temporels de l'ordre de 6 heures et des résolutions spatiales de moins de 250 km. Ces corrélations traduisent des propagations d'informations relatives aux perturbations météorologiques (l'effet de la course du soleil ayant été enlevé).

Les corrélations spatio-temporelles ayant été mises en évidence, la deuxième partie de notre travail a été de proposer des modèles de prévision qui exploitent ces corrélations. Le modèle spatio-temporel proposé pour la prévision déterministe de la production PV a été présenté au chapitre 3. Des modèles qui n'exploitent que les informations temporelles ont aussi été définis et utilisés comme référence afin d'évaluer l'apport de l'utilisation des informations spatio-temporelles. Le modèle spatio-temporel proposé est inspiré du modèle vectoriel autorégressif. Les entrées identifiées pour ce modèle pour exploiter les informations spatio-temporelles sont les mesures de production sur le site d'intérêt, les séries de production des sites voisins, les séries retardées associées à chacun de ces sites. Deux extensions du modèle spatio-temporel ont été proposées dans la deuxième partie du chapitre 3. La première a été d'intégrer au modèle de prévision spatio-temporel des informations sur les conditions météorologiques au voisinage de la centrale d'intérêt. Cette intégration a été faire par conditionnement des coefficients du modèle à une variable décrivant les conditions météorologiques. Les coefficients du modèle spatio-temporel sont donc calculés suivant les valeurs de la variable météorologique locale retenue. La deuxième amélioration concerne la dimension du modèle et sa parcimonie. En effet, un nombre important de centrales voisines entraîne une augmentation significative du nombre de variables en entrée du modèle. La deuxième extension du modèle spatio-temporel permet d'apporter une solution à ce problème par une méthode de sélection de variables. La régularisation Lasso est la méthode de sélection de variables que nous proposons. Elle est directement intégrée à la procédure d'estimation des coefficients du modèle spatio-temporel. Les critères d'évaluation des prévisions déterministes ont permis de comparer les performances des différents modèles proposés par rapport aux références pour des horizons allant jusqu'à 6 heures. Il en ressort que le modèle spatio-temporel permet une amélioration moyenne de 20% du critère du RMSE par rapport au modèle de référence autorégressif. De plus, le conditionnement par la vitesse du vent et la sélection de variables Lasso permettent aussi de réduire les erreurs de prévision pour l'ensemble des critères évalués. Dans la dernière partie du chapitre 3, nous avons évalué pour des horizons allant jusqu'à 12 heures, l'apport de l'intégration des données issues de modèles numériques de prévision de type NWP sur les performances de prévision. Cette analyse a montré que pour des horizons inférieurs à 6 heures, le modèle spatio-temporel avec uniquement les données de production est plus précis que celui avec les prévisions météorologiques NWP. De plus, pour ces horizons, les prévisions NWP n'apportent pas d'améliorations importantes (1% en moyenne). Pour les horizons supérieurs à 6 heures, les performances du modèle spatio-temporel sans les données NWP se dégradent significativement. Cette dégradation est corrigée par l'intégration des données NWP. Cela traduit l'importance de l'utilisation des données NWP pour les prévisions de production PV pour les horizons de plus de 6 heures. Le modèle spatio-temporel déterministe proposé a fait l'objet d'un article publié qui est présenté dans l'annexe A.

La troisième partie de notre travail a été de proposer un modèle de prévision qui exploite les informations spatio-temporelles pour fournir des prévisions probabilistes de la production PV. Les prévisions probabilistes permettent d'avoir plus d'informations sur la distribution future de la production PV et ainsi de pouvoir quantifier les erreurs associées à ces prévisions. Nous proposons dans le chapitre 4 un modèle de prévision probabiliste basé sur la régression quantile et la pénalisation Lasso. Ce modèle exploite

les corrélations spatio-temporelles entre les données de production mais aussi les prévisions météorologiques. Ce modèle permet de fournir des prévisions des quantiles conditionnels de la distribution future de la production PV pour des horizons de 6 heures. Ce modèle est comparé avec le modèle d'estimation par noyau qui est couramment utilisé dans la littérature. L'évaluation des performances avec les critères de prévision probabiliste a mis en évidence de meilleures performances de prévision avec le modèle spatio-temporel probabiliste proposé.

La dernière partie de notre travail a été d'utiliser les images satellites comme nouvelle source de données. Nous avons proposé dans le chapitre 5 une méthode d'intégration des images satellites dans un modèle spatio-temporel de prévision. Cette méthode permet de sélectionner les pixels les plus pertinents pour la prévision de la production d'une centrale et de les intégrer à un modèle qui exploite à la fois ces pixels et les mesures des centrales voisines. L'utilisation des images satellites permet d'améliorer la qualité des prévisions de production PV de 10% en moyenne absolue. Dans la dernière partie du chapitre 5 une analyse comparative de l'apport des différentes sources de données en fonction de l'horizon de prévision a été menée.

En conclusion générale nous pouvons dire qu'avec les méthodes spatio-temporelles, la prédictibilité à court-terme de la production PV peut être améliorée de façon significative, c'est-à-dire jusqu'à 20% par rapport à une méthode classique. Les prévisions probabilistes proposées ont de bonnes propriétés. En plus, le problème de dimension a été traité afin de pouvoir prendre en compte de grandes quantités de données.

6.2 Perspectives

Les travaux réalisés dans le cadre de cette thèse ont permis d'identifier d'autres pistes d'études. Une première piste d'étude pourrait être de proposer un modèle de prévision probabiliste "intelligent" au sens du choix des variables d'entrée. Un tel modèle serait capable de choisir entre les mesures sur sites, les mesures de centrales voisines, les prévisions NWP, les mesures de station météo et les images satellites les données adéquates en fonction de l'horizon pour fournir des prévisions de la distribution future de la production PV. Ce choix serait fait de façon automatique à l'intérieur du modèle par des méthodes de sélection de variables appropriées éliminant ainsi non seulement la possibilité de perte d'informations (par un choix subjectif/aléatoire des sources de données utilisées) mais aussi les problèmes de dimension du modèle. La mise en œuvre d'un tel modèle de prévision probabiliste avec un accent particulier sur les queues de distribution de la production pourrait être une application intéressante dans le cas de la fourniture de services systèmes au réseau.

Une perspective complémentaire pour un tel modèle serait de fournir des prévisions de production PV pour des horizons plus longs pouvant atteindre une journée. Du point de vue spécifique de l'utilisation des images satellites, plusieurs pistes peuvent être étudiées. Une première serait d'identifier les nuages susceptibles d'affecter la production PV et d'extraire des images les informations de vitesse et direction de ces nuages. Ces informations seraient ensuite intégrées dans un modèle global avec choix automatique des informations pertinentes. Une seconde piste d'amélioration liée aux images satellites serait d'évaluer les incertitudes associées aux informations extraites des images satellites. De plus, la part de ces incertitudes dans les erreurs de prévision pourrait être évaluée.

Bibliographie

[1] Jan Kleissl. *Solar energy forecasting and resource assessment.* Academic Press, 2013.

[2] Rich H. Inman, Hugo T. C. Pedro, and Carlos F. M. Coimbra. Solar forecasting methods for renewable energy integration. *Progress in Energy and Combustion Science*, 39(6) :535–576, December 2013.

[3] International energy agency. World energy statistics.

[4] United Nations. Convention-cadre sur les changements climatiques.

[5] International energy agency. Renewables energy statistics.

[6] LOI n 2015-992 du 17 août 2015 relative à la transition énergétique pour la croissance verte, August 2015.

[7] Direction générale de l'énergie et du climat. Programmation pluriannuelle de l'énergie période 2009-2020.

[8] Cameron W. Potter, Allison Archambault, and Kenneth Westrick. Building a Smarter Smart Grid Through Better Renewable Energy Information. In *Proceedings of Power Systems Conference and Exposition*, Seattle, WA, USA, March 2009.

[9] P. Pinson, C. Chevallier, and G.N. Kariniotakis. Trading Wind Generation From Short-Term Probabilistic Forecasts of Wind Power. *Power Systems, IEEE Transactions on*, 22(3) :1148–1156, August 2007.

[10] Georges Kariniotakis. *Renewable Energy Forecasting : From Models to Applications.* Woodhead Publishing, June 2017. Google-Books-ID : 6y_ZCgAAQBAJ.

[11] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de Pison, and F. Antonanzas-Torres. Review of photovoltaic power forecasting. *Solar Energy*, 136 :78–111, October 2016.

[12] Hugo T. C. Pedro and Carlos F. M. Coimbra. Assessment of forecasting techniques for solar power production with no exogenous inputs. *Solar Energy*, 86(7) :2017–2028, July 2012.

[13] V. Kostylev, A. Pavlovski, and others. Solar power forecasting performance–towards industry standards. In *1st International Workshop on the Integration of Solar Power into Power Systems, Aarhus, Denmark*, 2011.

[14] Simone Sperati, Stefano Alessandrini, Pierre Pinson, and George Kariniotakis. The weather intelligence for renewable energies. benchmarking exercise on short-term forecasting of wind and solar power generation. *Energies*, 8(9) :9594–9619, September 2015.

[15] M. Zamo, O. Mestre, P. Arbogast, and O. Pannekoucke. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I : Deterministic forecast of hourly production. *Solar Energy*, 105 :792–803, July 2014.

[16] M. Zamo, O. Mestre, P. Arbogast, and O. Pannekoucke. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II : Probabilistic forecast of daily production. *Solar Energy*, 105 :804–816, July 2014.

[17] Frederick G. Shuman. Numerical Weather Prediction. *Bulletin of the American Meteorological Society*, 59(1) :5–17, January 1978.

[18] Sophie Pelland, George Galanis, and George Kallos. Solar and photovoltaic forecasting through post-processing of the Global Environmental Multiscale numerical weather prediction model. *Progress in Photovoltaics : Research and Applications*, 21(3) :284–296, May 2013.

[19] Richard Perez, Sergey Kivalov, James Schlemmer, Karl Hemker Jr., David Renné, and Thomas E. Hoff. Validation of short and medium term operational solar radiation forecasts in the US. *Solar Energy*, 84(12) :2161–2172, December 2010.

[20] Richard Perez, Kathleen Moore, Steve Wilcox, David Renné, and Antoine Zelenka. Forecasting solar radiation – Preliminary evaluation of an approach based upon the national forecast database. *Solar Energy*, 81(6) :809–812, June 2007.

[21] Pierre Ineichen. Comparison of eight clear sky broadband models against 16 independent data banks. *Solar Energy*, 80(4) :468–478, April 2006.

[22] Christelle Rigollier, Olivier Bauer, and Lucien Wald. On the clear sky model of the esra — european solar radiation atlas — with respect to the heliosat method. *Solar Energy*, 68(1) :33–48, January 2000.

[23] R. E. Bird and R. L. Hulstrom. Simplified clear sky model for direct and diffuse insolation horizontal surfaces. Technical report, February 1981.

[24] Benjamin Y. H. Liu and Richard C. Jordan. The interrelationship and characteristic distribution of direct, diffuse and total solar radiation. *Solar Energy*, 4 :1–19, July 1960.

[25] J. N. Black, C. W. Bonython, and J. A. Prescott. Solar radiation and the duration of sunshine. *Quarterly Journal of the Royal Meteorological Society*, 80(344) :231–235, April 1954.

[26] Peder Bacher, Henrik Madsen, and Henrik Aalborg Nielsen. Online short-term solar power forecasting. *Solar Energy*, 83(10) :1772–1783, October 2009.

[27] N. A. Engerer and F. P. Mills. KPV : A clear-sky index for photovoltaics. *Solar Energy*, 105 :679–693, July 2014.

[28] James Douglas Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.

[29] E. Lorenz, J. Hurka, D. Heinemann, and H. G. Beyer. Irradiance Forecasting for the Power Prediction of Grid-Connected Photovoltaic Systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2(1) :2–10, March 2009.

[30] Peder Bacher, Henrik Madsen, Bengt Perers, and Henrik Aalborg Nielsen. A non-parametric method for correction of global radiation observations. *Solar Energy*, 88 :13 – 22, 2013.

[31] Claudio Monteiro, Tiago Santos, L. Alfredo Fernandez-Jimenez, Ignacio J. Ramirez-Rosado, and M. Sonia Terreros-Olarte. Short-Term Power Forecasting Model for Photovoltaic Plants Based on Historical Similarity. *Energies*, 6(5), May 2013.

[32] V. Berdugo, Christophe Chaussin, Laurent Dubus, Georges Hebrail, and Viviane Leboucher. Analog method for collaborative very-short-term forecasting of power generation from photovoltaic systems. *Next Generation Data Mining Summit (NGDM'11)*, 2011.

[33] Elke Lorenz, Detlev Heinemann, and Christian Kurz. Local and regional photovoltaic power prediction for large scale grid integration : Assessment of a new algorithm for snow detection. *Progress in Photovoltaics : Research and Applications*, 20(6) :760–769, September 2012.

[34] Marco Pierro, Francesco Bucci, Matteo De Felice, Enrico Maggioni, David Moser, Alessandro Perotto, Francesco Spada, and Cristina Cornaro. Multi-Model Ensemble for day ahead prediction of photovoltaic power generation. *Solar Energy*, 134 :132–146, September 2016.

[35] M. G. De Giorgi, P. M. Congedo, and M. Malvoni. Photovoltaic power forecasting using statistical methods : impact of weather data. *IET Science, Measurement Technology*, 8(3) :90–97, May 2014.

[36] Simon Haykin. *Neural Networks : A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 1998.

[37] A. Yona, T. Senjyu, A. Y. Saber, T. Funabashi, H. Sekine, and C. H. Kim. Application of Neural Network to One-Day-Ahead 24 hours Generating Power Forecasting for Photovoltaic System. In *International Conference on Intelligent Systems Applications to Power Systems, 2007. ISAP 2007*, pages 1–6, November 2007.

[38] C. Tao, D. Shanxu, and C. Changsong. Forecasting power output for grid-connected photovoltaic power system without using solar radiation measurement. In *The 2nd International Symposium on Power Electronics for Distributed Generation Systems*, pages 773–777, June 2010.

[39] Alberto Dolara, Francesco Grimaccia, Sonia Leva, Marco Mussetta, and Emanuele Ogliari. A physical hybrid artificial neural network for short term forecasting of pv plant power output. *Energies*, 8(2), February 2015.

[40] Y. Huang, J. Lu, C. Liu, X. Xu, W. Wang, and X. Zhou. Comparative study of power forecasting methods for PV stations. In *2010 International Conference on Power System Technology (POWERCON)*, pages 1–6, October 2010.

[41] L. Alfredo Fernandez-Jimenez, Andrés Muñoz-Jimenez, Alberto Falces, Montserrat Mendoza-Villena, Eduardo Garcia-Garrido, Pedro M. Lara-Santillan, Enrique Zorzano-Alba, and Pedro J. Zorzano-Santamaria. Short-term power forecasting system for photovoltaic plants. *Renewable Energy*, 44 :311–317, August 2012.

[42] Adel Mellit and Alessandro Massi Pavan. A 24-h forecast of solar irradiance using artificial neural network : Application for performance prediction of a grid-connected PV plant at Trieste, Italy. *Solar Energy*, 84(5) :807–821, May 2010.

[43] A. Mellit, A. Massi Pavan, and V. Lughi. Short-term forecasting of power production in a large-scale photovoltaic plant. *Solar Energy*, 105 :401–413, July 2014.

[44] Stylianos I. Vagropoulos, Evaggelos G. Kardakos, Christos K. Simoglou, Anastasios G. Bakirtzis, and João P. S. Catalão. ANN-based scenario generation methodology for stochastic variables of electric power systems. *Electric Power Systems Research*, 134 :9–18, May 2016.

[45] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines : And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA, 2000.

[46] Joao Gari da Silva Fonseca, Takashi Oozeki, Takumi Takashima, Gentarou Koshimizu, Yoshihisa Uchida, and Kazuhiko Ogimoto. Use of support vector regression and numerically predicted cloudiness to forecast power output of a photovoltaic power plant in Kitakyushu, Japan. *Progress in Photovoltaics : Research and Applications*, 20(7) :874–882, November 2012.

[47] Navin Sharma, Pranshu Sharma, David Irwin, and Prashant Shenoy. Predicting Solar Generation from Weather Forecasts Using Machine Learning. In *Proceedings of IEEE International Conference on Smart Grid Communications, (IEEE SmartGridComm)*, Brussels, Belgium, October 2011.

[48] O. Perpiñán and E. Lorenzo. Analysis and synthesis of the variability of irradiance and PV power time series with the wavelet transform. *Solar Energy*, 85(1) :188–197, January 2011.

[49] Jie Shi, Wei-Jen Lee, Yongqian Liu, Yongping Yang, and Peng Wang. Forecasting power output of photovoltaic system based on weather classification and support vector machine. In *2011 IEEE Industry Applications Society Annual Meeting (IAS)*, pages 1–6, October 2011.

[50] Matteo De Felice, Marcello Petitta, and Paolo M. Ruti. Short-term predictability of photovoltaic production over Italy. *Renewable Energy*, 80 :197–204, August 2015.

[51] M. Bouzerdoum, A. Mellit, and A. Massi Pavan. A hybrid model (SARIMA–SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant. *Solar Energy*, 98, Part C :226–235, December 2013.

[52] Jiaming Li, John K. Ward, Jingnan Tong, Lyle Collins, and Glenn Platt. Machine learning for solar irradiance forecasting of photovoltaic system. *Renewable Energy*, 90 :542–553, May 2016.

[53] Jing Huang and Matthew Perry. A semi-empirical approach using gradient boosting and -nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting. *International Journal of Forecasting*, 32(3) :1081–1086, July 2016.

[54] Qingyu Yang, Dou An, and Yuanli Cai. A Novel Evolution Kalman Filter Algorithm for Short-Term Climate Prediction. *Asian Journal of Control*, 18(1) :400–405, January 2016.

[55] Roger Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[56] James W. Taylor and Jooyoung Jeon. Forecasting wind power quantiles using conditional kernel estimation. *Renewable Energy*, 80 :370–379, August 2015.

[57] F. Golestaneh, P. Pinson, and H. B. Gooi. Very Short-Term Nonparametric Probabilistic Forecasting of Renewable Energy Generation #x2014 ; With Application to Solar Energy. *IEEE Transactions on Power Systems*, 31(5) :3850–3863, September 2016.

[58] Enrica Scolari, Fabrizio Sossan, and Mario Paolone. Irradiance prediction intervals for PV stochastic generation in microgrid applications. *Solar Energy*, 139 :116–129, December 2016.

[59] Yaoyao He, Qifa Xu, Jinhong Wan, and Shanlin Yang. Short-term power load probability density forecasting based on quantile regression neural network and triangle kernel function. *Energy*, 114 :498–512, November 2016.

[60] S. Alessandrini, L. Delle Monache, S. Sperati, and G. Cervone. An analog ensemble for short-term probabilistic solar power forecast. *Applied Energy*, 157 :95–110, November 2015.

[61] Gábor I. Nagy, Gergő Barta, Sándor Kazi, Gyula Borbély, and Gábor Simon. GEFCom2014 : Probabilistic solar and wind power forecasting using a generalized additive tree ensemble approach. *International Journal of Forecasting*, 32(3) :1087–1093, July 2016.

[62] Y. Chu and C.F.M. Coimbra. Short-term probabilistic forecasts for Direct Normal Irradiance. *Renewable Energy*, 101 :526–536, 2017.

[63] Simone Sperati, Stefano Alessandrini, and Luca Delle Monache. An application of the ECMWF Ensemble Prediction System for short-term solar power forecasting. *Solar Energy*, 133 :437–450, August 2016.

[64] Yuanyuan Liu, Susumu Shimada, Jun Yoshino, Tomonao Kobayashi, Yasushi Miwa, and Kiyotaka Furuta. Ensemble forecasting of solar irradiance by applying a mesoscale meteorological model. *Solar Energy*, 136 :597–605, October 2016.

[65] M. David, F. Ramahatana, P. J. Trombe, and P. Lauret. Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models. *Solar Energy*, 133 :55–72, August 2016.

[66] Adrian Grantham, Yulia R. Gel, and John Boland. Nonparametric short-term probabilistic forecasting for solar radiation. *Solar Energy*, 133 :465–475, August 2016.

[67] Jie Zhang, Bri-Mathias Hodge, and Anthony Florita. Joint Probability Distribution and Correlation Analysis of Wind and Solar Power Forecast Errors in the Western Interconnection. *Journal of Energy Engineering*, 141(1) :B4014008, 2015.

[68] G. Tina and S. Gagliano. Probabilistic analysis of weather data for a hybrid solar/wind energy system. *International Journal of Energy Research*, 35(3) :221–232, March 2011.

[69] M. J. Sanjari and H. B. Gooi. Probabilistic Forecast of PV Power Generation Based on Higher-Order Markov Chain. *IEEE Transactions on Power Systems*, PP(99) :1–1, 2016.

[70] Emil B. Iversen, Juan M. Morales, Jan K. Møller, and Henrik Madsen. Short-term probabilistic forecasting of wind speed using stochastic differential equations. *International Journal of Forecasting*, 32(3) :981–990, July 2016.

[71] Emil B. Iversen, Juan M. Morales, Jan K. Møller, and Henrik Madsen. Probabilistic Forecasts of Solar Irradiance by Stochastic Differential Equations. *arXiv :1310.6904 [stat]*, October 2013.

[72] Cihan H. Dagli, ., Paras Mandal, Surya Teja Swarroop Madhira, Ashraf Ul haque, Julian Meng, and Ricardo L. Pineda. Forecasting power output of solar photovoltaic system using wavelet transform and artificial intelligence techniques. *Procedia Computer Science*, 12 :332–337, January 2012.

[73] A. Bracale, P. Caramia, G. Carpinelli, A. R. Di Fazio, and P. Varilone. A Bayesian-Based Approach for a Short-Term Steady-State Forecast of a Smart Grid. *IEEE Transactions on Smart Grid*, 4(4) :1760–1771, December 2013.

[74] Antonio Bracale, Pierluigi Caramia, and Guido Carpinelli et al. A Bayesian-Based Approach for a Short-Term Steady-State Forecast of a Smart Grid. In *Proceedings of IEEE TRANSACTIONS ON SMART GRID*, Brussels, Belgium, 2009.

[75] Jethro Dowell, Stephan Weiss, David Hill, and David Infield. Short-term spatio-temporal prediction of wind speed and direction. *Wind Energy*, 17(12) :1945–1955, 2014.

[76] Robin Girard and Denis Allard. Spatio-temporal propagation of wind power prediction errors. *Wind Energy*, 16(7) :999–1012, 2013.

[77] Julija Tastu, Pierre Pinson, Ewelina Kotwa, Henrik Madsen, and Henrik Aa. Nielsen. Spatio-temporal analysis and modeling of short-term wind power forecast errors. *Wind Energy*, 14(1) :43–60, January 2011.

[78] S. Jerez, R. M. Trigo, A. Sarsa, R. Lorente-Plazas, D. Pozo-Vázquez, and J. P. Montávez. Spatio-temporal Complementarity between Solar and Wind Power in the Iberian Peninsula. *Energy Procedia*, 40 :48–57, January 2013.

[79] J. Tastu, P. Pinson, P.-J. Trombe, and H. Madsen. Probabilistic Forecasts of Wind Power Generation Accounting for Geographically Dispersed Information. *Smart Grid, IEEE Transactions on*, 5(1) :480–489, January 2014.

[80] Miao He, Lei Yang, Junshan Zhang, and V. Vittal. A Spatio-Temporal Analysis Approach for Short-Term Forecast of Wind Farm Generation. *Power Systems, IEEE Transactions on*, 29(4) :1611–1622, July 2014.

[81] Michael Sherman. *Spatial Statistics and Spatio-Temporal Data*. Wiley, 2011.

[82] Akin Tascikaraoglu. Evaluation of spatio-temporal forecasting methods in various smart city applications. *Renewable and Sustainable Energy Reviews*, 82(Part 1) :424–435, February 2018.

[83] Michel Journée and Cédric Bertrand. Improving the spatio-temporal distribution of surface solar radiation data by merging ground and satellite measurements. *Remote Sensing of Environment*, 114(11) :2692–2704, November 2010.

[84] J. L. Bosch and J. Kleissl. Cloud motion vectors from a network of ground sensors in a solar power plant. *Solar Energy*, 95 :13–20, September 2013.

[85] Matthew Lave and Jan Kleissl. Cloud speed impact on solar variability scaling – Application to the wavelet variability model. *Solar Energy*, 91 :11–21, May 2013.

[86] S. Quesada-Ruiz, Y. Chu, J. Tovar-Pescador, H. T. C. Pedro, and C. F. M. Coimbra. Cloud-tracking methodology for intra-hour DNI forecasting. *Solar Energy*, 102 :2–10, April 2014.

[87] Elke Lorenz, Jan Kühnert, Detlev Heinemann, Kristian Pagh Nielsen, Jan Remund, and Stefan C. Müller. Comparison of global horizontal irradiance forecasts based on numerical weather prediction models with different spatio-temporal resolutions. *Progress in Photovoltaics : Research and Applications*, 24(12) :1626–1640, December 2016.

[88] Romain Dambreville, Philippe Blanc, Jocelyn Chanussot, and Didier Boldo. Very short term forecasting of the Global Horizontal Irradiance using a spatio-temporal autoregressive model. *Renewable Energy*, 72 :291–300, December 2014.

[89] C. A. Glasbey and D. J. Allcroft. A Spatiotemporal Auto-Regressive Moving Average Model for Solar Radiation. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 57(3) :343–355, 2008.

[90] Dazhi Yang, Chaojun Gu, Zibo Dong, Panida Jirutitijaroen, Nan Chen, and Wilfred M. Walsh. Solar irradiance forecasting using spatial-temporal covariance structures and time-forward kriging. *Renewable Energy*, 60 :235–245, December 2013.

[91] Massimo Lazzaroni, Stefano Ferrari, Vincenzo Piuri, Ayşe Salman, Loredana Cristaldi, and Marco Faifer. Models for solar radiation prediction based on different measurement sites. *Measurement*, 63 :346–363, March 2015.

[92] Joshua D. Patrick, Jane L. Harvill, and Clifford W. Hansen. A semiparametric spatio-temporal model for solar irradiance data. *Renewable Energy*, 87, Part 1 :15–30, March 2016.

[93] A. Tascikaraoglu, B. Sanandaji, G. Chicco, V. Cocina, F. Spertino, O. Erdinc, N. Paterakis, and J. P. Catalao. Compressive Spatio-Temporal Forecasting of Meteorological Quantities and Photovoltaic Power. *IEEE Transactions on Sustainable Energy*, PP(99) :1–1, 2016.

[94] C. Yang, A. A. Thatte, and L. Xie. Multitime-Scale Data-Driven Spatio-Temporal Forecast of Photovoltaic Generation. *IEEE Transactions on Sustainable Energy*, 6(1) :104–112, January 2015.

[95] Vincent P. A. Lonij, Adria E. Brooks, Alexander D. Cronin, Michael Leuthold, and Kevin Koch. Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors. *Solar Energy*, 97 :58–66, November 2013.

[96] Chen Yang and Le Xie. A novel ARX-based multi-scale spatio-temporal solar power forecast model. In *2012 North American Power Symposium (NAPS)*, pages 1–6, September 2012.

[97] Faranak Golestaneh, Hoay Beng Gooi, and Pierre Pinson. Generation and evaluation of space–time trajectories of photovoltaic power. *Applied Energy*, 176 :80–91, August 2016.

[98] J. R. Andrade and R. J. Bessa. Improving Renewable Energy Forecasting with a Grid of Numerical Weather Predictions. *IEEE Transactions on Sustainable Energy*, PP(99) :1–1, 2017.

[99] R. J. Bessa, A. Trindade, Cátia S. P. Silva, and V. Miranda. Probabilistic solar power forecasting in smart grids using distributed information. *International Journal of Electrical Power & Energy Systems*, 72 :16–23, November 2015.

[100] Thomas E. Hoff and Richard Perez. Quantifying PV power Output Variability. *Solar Energy*, 84(10) :1782–1793, October 2010.

[101] Hoyt C. Hottel. A simple model for estimating the transmittance of direct solar radiation through clear atmospheres. *Solar Energy*, 18(2) :129–134, January 1976.

[102] D. A. Dickey, D. P. Hasza, and W. A. Fuller. Testing for Unit Roots in Seasonal Time Series. *Journal of the American Statistical Association*, 79(386) :355–367, June 1984.

[103] Said E. Said and David A. Dickey. Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika*, 71(3) :599–607, 1984.

[104] P. A. P. Moran. The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2) :243–251, 1948.

[105] Koenig WD. Spatial autocorrelation of ecological phenomena. *Trends in Ecology & Evolution*, 14(1) :22–26, January 1999.

[106] Leo Breiman. Random Forests. *Machine Learning*, 45(1) :5–32, 2001.

[107] Yi Lin and Yongho Jeon. Random Forests and Adaptive Nearest Neighbors. *Journal of the American Statistical Association*, 101(474) :578–590, 2006.

[108] Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58 :267–288, 1994.

[109] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[110] S. Ferrari, M. Lazzaroni, V. Piuri, L. Cristaldi, and M. Faifer. Statistical models approach for solar radiation prediction. In *2013 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1734–1739, May 2013.

[111] Jie Zhang, Anthony Florita, Bri-Mathias Hodge, Siyuan Lu, Hendrik F. Hamann, Venkat Banunarayanan, and Anna M. Brockway. A suite of metrics for assessing the performance of solar power forecasting. *Solar Energy*, 111 :157–175, January 2015.

[112] S. Mekhilef, R. Saidur, and M. Kamalisarvestani. Effect of dust, humidity and air velocity on efficiency of photovoltaic cells. *Renewable and Sustainable Energy Reviews*, 16(5) :2920–2925, June 2012.

[113] E. Skoplaki and J. A. Palyvos. On the temperature dependence of photovoltaic module electrical performance : A review of efficiency/power correlations. *Solar Energy*, 83(5) :614–624, May 2009.

[114] ECMWF. Medium-range forecasts, 2017.

[115] ECMWF. Atmospheric model high resolution 10-day forecast (hres), 2017.

[116] Bei Zhang, P. Dehghanian, and M. Kezunovic. Spatial-temporal solar power forecast through use of Gaussian Conditional Random Fields. In *2016 IEEE Power and Energy Society General Meeting (PESGM)*, pages 1–5, July 2016.

[117] Yao Zhang and Jianxue Wang. K-nearest neighbors and a kernel density estimator for GEFCom2014 probabilistic wind power forecasting. *International Journal of Forecasting*, 32(3) :1074–1080, July 2016.

[118] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. CRC Press, April 1986.

[119] George R. Terrell and David W. Scott. Variable Kernel Density Estimation. *The Annals of Statistics*, 20(3) :1236–1265, 1992.

[120] M. P. Wand and M. C. Jones. *Kernel Smoothing*. CRC Press, December 1994. Google-Books-ID : GTOOi5yE008C.

[121] David W. Scott. *Multivariate Density Estimation : Theory, Practice, and Visualization*. John Wiley & Sons, March 2015. Google-Books-ID : XZ03BwAAQBAJ.

[122] A. R Mugdadi and Ibrahim A Ahmad. A bandwidth selection for kernel density estimation of functions of random variables. *Computational Statistics & Data Analysis*, 47(1) :49–62, August 2004.

[123] Travis A. O'Brien, Karthik Kashinath, Nicholas R. Cavanaugh, William D. Collins, and John P. O'Brien. A fast and objective multidimensional kernel density estimation method : fastKDE. *Computational Statistics & Data Analysis*, 101 :148–160, September 2016.

[124] Tarn Duong and Martin L. Hazelton. Cross-validation Bandwidth Matrices for Multivariate Kernel Density Estimation. *Scandinavian Journal of Statistics*, 32(3) :485–506, September 2005.

[125] Xavier D'HAULTFOEUILE & Pauline GIVORD. *La régression quantile en pratique*. Document de travail INSEE, Janvier 2013.

[126] T. M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley, August 1991.

[127] J. Juban, N. Siebert, and G. N. Kariniotakis. Probabilistic Short-term Wind Power Forecasting for the Optimal Management of Wind Generation. In *2007 IEEE Lausanne Power Tech*, pages 683–688, July 2007.

[128] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[129] Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5) :559–570, October 2000.

[130] M. Noia, C. F. Ratto, and R. Festa. Solar irradiance estimation from geostationary satellite data : I. Statistical models. *Solar Energy*, 51(6) :449–456, December 1993.

[131] M. Noia, C. F. Ratto, and R. Festa. Solar irradiance estimation from geostationary satellite data : II. Physical models. *Solar Energy*, 51(6) :457–465, December 1993.

[132] Richard Perez, Pierre Ineichen, Kathy Moore, Marek Kmiecik, Cyril Chain, Ray George, and Frank Vignola. A new operational model for satellite-derived irradiances : description and validation. *Solar Energy*, 73(5) :307–317, November 2002.

[133] Annette Hammer, Detlev Heinemann, Carsten Hoyer, Rolf Kuhlemann, Elke Lorenz, Richard Müller, and Hans Georg Beyer. Solar energy assessment using remote sensing technologies. *Remote Sensing of Environment*, 86(3) :423–432, August 2003.

[134] C Rigollier, M Lefèvre, and L Wald. The method Heliosat-2 for deriving shortwave solar radiation from satellite images. *Solar Energy*, 77(2) :159–169, January 2004.

[135] Viorel Badescu, Christian A. Gueymard, Sorin Cheval, Cristian Oprea, Madalina Baciu, Alexandru Dumitrescu, Flavius Iacobescu, Ioan Milos, and Costel Rada. Computing global and diffuse solar hourly irradiation on clear sky. Review and testing of 54 models. *Renewable and Sustainable Energy Reviews*, 16(3) :1636–1656, April 2012.

[136] Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets. *Solar Energy*, 132 :25–37, July 2016.

[137] Philippe Blanc, Jan Remund, and Loic Vallance. Short-term solar power forecasting based on satellite images. In George Kariniotakis, editor, *Renewable Energy Forecasting*, Woodhead Publishing Series in Energy, pages 179 – 198. Woodhead Publishing, 2017.

[138] Zhenzhou Peng, Dantong Yu, Dong Huang, John Heiser, and Paul Kalb. A hybrid approach to estimate the complex motions of clouds in sky images. *Solar Energy*, 138(Supplement C) :10–25, November 2016.

[139] Zibo Dong, Dazhi Yang, Thomas Reindl, and Wilfred M. Walsh. Satellite image analysis and a hybrid ESSS/ANN model to forecast solar irradiance in the tropics. *Energy Conversion and Management*, 79(Supplement C) :66–73, March 2014.

[140] S. Cros, O. Liandrat, N. Sébastien, and N. Schmutz. Extracting cloud motion vectors from satellite images for solar power forecasting. In *2014 IEEE Geoscience and Remote Sensing Symposium*, pages 4123–4126, July 2014.

[141] H. Escrig, F. J. Batlles, J. Alonso, F. M. Baena, J. L. Bosch, I. B. Salbidegoitia, and J. I. Burgaleta. Cloud detection, classification and motion estimation using geostationary satellite imagery for cloud cover forecast. *Energy*, 55(Supplement C) :853–859, June 2013.

[142] Alvaro Linares-Rodriguez, Samuel Quesada-Ruiz, David Pozo-Vazquez, and Joaquin Tovar-Pescador. An evolutionary artificial neural network ensemble model for estimating hourly direct normal irradiances from meteosat imagery. *Energy*, 91(Supplement C) :264–273, November 2015.

[143] Cyril Voyant, Pierrick Haurant, Marc Muselli, Christophe Paoli, and Marie-Laure Nivet. Time series modeling and large scale global solar radiation forecasting from geostationary satellites data. *Solar Energy*, 102 :131–142, April 2014.

[144] Ricardo Marquez, Hugo T. C. Pedro, and Carlos F. M. Coimbra. Hybrid solar forecasting method uses satellite imaging and ground telemetry as inputs to ANNs. *Solar Energy*, 92(Supplement C) :176–188, June 2013.

[145] H. S. Jang, K. Y. Bae, H. S. Park, and D. K. Sung. Solar Power Prediction Based on Satellite Images and Support Vector Machine. *IEEE Transactions on Sustainable Energy*, 7(3) :1255–1263, July 2016.

[146] Björn Wolff, Jan Kühnert, Elke Lorenz, Oliver Kramer, and Detlev Heinemann. Comparing support vector regression for PV power forecasting to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data. *Solar Energy*, 135 :197–208, October 2016.

[147] L. Mazorra Aguiar, B. Pereira, M. David, F. Díaz, and P. Lauret. Use of satellite data to improve solar radiation forecasting with Bayesian Artificial Neural Networks. *Solar Energy*, 122(Supplement C) :1309–1324, December 2015.

[148] L. Mazorra Aguiar, B. Pereira, P. Lauret, F. Díaz, and M. David. Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting. *Renewable Energy*, 97(Supplement C) :599–610, November 2016.

[149] Philippe Blanc, Benoît Gschwind, Mireille Lefèvre, and Lucien Wald. The HelioClim Project : Surface Solar Irradiance Data for Climate Applications. *Remote Sensing*, 3(2) :343–361, February 2011.

[150] Lucien Wald, Michel Albuisson, Clive Best, Catherine Delamare, Dominique Dumortier, Elena Gaboardi, Anette Hammer, Detlev Heinnemann, Richard Kift, Stefan Kunz, Mireille Lefèvre, Sébastien Leroy, Mario Martinoli, Lionel Ménard, John H. Page, Tamas Prager, Corrado Ratto, Christian Reise, Jan Remund, Aniko Rimoczi-Paal, Eric Van Der Goot, Franz Vanroy, and Ann Webb. SoDa : a Web service on solar radiation. In *Eurosun 2004*, volume 3, pages 921–927, Freiburg, Germany, 2004. PSE GmbH, Freiburg, Germany. ISBN 3-9809656-4-3.

[151] Christophe Vernay, Philippe Blanc, and Sébastien Pitaval. Characterizing measurements campaigns for an innovative calibration approach of the global horizontal irradiation estimated by HelioClim-3. *Renewable Energy*, 57(Supplement C) :339–347, September 2013.

[152] Claire Thomas, Laurent Saboret, Etienne Wey, Philippe Blanc, and Lucien Wald. Preliminary assessment of a new SoDa service for real-time estimates and short-term forecasts of the solar radiation. In *15th EMS Annual Meeting / 12th ECAM*, USKKey, page 335, Sofia, Bulgaria, September 2015. EMS European Meteorological Society.

[153] Benoît Gschwind and Lucien Wald. "HC-1+HC-3" a long-term data set of daily solar radiation at surface. EMS Annual Meeting, September 2017. Poster.

[154] Romain Dambreville. *Nowcasting and very short term forecasting of the global horizontal irradiance at ground level : application to photovoltaic output forecasting.* Theses, Université de Grenoble, October 2014.

[155] Daniel Wartenberg. Multivariate Spatial Correlation : A Method for Exploratory Geographical Analysis. *Geographical Analysis*, 17(4) :263–283, October 1985.

[156] F. Glineur. *Etude des méthodes de point intérieur appliquées à la programmation linéaire et à la programmation semidéfinie.*

[157] GEORGE B. DANTZIG. *Linear Programming and Extensions.* Princeton University Press, 1991.

[158] Roger Koenker and Jose A. F. Machado. Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association*, 94(448) :1296–1310, 1999.

Annexe A

Les articles soumis à des journaux à comité de lecture

Les articles présentés ici sont en version preprint. La version finale de ces articles et les références de citation peuvent être retrouvées sur les sites respectifs des journaux à qui les droits ont été cédés.

Short-Term Spatio-Temporal Forecasting of Photovoltaic Power Production

Xwégnon Ghislain Agoua, Robin Girard and George Kariniotakis, *Senior Member, IEEE*

***Abstract*—In recent years, the penetration of photovoltaic (PV) generation in the energy mix of several countries has significantly increased thanks to policies favoring development of renewables and also to the significant cost reduction of this specific technology. The PV power production process is characterized by significant variability, as it depends on meteorological conditions, which brings new challenges to power system operators. To address these challenges it is important to be able to observe and anticipate production levels. Accurate forecasting of the power output of PV plants is recognized today as a prerequisite for large-scale PV penetration on the grid. In this paper, we propose a statistical method to address the problem of stationarity of PV production data, and develop a model to forecast PV plant power output in the very short term (0-6 hours). The proposed model uses distributed power plants as sensors and exploits their spatio-temporal dependencies to improve forecasts. The computational requirements of the method are low, making it appropriate for large-scale application and easy to use when on-line updating of the production data is possible. The improvement of the normalized root mean square error (nRMSE) can reach 20% or more in comparison with state-of-the-art forecasting techniques.**

***Index Terms*—Autoregressive processes, forecasting, photovoltaic systems, smart grids, spatial correlation, stationarity, time series.**

I. Introduction

GROWING global energy demand and increased awareness of the consequences of climate change have put renewable energy in the spotlight. Renewable energy generation, and particularly photovoltaic (PV) energy, is continuously increasing in several countries, especially in Europe. The power output of a PV plant depends on meteorological conditions. In regions subject to active weather changes, it is characterized by high variability and low short-term predictability. These characteristics challenge power system operators, since they introduce uncertainties into the various functions of power system management, especially for large-scale PV integration.

The PV production expected in the next few minutes, hours or days needs to be accurately forecasted in order to efficiently perform functions like scheduling power systems, minimizing reserve costs [1], trading PV production in electricity markets and coordinating PV plants with storage, and in general to contribute to increasing the competitiveness of renewable energy technologies [2]. In the context of smart grids, PV forecasts are necessary to manage distribution networks, microgrids or smart homes, where other options like active demand, storage, electric vehicles etc., coexist with PV generation [1], [3].

The literature proposes several methods to forecast PV production [4]. These methods can be classified according to their specific forecast horizon [5]. The final choice of forecasting technique is related to this horizon and the available data. The most common statistical methods are regression methods like linear regression, regression trees, boosting, bagging, random forests, Support Vector Machines [6]–[9], and semi-parametric models. These techniques investigate the correlation between the historical production and the related meteorological measurements [10]. The Box and Jenkins time series treatment methods (ARIMA, ARMA, SARIMA, ...) are also widely used in PV power forecasting. The question of the series stationarity is treated by pre-processing steps using either clear sky modeling, [11]–[13] or certain normalization techniques employing Top of Atmosphere (TOA) or Global Horizontal Irradiance (GHI). In [14], [15], regression-based methods are also used. Data mining techniques are employed to cluster past events into historical data on production and/or meteorological variables. This same idea of similarity is used to forecast production when PV panels are covered by snow [16].

Neural networks have been used to forecast PV production with different types of activation functions [17]. They are often compared or coupled with physical models [18], [19]. They can also be used as a second step in a two-step modeling chain, where the first step is to predict meteorological variables using Numerical Weather Predictions (NWP) [20], [21].

Recent years have seen increasing interest in techniques that can take into account not only historical data about the site that is the object of forecasting, but also other spatially distributed data. These methods, initially proposed for wind power forecasting, are developed for different applications, like identifying regions with high energy production potential [22], [23], studying the spatial propagation of forecasting errors [24], [25], and even "geographically intelligent" prediction [26]–[28].

Most references refer to spatio-temporal solar irradiation forecasts. Spatial information from sky cameras or satellite images is used and described in 2D or even 3D with cloud motion vectors. Cloud movement predictions lead to solar radiation forecasts for very short-term horizons (a few minutes up to 2-4 hours ahead) [29]–[31]. NWP models and cloud motion vectors can also be combined for short-term forecasts (a few hours up to 2 or 3 days ahead) [32]. Solar radiation can also be forecasted with auto-regressive models in time and

The authors are with MINES ParisTech, PSL Research University, PERSEE - Centre for Processes, Renewable Energies and Energy Systems CS 10207 rue Claude Daunesse, 06904 Sophia Antipolis Cedex, France. (e-mails: xwegnon.agoua@mines-paristech.fr, robin.girard@mines-paristech.fr, georges.kariniotakis@mines-paristech.fr)

Manuscript received February 13, 2017; revised May 31, 2017 and July 20, 2017; accepted August 26, 2017.

space or kriging [33]–[36]. These methods employed in solar radiation forecasts can be costly due to the complexity of the required measuring infrastructure and data, and the modeling chain that has to be developed.

In this paper we propose a forecasting methodology that exploits the spatial and temporal correlations in existing data from geographically dispersed PV installations to predict the power output of a specific plant. Short-term forecast horizons of a few minutes up to 6 hours are considered. The models investigated here directly use geographically dispersed power plants as a network of sensors. This differentiates the approach from methods that use off-site data from meteorological stations and ground-based irradiance sensors as in [37]. The proposed model does not consider input from a NWP model, and forecasts are made based on the production data and not global irradiance data as is the case in [38], [39].

In a preceding conference paper [40], the authors have proposed a spatio-temporal methodology. In this paper that methodology is significantly improved on several points. The first improvement consists in proposing a new stationarization process, that unlike [41], does not involve modeling for the clear sky generation. The proposed approach aims to overcome weaknesses of the clear sky based normalization especially for early and late hours of the day when solar irradiation is low. The second improvement proposed here permits to take into account the local meteorological conditions in the spatio-temporal model. This is done by defining model coefficients dependent on the weather variables in the estimation process. The third improvement is to propose a model that integrates an automatic selection of the appropriate input variables. This is particularly adapted to highly dimensional problems, as can be the case for spatio-temporal PV forecasting. Finally, the spatial density of the considered PV plants in real-world cases can be variable, and for this reason we illustrate the usefulness of the proposed methodology with two test cases featuring a low and high number of PV plants. The dimensionality problem and the importance of the proposed variable selection process are highlighted through the test case with high number of PV installations (185 PV plants). The benefits in terms of performance of all the above contributions with respect to [40] are presented in section IV-C.

The paper is structured as follows: the potential of making use of spatio-temporal information is investigated in section II with a focus on the proposed stationarization procedure, the data and the evaluation criteria. The proposed spatio-temporal models are presented in section III. The results are presented and discussed in section IV. Finally, the conclusions of the study are discussed in section V.

II. Analysis of the Interest of Spatio-temporal Modeling

The aim in this section is to demonstrate the interest of using spatio-temporal information for PV forecasting purposes. This is done through an analysis of the correlations between data from PV plants. However, given that these data are dominated by the daily sun cycle, which biases correlation analysis, it is necessary to subtract the periodic components in the series through appropriate stationarization. To achieve this, we propose a new method to stationarize PV production series that is not only useful for analyzing correlations but also for building the forecasting models themselves. Initially, two test cases are introduced that provide real world data, used in this section to assess the proposed stationarization process, and in later sections to evaluate the proposed forecasting models.

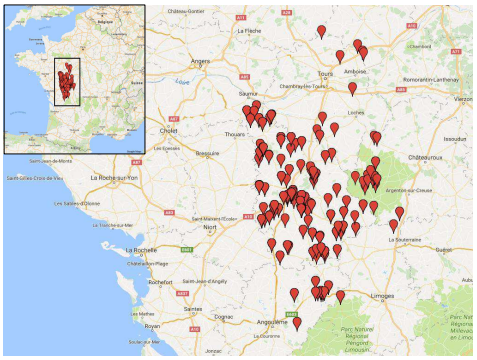


Fig. 1. The power plants of the second data set d_2 located in west central France. The distance between the power plants ranges from 1 km to 230 km.

A. Test Cases

Two data sets are considered in this paper corresponding to relatively different climatic conditions and different spatial densities of installed PV plants as well as the distances between them. The first data set labeled d_1, consists of time series of the measured PV generation of a set of 9 power plants located in the south of France. The power plants are labeled P1-P9 with peak power ranging from 45 kWc to 5 MWc. The distance between the power plants ranges from 5 km to 465 km. The measurements cover a period of 20 months starting from July 2013 with a resolution of 6 min to 15 min depending on the PV plant. The data quality has been checked to remove inconsistencies and then interpolated to produce series with a 15 min temporal resolution that are used hereafter.

The second data set labeled d_2 is a good illustration of a case featuring a high number of power plants and significant geographic density. It comprises the output of 905 PV power inverters in the mid-west region of France with peak power ranging from 3.2 kWc to 58 kWc. This amount of inverters corresponds to 185 different PV power plants (set of inverters at the same location). The distance between them varies from 1 km to 230 km. The data relate to the period from November 2014 to March 2016. The original time resolution of the data is 5 min, which was averaged to produce series with a 15 min temporal resolution as with the previous test case. The locations of the power plants in the test case d_2 are represented in Figure 1.

B. The Stationarization Procedure

Most of the time series analysis methods require stationary series. The photovoltaic production series are not stationary because the average production depends on the time of day, while the variability, as expressed by the variance of the

production, depends on the level of production and indirectly on the time of day. A simple differentiation of the series is not efficient in producing stationary time series because the non-stationarity in the variance remains.

Here we propose a procedure to stationarize a PV production series. The aim is to decompose the production series using a deterministic component that describes the movement of the sun. It is inspired from the clear sky index for solar radiation [42]–[44].

The clear sky index for solar radiation represents the way that the atmosphere attenuates light on an hour-to-hour or day-to-day basis as a function of the movement of the earth around sun. It is defined as the quotient of radiation actually measured by the radiation simulated with a clear sky model.

This index makes it possible to remove the diurnal and seasonal pattern from irradiation data, which is expected to improve the performance of the statistical techniques applied thereafter. Here, we define it as the ratio between irradiation measurements and an advanced clear sky estimate at time t:

$$k_t^{irr} = \frac{I_t^{meas}}{I_t^{sim}}.$$

In a similar way, we define a clear sky index for photovoltaic power k^{pv} as

$$k_t^{pv} = \frac{P_t^{meas}}{P_t^{sim}} \tag{1}$$

where P_t^{meas} is the PV production measured at time t, P_t^{sim} is the simulated production output for time t. P_t^{sim} is constructed as the product of the PV overall system efficiency parameter η and the simulated irradiation I_t^{sim} at either the Top of Atmosphere (ToA) level or under clear sky conditions as proposed by the European Solar Radiation Atlas (ESRA) model [43]. The parameter η embedded the efficiency of the generator and the active surface.

Although intuitively, the index k_t^{pv} would be expected to be adequate for de-trending, in practice appropriate stationarity tests on the resulting series (i.e. unit roots tests) indicate that the results are not satisfactory. For this reason we propose a new relation between the actual production and P_t^{sim} using a function f that would explain more accurately the link between the two productions. This function would also help to reduce the non-stationarity when defining the new working series u_t for the hours at which P_t^{sim} is not zero as

$$u_t = P_t / f(P_t^{sim}). \tag{2}$$

The irradiation considered for defining P_t^{sim} is the simulated ESRA series as it embeds more information about the atmospheric characteristics than the ToA, such as albedo, air mass, the Linke turbidity factor and other atmospheric conditions. The simulation of irradiation was done under the hypothesis of a horizontal surface; this is because the inclination does not affect the stationarization since the variation in the output level it produced would be assimilated by η. Different types of relation can be conceived for f including linear, quadratic and piecewise linear.

The choice of the appropriate function was made using a quantitative criterion based on the evolution of the daily standard deviation of the series u_t. The retained function is piecewise linear in the simulated production and depends on the direction of the productions daily evolution (either increasing at the beginning of the day or decreasing after solar noon). It can be expressed as:

$$f(P_t^{sim}) = P_t^{sim} + f_a(P_t^{sim}) + f_b(P_t^{sim}) \tag{3}$$

where the function f_a is defined from sunrise to noon and f_b from noon to sunset. The goal of f_a and f_b is to improve the treatment at the beginning and at the end of the day. Their definition on a daily basis is:

$$\begin{cases} f_a(0) = \alpha_a \\ f_a\left(\beta_a \frac{P_{max}^{sim}}{2}\right) = 0 \\ f_a(P_{max}^{sim}) = \gamma \end{cases} \quad \begin{cases} f_b\left(P_{max}^{sim}\right) = \gamma \\ f_b\left(\beta_b \frac{P_{max}^{sim}}{2}\right) = 0 \\ f_b(0) = \alpha_b \end{cases} \tag{4}$$

where P_{max}^{sim} is the maximum production simulated for the day. The values of the coefficients $\alpha_{a,b}, \beta_{a,b}, \gamma$ are obtained through an optimization process that aims to minimize the standard deviation criterion. The optimization is made under the constraints $\beta_{a,b} \in (0, 2)$. The coefficients are randomly initialized and then optimized considering a sliding window of one-month over the ESRA irradiation time series. The sliding window covers the period prior to the day of interest. The stationarity of the normalized form of u_t was evaluated by analyzing its autocorrelogram and computing unit root test.

The procedure can be summarized in the following steps for a power plant:

1) Clean the spurious data from the PV production series.
2) Simulate the ESRA clear sky irradiation series and the corresponding power series using the plant's efficiency.
3) Determine the appropriate coefficients of the functions (f_a, f_b) using an optimization process on a sliding interval of simulated irradiation values.
4) Normalize the measured series P_t^{meas} to obtain the series u_t.

C. Analysis of Spatial Correlations

To investigate the existence of spatio-temporal patterns, we evaluate the cross-correlation between the lagged production series. However, this requires eliminating the effect of East to West correlation transfer by considering the stationarized series for the PV plants.

Figure 2 presents the empirical cumulative distribution of the cross-correlation values for the power plants in the data set d_2. Three distributions are plotted for three classes of distance between the power plants (from the closest to the farthest). The figure shows that the cross-correlation values are higher for the first class of distance (less 50 km) than for the last class (more than 100 km). As the effect of the bell-shape in the stationarized production data is absent, we can assume that the link described by these correlation values is due to a spatial transfer of information between the power plants mainly due to cloud movements. This analysis confirms the interest of a forecasting solution that takes into account both the temporal and spatial variability of the production series.

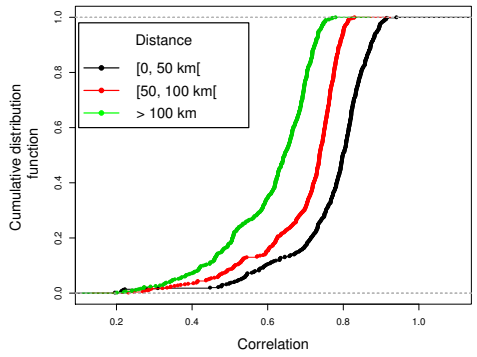


Fig. 2. Data set d_2: Cumulative distribution function (CDF) of the cross-correlation between the lagged production series. Green, red and black CDFs respectively described three classes of distance between the power plants.

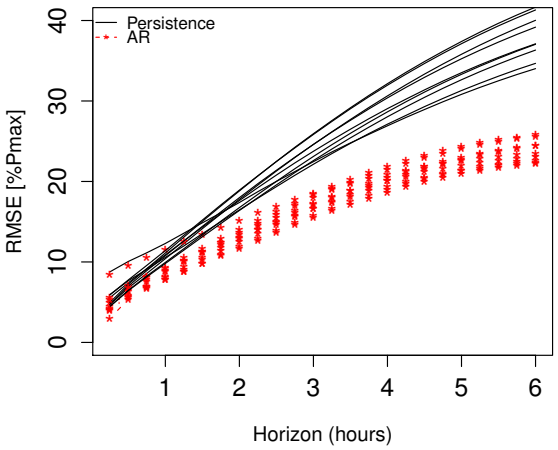


Fig. 3. Data set d_1: Comparison of the normalized RMSE of AR and persistence models over the testing set. Solid and dotted lines represent respectively the performance of persistence AR models. The forecast time step is 15 min.

III. A Model for Spatio-Temporal PV Forecasting

A. The Reference Model

In order to be able to compare the advantages of a spatio-temporal approach for PV forecasting, we introduce reference models for benchmarking that does not use such geographically distributed information. Several methods can be used to forecast PV generation as presented in the introduction. The persistence model is often used as a reference in the literature on renewable energy forecasting to compare the performance of advanced models, as it is easy to compute, is based only on measured data, and does not involve any modeling processes. Thus, the persistence results are easily replicable. Moreover, in practical applications of PV forecasting, persistence is often chosen as a fallback model to provide forecasts in case advanced models fail. We define here as persistence a model that considers that the power production of a PV plant at time $t+h$ is the same as the production of that plant at the same time on the previous day. This approach does not consider any off-site data. Despite its popularity as a reference model in the literature, its overall performance is poor [4]. To account for the different factors that affect PV production one could adjust persistence as a function of the observed values on the current day. However, this already involves some data manipulation, and different options could be considered, but such empirical adjustments are out of the scope of this paper. To avoid obtaining overoptimistic results from a spatio-temporal method, it is also necessary to use an advanced reference" model featuring state-of-the-art performance and reasonable complexity so that results can be easily reproduced. For this purpose we consider autoregressive (AR) models described as:

$$\hat{P}^x_{t+h|t} = \hat{\beta}^0_h + \sum_{l=0}^{L} \hat{\beta}^l_h P^x_{t-l} \tag{5}$$

where P^x_t is the production of the power plant x at time t and $\hat{P}^x_{t+h|t}$ the prediction for horizon h. The appropriate maximum time lag L is chosen by minimizing the Akaike Information Criterion (AIC). We applied this model to the data set d_1 using 15 months for learning and 5 months for the tests. Forecasts are updated at each 15-minute time step.

For the 5 months of testing set for each power plant, we also applied the persistence and compared its performance with the AR model. Figure 3 presents the normalized root mean square error RMSE for the AR and persistence models for the d_1 power plants as a function of the prediction horizon. The figure shows that the best model is the AR model, as its RMSE levels are the lowest. We thus retain the AR model as a reference in this paper to evaluate the performance of the spatio-temporal forecasting models. With our reference model thus defined, we can evaluate the contribution of integrating additional information from neighboring plants.

B. The Proposed Spatio-Temporal Model

The correlation analysis carried out in subsection II-C confirms the interest of using measurements from other power plants to increase the quality of the PV power forecasts. We propose here a spatio-temporal model that produces PV power forecasts for a power plant using measurements from other plants nearby.

Let $\mathcal{X}$ be the set of N PV plants and Ls the appropriate maximum lag. The forecast model for a power plant of interest x is then defined as:

$$P^x_t = \beta^0 + \sum_{l=0}^{Ls} \sum_{y \in \mathcal{X}} \beta^{l,y} P^y_{t-l} \quad . \tag{6}$$

For a selected horizon h, the coefficients $\beta = (\beta^0, \beta_r)$ with $\beta_r = (\beta^{l,y})_{0 \le l \le Ls,\, y \in \mathcal{X}}$ are estimated using a least squares method that involves minimizing the Residual Sum of Squares (RSS):

$$RSS(\beta) = \|\mathbf{P^x} - \mathbf{X}\,\beta\|^2, \tag{7}$$

where $\mathbf{P^x}$ is the measurement for power plant x.

$\mathbf{X}$ is a $N \times (Ls+1)$ matrix the lines of which are the current and lagged production for the power plants y_i

$$\mathbf{X} = \begin{pmatrix} 1 & P_t^{y_1} & \dots & P_{t-Ls}^{y_1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & P_t^{y_N} & \dots & P_{t-Ls}^{y_N} \end{pmatrix}. \quad (8)$$

The forecast at time t for the horizon h for a power plant x is then defined by:

$$\hat{P}_{t+h|t}^{x} = \hat{\beta}_h^0 + \sum_{l=0}^{Ls} \sum_{y \in \mathcal{X}} \hat{\beta}_h^{l,y} P_{t-l}^{y}. \quad (9)$$

The first issue related to the above model is the dimensionality problem when there is a high number of PV plants. To reduce the complexity of the model in such cases, we propose a two-step variables selection procedure. Let us call x the power plant of interest for which the forecasts are made. The first step is to compute the distance between the plant x and the other plants and select the n_p closest plants to x. The second step is to apply a stepwise selection procedure based on the AIC criterion. The selection is made backward; the n_p variables and their respective lags are integrated into the model and then removed one by one and the AIC is recalculated each time. The model with the minimum AIC is retained.

C. Extension of the Model: Spatio-Temporal Model using Clusters of Meteorological Conditions

The previous model is purely based on the historical production data. Here, we propose a variant of the model that allows a smooth dependency of the linear model coefficients on local meteorological conditions. The meteorological variables can be temperature, wind speed or direction, or another variable. These measurements are obtained from the closest weather station. With the previous notation, the forecast for the horizon h is denoted as:

$$\hat{P}_{t+h|t}^{x} = \hat{\beta}_h^0(Z) + \sum_{l=0}^{L} \sum_{y \in \mathcal{X}} \hat{\beta}_h^{l,y}(Z) P_{t-l}^{y}, \quad (10)$$

where Z represent the meteorological variables. The coefficients are estimated by weighted least square regression by:

$$\hat{\beta}_h^{l,y}(z) = Arg\min \sum_t \phi\left(\frac{Z_{t,y} - z}{\gamma}\right)\left(P_t^y - P_{t+h}^y\right)^2 \quad (11)$$

where the coefficient γ is the mean of the random variable Z. The weights function is exponential:

$$\phi(x) = \exp\left(-||x||^2/2\right). \quad (12)$$

The weights are calculated using the measurements from the closest available meteorological station.

D. Improved Variable Selection Procedure

In the model presented above, the dimensionality problem (i.e. high number of variables) is treated with a simple selection variable procedure. The model can be modified to directly treat the variable selection issue using LASSO. The Least Absolute Shrinkage and Selection Operator [45] regression integrates a penalty into the minimization problem by applying a constraint on the sum of the absolute values of the coefficients. The estimator is defined as:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} RSS(\beta) + \lambda \|\beta\|_1 \right\}. \quad (13)$$

Some bias is introduced but the variance is reduced. The selection of the coefficients is automatic and some of them are set to zero for high values of the penalization parameter λ. The regularization parameter λ is obtained by cross-validation and the path of the solutions of β is piecewise linear in λ.

IV. Evaluation

The proposed models are applied to the data sets d_1 and d_2 for a 6-hour horizon with a 15-min time step, and with a sliding window scheme that updates forecasts every 15 min. The forecasts are compared to those of the reference model. The models were developed using the software R [46].

A. Impact of the Stationarity Procedure on Forecast Errors

The reference AR model was applied to the two types of production series of d_1: the raw series and the series that was stationarized following the procedure proposed in Section II. The RMSE for the respective series was computed for each power plant and the improvement due to the stationarization was calculated. For all the plants except P4, there is a significant improvement in RMSE when stationarized series are used. The average improvement in terms of RMSE is 7%. The stationarized series perform better than the raw inputs. The case of P4 can be explained by the fact that the AR model efficiently captures the temporal variability with standard normalization.

The same analysis was made of the power plants in data set d_2, where 136 power plants were retained after data cleaning. Figure 4 represents the improvement of the RMSE achieved with the stationary procedure for d_2. The mean improvement for 3-hour horizons is 10% and can reach 15%. This significant reduction in forecasting errors confirms the efficiency of the stationarization method and the interest of using it to pre-process data before integrating them into the forecasting model. Thus hereafter, we use the stationarized series.

B. Performance of the Spatio-Temporal Model

For each of the power plants of d_1, we apply the spatio-temporal model in its form defined in part III-B. The standardized errors are computed at time t for look-ahead time h ranging from 15 min to 6 hours. The densities of the prediction errors are computed using kernel density estimation and are presented in figure 5 for two power plants and different

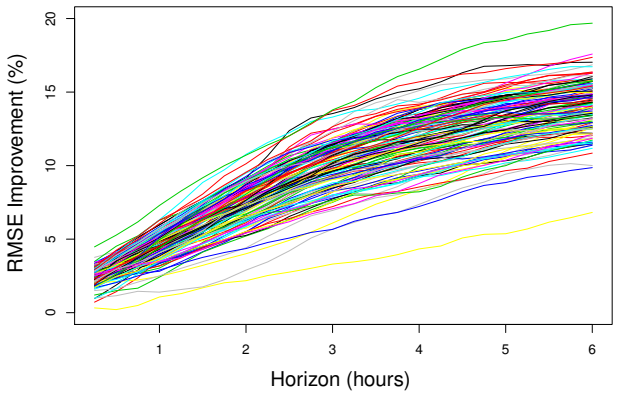


Fig. 4. Data set d_2: RMSE Improvement of the AR model with stationary series over non-transformed data. Each line represents the improvement obtained for a power plant. The time step is 15 min.

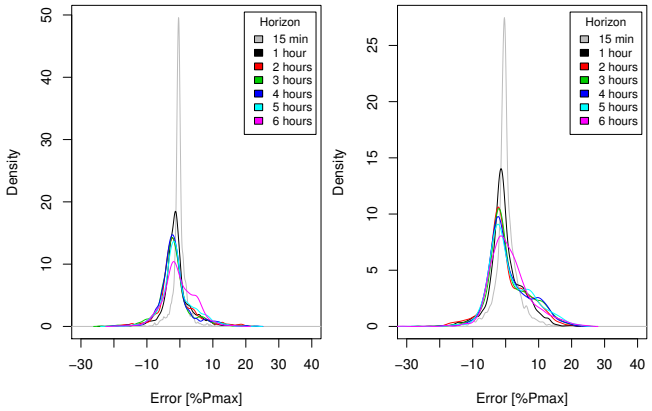


Fig. 5. Densities of the forecasting errors post spatio-temporal model for 2 power plants (Kernel estimation). The horizons range from 15 min to 6 hours.

horizons. Note that for both power plants the distributions are not Gaussian, as the modes and averages are significantly different. The averages are close to zero and the skew is negative. As the horizon increases, the distribution mode shifts to the left. The same analysis was performed on the other power plants of d_1 with the same conclusions.

To obtain a more complete overview of the proposed models performance, we compare it to random forest (RF) models. RF models are shown in the literature [10] to be one of the most efficient models to produce accurate forecasts of PV power production. We thus computed an RF model and compared its performance to the spatio-temporal model. Table I presents the minimum, mean and maximum RMSE improvement over the 6-hour time horizons of the spatio-temporal model (ST) w.r.t. the AR and RF models for a sample of five power plants of d_1. The table shows an average improvement of around 10% for the ST model compared with the AR model and 6% compared with the RF one. The improvement compared to the AR and RF models can reach respectively 20% and 15%. The improvement values are quite similar for all of the power plants except for plant P8, for which there is no improvement. This is the most distant power plant, and the spatial correlation does not reach it.

TABLE I
RMSE IMPROVEMENT OF THE SPATIO-TEMPORAL (ST) MODEL OVER THE REFERENCE AR MODEL AND THE RANDOM FOREST (RF) MODEL FOR 5 POWER PLANTS OF DATA SET d_1.

Improvement of RMSE (%)		P1	P2	P4	P5	P6
ST vs AR	min	0.4	3.02	0.61	-0.46	0.83
	mean	9.49	13.05	7.36	8.69	12.57
	max	16.81	19.27	12.5	15.71	20.13
ST vs RF	min	0.17	2.94	0.32	-0.72	2.14
	mean	6.52	10.27	4.5	5.03	7.84
	max	15.3	16.6	9.03	11.12	11.39

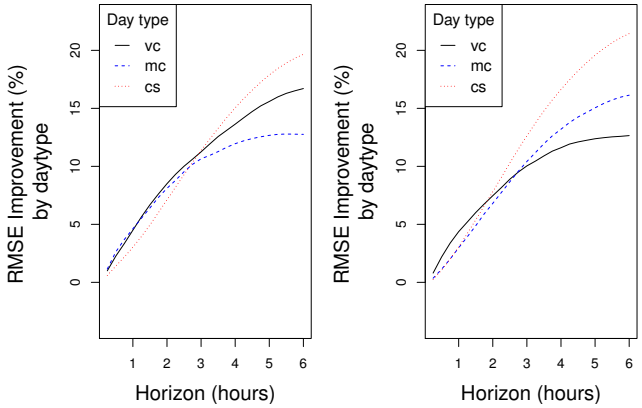


Fig. 6. RMSE Improvement of spatio-temporal model compared to the reference model by day type for two power plants of data set d_1.
The day types are very cloudy (vc), moderately cloudy (mc) and clear sky(cs).

The analysis of the performance of the spatio-temporal model can be related to the sky cover. The days of the testing set can be clustered according to sky cover level. We then define three levels of sky coverage: clear sky (cs), moderately cloudy (mc) and very cloudy (vc). These levels were computed using an index based on the ratio of the sum of the daily production to the sum of the simulated irradiation using the ESRA model. Figure 6 presents, for two power plants of d_1, the improvement of the spatio-temporal model compared to the reference model by type of day.

We observe that for the first two hours the improvement on cloudy days exceeds that of clear days. This observation shows that the spatio-temporal model helps to capture the movement of the clouds. The graphs also show that the improvement is greater for clear sky days for the longer horizons and that even on the cloudiest days, the improvement exceeds 5%. This analysis produced similar results for the other power plants.

C. Wind Speed Effect on the Model Performances

We choose the wind speed for the meteorological variable Z as presented in the model extension in part III-C. This choice is motivated by the fact that surface wind speed affects the performance of PV modules given its relation with ambient temperature. Also, wind conditions are generally related to cloud movement, which affects PV production. Note however that by considering surface wind speed, which is in general considerably different from wind speeds in upper layers of the

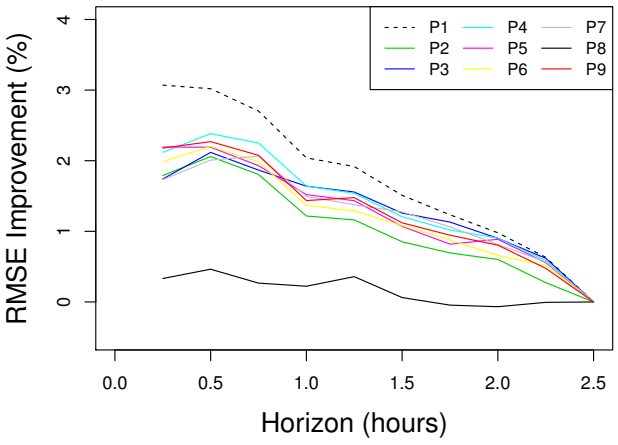


Fig. 7. Data set d_1: RMSE improvement of the spatio-temporal model conditioned by wind speed in comparison with the model without conditioning. Each line represents the improvement obtained for a power plant.

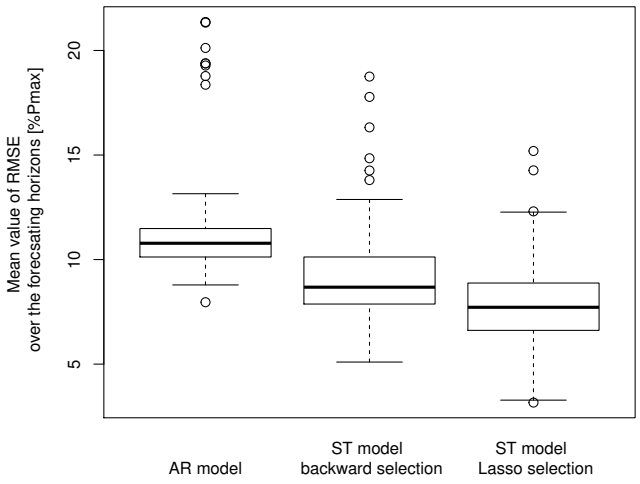


Fig. 8. Data set d_2: Distribution of the mean value (over the 6-hour prediction horizon) of RMSE for the reference model, the spatio-temporal model (ST) with backward selection, and the spatio-temporal model with Lasso selection.

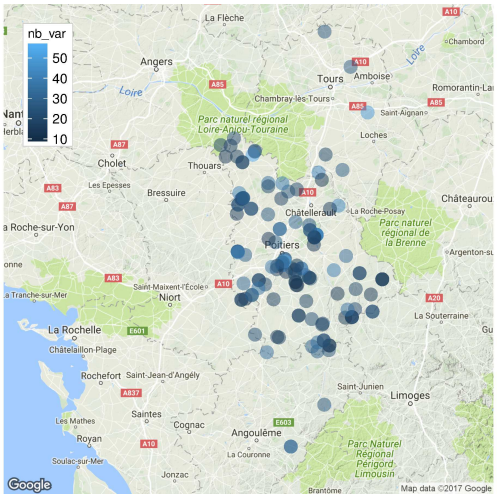


Fig. 9. Data set d_2: Map of the power plants. For each plant, the color is defined by the number of neighboring plants selected by the Lasso.

atmosphere, the aim is not to make an explicit relation with cloud movement.

The spatio-temporal model with a conditioned wind speed parameter was then applied to d_1. Compared to the spatio-temporal model with fixed parameters, this model shows a reduction in RMSE for the first two hours as shown in figure 7. The mean value of this improvement is 2% and the most significant reduction is noted for the first forecasting hour. After two hours, the model with conditioning shows no improvement compared to the model without conditioning. These results are promising and show that there is a potential for improving forecast quality by using adequate meteorological variables within the model.

In the paper [40], the average RMSE improvement of the proposed spatio-temporal model over the reference AR model was about 6% and the maximum RMSE improvement was about 13%. These values are respectively 12% and 20% when applying the spatio-temporal model proposed here.

D. The Variable Selection Contribution: Lasso and AIC

In this section the impact of the different variable selection methods is evaluated. Here we consider the second data set d_2 because the high number of power plants amplifies the dimensionality problem. The spatio-temporal model with the variable selection procedure based on the AIC as described in part III-B was evaluated. The extension of the model with a selection variable procedure based on Lasso regularization (part III-D) was also computed and evaluated on the same data set d_2. Figure 8 represents the dispersion of the mean value (over all prediction horizons) of the RMSE for the reference model and the spatio-temporal model resulting from the two variable selection procedures.

The figure shows that the spatio-temporal model significantly reduces prediction errors compared to the reference model (around 28% reduction in average performance). Moreover, the Lasso variable selection procedure presents lower prediction errors than the selection based on the AIC, showing that the Lasso procedure is more efficient (22% reduction in average performance).

The performance of the Lasso selection variable procedure can also be analyzed by the level of reduction of the dimensionality problem. For each of the power plants of the data set d_2, figure 9 represents the number of neighboring power plants (among the other 135) retained by the Lasso selection. In 75% of cases, the number of variables used is less than 30, while the maximum number used is 57. These numbers show that the Lasso selection variable procedure is quite successful in reducing the dimension of the problem. The results emphasize the interest for the neighboring plants of improving the quality of the PV production forecasts.

V. Conclusion

In this paper we proposed a statistical spatio-temporal model to improve short-term forecasting of photovoltaic production. The non-stationarity issue of the production series was addressed by a new stationarization process. This process demonstrated a clear improvement in terms of forecasting error reduction in comparison with a case in which raw inputs are used. The spatio-temporal model was applied to the stationarized series and showed a significant reduction in forecasting errors compared to regular forecasting techniques. The problem of high dimension data was also addressed by two different variable selection procedures for dimension reduction. The Lasso regularization applied to the spatio-temporal model presents the highest reduction for the forecasts. Moreover, we demonstrate that including the effects of meteorological variables such as wind speed in the spatio-temporal results in an additional reduction of the forecasting error level of PV production.

Further work could investigate beyond the linear modeling of the spatio-temporal data using more complex relations like polynomial estimations or splines. The integration of meteorological data could also be investigated, either as a parameter of the coefficient estimated in the spatio-temporal model, or by integrating sky images obtained by cameras or satellites. A probabilistic model that uses information on geographically distributed power plants to produce forecasts could also be investigated.

Acknowledgment

The authors would like to thank the French industrials Coruscant and Hespul for providing the PV data. They would also like to thank Prof. Philippe Blanc (MINES ParisTech) for his advice on solar radiation data treatment and clear sky models and for providing ESRA data.

References

[1] C. W. Potter, A. Archambault, and K. Westrick, "Building a smarter smart grid through better renewable energy information," in *Proceedings of Power Systems Conference and Exposition*, Seattle, WA, USA, March 2009.

[2] P. Pinson, C. Chevallier, and G. N. Kariniotakis, "Trading Wind Generation From Short-Term Probabilistic Forecasts of Wind Power," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1148–1156, Aug. 2007.

[3] X. Wu, X. Hu, S. Moura, X. Yin, and V. Pickert, "Stochastic control of smart home energy management with plug-in electric vehicle battery energy storage and photovoltaic array," *Journal of Power Sources*, vol. 333, pp. 203–212, Nov. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037877531631357X

[4] R. H. Inman, H. T. C. Pedro, and C. F. M. Coimbra, "Solar forecasting methods for renewable energy integration," *Progress in Energy and Combustion Science*, vol. 39, no. 6, pp. 535–576, Dec. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360128513000294

[5] V. Kostylev and A. Pavlovski, "Solar power forecasting performances - towards industry standards," in *Proceedings of 1st International Workshop on the Integration of Solar Power into Power Systems*, Aarhus, Denmark, October 2011.

[6] J. Shi, W.-J. Lee, Y. Liu, Y. Yang, and P. Wang, "Forecasting power output of photovoltaic system based on weather classification and support vector machine," in *Industry Applications Society Annual Meeting (IAS), 2011 IEEE*, Oct 2011, pp. 1–6.

[7] J. G. da Silva Fonseca, T. Oozeki, T. Takashima, G. Koshimizu, Y. Uchida, and K. Ogimoto, "Use of support vector regression and numerically predicted cloudiness to forecast power output of a photovoltaic power plant in kitakyushu, Japan," *Progress in Photovoltaics: Research and Applications*, vol. 20, no. 7, pp. 874–882, 2012. [Online]. Available: http://dx.doi.org/10.1002/pip.1152

[8] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy, "Predicting solar generation from weather forecasts using machine learning," in *Proceedings of IEEE International Conference on Smart Grid Communications, (IEEE SmartGridComm)*, Brussels, Belgium, October 2011. [Online]. Available: http://ieeexplore.ieee.org/

[9] O. Perpin and E. Lorenzo, "Analysis and synthesis of the variability of irradiance and {PV} power time series with the wavelet transform," *Solar Energy*, vol. 85, no. 1, pp. 188 – 197, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X10002811

[10] M. Zamo, O. Mestre, P. Arbogast, and O. Pannekoucke, "A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part i: Deterministic forecast of hourly production," *Solar Energy*, vol. 105, pp. 792 – 803, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X13005239

[11] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," *Solar Energy*, vol. 83, no. 10, pp. 1772 – 1783, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X09001364

[12] P. Bacher, H. Madsen, B. Perers, and H. A. Nielsen, "A non-parametric method for correction of global radiation observations," *Solar Energy*, vol. 88, pp. 13 – 22, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X12003891

[13] H. T. Pedro and C. F. Coimbra, "Assessment of forecasting techniques for solar power production with no exogenous inputs," *Solar Energy*, vol. 86, no. 7, pp. 2017 – 2028, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X12001429

[14] C. Monteiro, T. Santos, L. A. Fernandez-Jimenez, I. J. Ramirez-Rosado, and M. S. Terreros-Olarte, "Short-term power forecasting model for photovoltaic plants based on historical similarity," *Energies*, vol. 6, no. 5, p. 2624, 2013. [Online]. Available: http://www.mdpi.com/1996-1073/6/5/2624

[15] V. G. Berdugo, C. Chaussin, and L. D. et al., "Analog method for collaborative very-short-term forecasting of power generation from photovoltaic systems," Patent 1 154 438.

[16] E. Lorenz, D. Heinemann, and C. Kurz, "Local and regional photovoltaic power prediction for large scale grid integration: Assessment of a new algorithm for snow detection," *Progress in Photovoltaics: Research and Applications*, vol. 20, no. 6, pp. 760–769, 2012. [Online]. Available: http://dx.doi.org/10.1002/pip.1224

[17] Y. A., S. T., and S. A. et al., "Application of neural network to one-day-ahead 24 hours generating power forecasting for photovoltaic system," in *Proceedings of the International Conference on Intelligent Systems Applications to Power Systems*, Kaohsiung, Taiwan, November 2007.

[18] Y. HUANG, J. LU, and C. L. et al., "Comparative study of power forecasting methods for pv stations," in *Proceedings of the International Conference on Power System Technology*, Zhejiang, Zhejiang, China, October 2010.

[19] C. Tao, D. Shanxu, and C. Changsong, "Forecasting power output for grid-connected photovoltaic power system without using solar radiation measurement," in *Proceedings of the IEEE International Symposium on Power Electronics for Distributed Generation Systems*, Hefei, China, June 2010. [Online]. Available: http://ieeexplore.ieee.org/

[20] L. A. Fernandez-Jimenez, A. Muoz-Jimenez, A. Falces, M. Mendoza-Villena, E. Garcia-Garrido, P. M. Lara-Santillan, E. Zorzano-Alba, and P. J. Zorzano-Santamaria, "Short-term power forecasting system for photovoltaic plants," *Renewable Energy*, vol. 44, pp. 311 – 317, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960148112001516

[21] A. Mellit and A. M. Pavan, "A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected {PV} plant at trieste, italy," *Solar Energy*, vol. 84, no. 5, pp. 807 – 821, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X10000782

[22] M. Khn, C. Juhlin, H. Held, V. Bruckman, T. Tambach, T. Kempka, S. Jerez, R. Trigo, A. Sarsa, R. Lorente-Plazas, D. Pozo-Vzquez, and J. Montvez, "European geosciences union general assembly 2013, egudivision energy, resources & the environment, ere spatio-temporal complementarity between solar and wind power in the iberian peninsula," *Energy*

Procedia, vol. 40, pp. 48 – 57, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1876610213016019
[23] J. Dowell, S. Weiss, D. Hill, and D. Infield, "Short-term spatio-temporal prediction of wind speed and direction," *Wind Energy*, vol. 17, no. 12, pp. 1945–1955, 2014. [Online]. Available: http://dx.doi.org/10.1002/we.1682
[24] J. Tastu, P. Pinson, E. Kotwa, H. Madsen, and H. A. Nielsen, "Spatio-temporal analysis and modeling of short-term wind power forecast errors," *Wind Energy*, vol. 14, no. 1, pp. 43–60, 2011. [Online]. Available: http://dx.doi.org/10.1002/we.401
[25] R. Girard and D. Allard, "Spatio-temporal propagation of wind power prediction errors," *Wind Energy*, vol. 16, no. 7, pp. 999–1012, 2013. [Online]. Available: http://dx.doi.org/10.1002/we.1527
[26] M. He, L. Yang, J. Zhang, and V. Vittal, "A Spatio-Temporal Analysis Approach for Short-Term Forecast of Wind Farm Generation," *IEEE Transactions on Power Systems*, vol. 29, no. 4, pp. 1611–1622, Jul. 2014.
[27] J. Tastu, P. Pinson, P. J. Trombe, and H. Madsen, "Probabilistic Forecasts of Wind Power Generation Accounting for Geographically Dispersed Information," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 480–489, Jan. 2014.
[28] M. Sherman, *Spatial Statistics and Spatio-Temporal Data*. Wiley, 2011.
[29] J. Bosch and J. Kleissl, "Cloud motion vectors from a network of ground sensors in a solar power plant," *Solar Energy*, vol. 95, pp. 13 – 20, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X13002193
[30] M. Lave and J. Kleissl, "Cloud speed impact on solar variability scaling application to the wavelet variability model," *Solar Energy*, vol. 91, pp. 11 – 21, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X13000406
[31] S. Quesada-Ruiz, Y. Chu, J. Tovar-Pescador, H. Pedro, and C. Coimbra, "Cloud-tracking methodology for intra-hour {DNI} forecasting," *Solar Energy*, vol. 102, pp. 267 – 275, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X14000486
[32] R. Perez, S. Kivalov, J. Schlemmer, K. Hemker Jr., D. Renn, and T. E. Hoff, "Validation of short and medium term operational solar radiation forecasts in the US," *Solar Energy*, vol. 84, no. 12, pp. 2161–2172, Dec. 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X10002823
[33] C. A. Glasbey and D. J. Allcroft, "A Spatiotemporal Auto-Regressive Moving Average Model for Solar Radiation," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 57, no. 3, pp. 343–355, 2008. [Online]. Available: http://www.jstor.org/stable/20492608
[34] D. Yang, C. Gu, Z. Dong, P. Jirutitijaroen, N. Chen, and W. M. Walsh, "Solar irradiance forecasting using spatial-temporal covariance structures and time-forward kriging," *Renewable Energy*, vol. 60, pp. 235–245, Dec. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960148113002759
[35] A. Tascikaraoglu, B. Sanandaji, G. Chicco, V. Cocina, F. Spertino, O. Erdinc, N. Paterakis, and J. P. Catalao, "Compressive Spatio-Temporal Forecasting of Meteorological Quantities and Photovoltaic Power," *IEEE Transactions on Sustainable Energy*, vol. PP, no. 99, pp. 1–1, 2016.
[36] R. Dambreville, P. Blanc, J. Chanussot, and D. Boldo, "Very short term forecasting of the Global Horizontal Irradiance using a spatio-temporal autoregressive model," *Renewable Energy*, vol. 72, pp. 291–300, Dec. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S096014811400398X
[37] V. P. A. Lonij, A. E. Brooks, A. D. Cronin, M. Leuthold, and K. Koch, "Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors," *Solar Energy*, vol. 97, pp. 58–66, Nov. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X13003125
[38] J. D. Patrick, J. L. Harvill, and C. W. Hansen, "A semiparametric spatio-temporal model for solar irradiance data," *Renewable Energy*, vol. 87, Part 1, pp. 15–30, Mar. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960148115303542
[39] C. Yang, A. A. Thatte, and L. Xie, "Multitime-Scale Data-Driven Spatio-Temporal Forecast of Photovoltaic Generation," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 1, pp. 104–112, Jan. 2015.
[40] X. G. Agoua, R. Girard, and G. Kariniotakis, "Spatio-temporal models for photovoltaic power short-term forecasting," in *Solar Integration workshop 2015*, Brussels, Belgium, Oct. 2015. [Online]. Available: https://hal-mines-paristech.archives-ouvertes.fr/hal-01220321
[41] R. J. Bessa, A. Trindade, C. S. P. Silva, and V. Miranda, "Probabilistic solar power forecasting in smart grids using distributed information," *International Journal of Electrical Power & Energy Systems*, vol. 72, pp. 16–23, Nov. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0142061515000897
[42] H. C. Hottel, "A simple model for estimating the transmittance of direct solar radiation through clear atmospheres," *Solar Energy*, vol. 18, pp. 129 – 134, 1976.
[43] C. Rigollier, O. Bauer, and L. Wald, "On the clear sky model of the ESRA European Solar Radiation Atlas with respect to the heliosat method," *Solar Energy*, vol. 68, no. 1, pp. 33–48, Jan. 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X99000559
[44] N. A. Engerer and F. P. Mills, "KPV: A clear-sky index for photovoltaics," *Solar Energy*, vol. 105, pp. 679–693, Jul. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X14002151
[45] T. Robert, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.
[46] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

Xwégnon Ghislain Agoua engineer graduate of ENSAI (Ecole nationale de la statistique et de l'analyse de l'information), France, in 2014.
He is currently a PhD student at MINES ParisTech, PSL - Research University, PERSEE - Centre for Processes, Renewable Energies and Energy Systems. He is mostly interested in statistical modeling, forecasting techniques, time series analysis, spatio-temporal regression models, and their applications to photovoltaic generation modeling.

Robin Girard received a Master's degree (2004) in Computer Science and Applied Mathematics from INPG in Grenoble, France and a PhD degree (2008)in applied Mathematics from Joseph Fourier University in Grenoble.
He is currently a Research Engineer at the Centre of Energy and Processes of the Ecole des Mines de Paris. His research interests include wind and solar power forecast, optimization in planning of energy production and spatio-temporal patterns of renewable power production.

George Kariniotakis (S95-M02-SM11) was born in Athens, Greece. He received his Eng. and M.Sc. degrees from Greece in 1990 and 1992 respectively, and his Ph.D. degree from Ecole des Mines de Paris in 1996. He is currently with the Centre PERSEE of MINES ParisTech as a senior scientist and head of the Renewable Energies and Smartgrids Group. He has authored more than 200 scientific publications in journals and conferences. He has been involved as participant or coordinator in more than 40 R&D projects in the fields of renewable energies and distributed generation. Among them, he was the coordinator of some major EU projects in the field of wind power forecasting such as Anemos, Anemos.plus and SafeWind projects. His scientific interests include among others timeseries forecasting, decision making under uncertainty, modelling, management and planning of power systems.

Probabilistic Model for Spatio-Temporal Photovoltaic Power Forecasting

Xwégnon Ghislain Agoua, Robin Girard and George Kariniotakis, *Senior Member, IEEE*

***Abstract*—Photovoltaic (PV) power generation is characterized by significant variability. Accurate PV forecasts are a prerequisite to securely and economically operating electricity networks, especially in the case of large-scale penetration. In this paper, we propose a probabilistic spatio-temporal model for the PV power production that exploits production information from neighboring plants. The model provides the complete future probability density function of PV production for very short-term horizons (0-6 hours). The method is based on quantile regression and a L_1 penalization technique for automatic selection of the input variables. The proposed modeling chain is simple, making the model fast and scalable to direct on-line application. The performance of the proposed approach is evaluated using a real-world test case, with a high number of geographically distributed PV installations and by comparison with state-of-the-art probabilistic methods.**

***Index Terms*—Lasso, photovoltaic generation, probabilistic forecasts, quantile regression, reliability, sharpness, spatio-temporal**

I. Introduction

RENEWABLE Energy Sources (RES), including photovoltaic (PV) production, are being developed in many countries as a response to the need for clean energy solutions. PV production is characterized by high variability and uncertainty due to its inherent dependence on changing meteorological conditions. Variability and uncertainty are a challenge for network operators especially in systems with large-scale RES integration. Short-term forecasting of PV production is a prerequisite for the economic and secure management of power systems, reduction of reserve costs [1] and market participation of PV producers. It is also important to ensure the competitiveness of renewable energy technologies [2].

The literature features several methods to forecast PV production. Detailed reviews of the state-of-the-art are provided in [3]–[5]. They can be classified according to the forecast horizon, the available data, and the type of approach, which may be based on statistics, physics or a hybrid combination [4], [6]. Although early methods were deterministic, probabilistic approaches are increasingly popular since they provide additional information about the distribution of future production and uncertainty in the forecasts. Some of these probabilistic approaches are based on Numerical Weather Prediction (NWP) models or sky imaging and provide ensemble forecasts of the future PV generation [7]–[9]. Analog ensembles [10], regression trees [11], [12] and k-nearest neighbors (kNN) [13]–[15] are also found in the related literature on probabilistic PV forecasting. A wide range of Artificial Neural Networks (ANN) based models also exist for short-term PV power production [16]–[21]. These models have evolved from simple neural networks, to radial neural networks (more suitable for time series prediction) and more recently to deep learning methods [22], [23].

Recently the focus has shifted to improving short-term predictability of solar irradiance or PV production by considering off-site information as input to the models. Such models are commonly referred to as spatio-temporal models. Models for solar radiation forecasting are based on autoregressive models [24]–[26], geostatistical models [27]–[29] or classification models [30]. Physical models combined with NWP forecasts [31], [32] and semi-parametric models [33] can also be used for spatio-temporal forecasting of solar radiation. ANN-based models are also used for spatio-temporal solar power forecasting [34]–[36]. For the case of PV power forecasting, most spatio-temporal models use off-site meteorological measurements and neighboring site information to improve the forecast for the site of interest. These models are autoregressive models [37]–[39] or classification models [30]. The classification model proposed in [40] handles data acquisition and privacy issues.

The spatial-temporal models listed above provide deterministic forecasts. In contrast to day-ahead forecasting, very few spatial-temporal models for short-term probabilistic forecasting feature in the literature. These models are based on regression trees [12], the kNN method [13], the combination of a vectorial autoregressive model and gradient boosting [41], multivariate predictive distributions, [42] and Gaussian random fields [43].

One of the biggest challenges in spatio-temporal forecasting is to propose models that are able to handle the dimensionality issues raised by the large amount of explanatory variables. This is a prerequisite for the models so that they can generalize efficiently when they “see” new data. In this paper, we propose a short-term probabilistic forecasting methodology that exploits information available from a large number of PV installations and includes such a process for the automated selection of explanatory variables. The aim is to improve predictability in the time range of 0-6 hours.

In a previous paper [44], we proposed a deterministic

This work was carried out within the research project entitled ”Improvement of PV power forecasting and predictive management including storage solutions”, supported by the company Coruscant SA in the frame of its participation to a tender of the French Energy Regulator CRE for the development of PV plants above 250 kWc.

The authors are with MINES ParisTech, PSL - Research University, PERSEE - Centre for Processes, Renewable Energies and Energy Systems CS 10207 rue Claude Daunesse, 06904 Sophia Antipolis Cedex, France. (e-mails: xwegnon.agoua, robin.girard, georges.kariniotakis each with @mines-paristech.fr)

spatio-temporal model for PV power forecasting. This paper makes several additional contributions, compared to both the previous paper and to the related literature on probabilistic PV forecasting. These contributions are: 1) the use of off-site data (neighboring production data) to generate probabilistic forecasts with improved properties and performance compared to the standard approach without off-site data 2) the introduction of an automated variable selection procedure for the PV power forecasts inside the model, which avoids the dimensionality problem 3) the integration of NWP forecasts as explanatory variables despite the short-term frame; this allows us to assess the effect of the NWP variables on the probabilistic forecasts. In probabilistic models in the literature [35], [41], [45] the variable selection is done either empirically or with the variable importance criteria for decision trees; 4) two probabilistic models chosen from the most efficient ones in the literature are used as references for our model evaluation.

In contrast to methods proposed in the literature for this time range, our approach is not based on information from sky cameras or satellite images. It uses available data from off-site PV plants. The model uses the transfer of spatial and temporal information within the network of power plants to capture the meteorological perturbations. Unlike in [46], these spatial and temporal correlations between power plants are useful for short-term horizons as they offer good spatial and temporal resolution.

Moreover, in the proposed model we use NWP forecasts as additional explanatory input, which is not standard practice in the literature for models developed especially for this time scale. NWPs are a standard input for models that address day-ahead or longer horizons. Such models also cover horizons shorter than 6h. It is however widely recognized in renewable generation forecasting that, for the first hours, NWPs do not obtain performance improvements w.r.t. persistence. To achieve such improvements, it is necessary to include recent measurements, such as input. This observation, coupled with the need for frequent forecast updates (i.e. every 30 minutes) for applications like intraday trading, have motivated the development of purely statistical models based only on measurements. The standard spatiotemporal approach comes into this category. However, we have observed that including NWPs in this family of models can provide information on weather condition trends for horizons around 6 hours ahead. In cases like wind prediction, this inclusion brings improvements w.r.t. persistence, amounting to double the improvements without NWPs for 6 hours ahead [3]. This inclusion of NWPs in the family of short-term models that operate with very frequent updates goes against the general tendency in the literature. However, this approach allows for better coupling between very short-term spatial-temporal forecasts and standard day-ahead forecasts since NWPs reduce? the mismatch between two types of forecast that can be confusing for end-users. The most relevant NWP forecasts were selected by measuring their dependence on production data.

The generation of the probabilistic forecast was done using a quantile regression scheme. Therefore the variable selection process using L1-penalization that we propose in this paper is different from the one used in [44] as it is done in a L1-optimization model (quantile regression). We have also added the constraints of different parameters for each quantile and each time step.

This paper puts forward the non-stationary characteristic of PV power series. The spatio-temporal approach exploits correlations in the data from geographically dispersed sites. It is thus necessary to efficiently eliminate the deterministic effect that is related to the course of the sun. This issue has been treated by applying an advanced stationarization process to the production series as proposed in [44]. The evaluation is carried out on a real world test case of 185 PV installations in France. This test case is a good illustration of a system with high penetration of renewable energy and also allows us to demonstrate the full potential of a spatio-temporal approach for probabilistic forecasting when considering a highly dimensional problem.

The paper is structured as follows: the data and an analysis of the spatio-temporal correlations are presented in section II. The proposed spatio-temporal probabilistic models are presented in section III. The evaluation and analysis of the performances of the forecasts are treated in section IV. Finally, the conclusions of the study are discussed in section V.

II. The Interest of Spatio-Temporal modeling

The aim of this section is to illustrate the interest and potential of spatio-temporal modeling in short-term PV power forecasting. A real-world case study is introduced followed by an analysis illustrating that useful information can be extracted from geographically distributed PV installations.

A. Data

The data set considered here, denoted as d, comprises the output of 905 PV power inverters in a mid-west region of France with peak power ranging from 3.2 kWc to 58 kWc. This number of inverters corresponds to 185 different PV power plants (set of inverters at the same location). The distance between them varies from 1 km to 230 km. The data relate to the period from November 2014 to March 2016 with a 15 min temporal resolution. The locations of the power plants are represented in Figure 1. After data cleaning, 136 power plants were retained. In the following, the power plants are labeled $P_{i,1\leq i\leq 136}$. The production series have been stationarized employing the same procedure as that proposed in [44], which is used thereafter.

B. Spatial-Temporal Correlation

The existence of a spatial-temporal pattern is evaluated by an analysis of the correlation between the stationarized production series. In these series, the effects of the sun position on cross-correlations are absent.

Figure 2 presents the histograms of the lagged cross-correlation values for three classes of distance. For a couple of stationarized series s_1, s_2 for example, the value of the correlation is obtained by evaluating the correlation between s_1 and the lagged series $lag(s_2, k)$. The maximum lag used is the maximum horizon considered for the forecasts, here

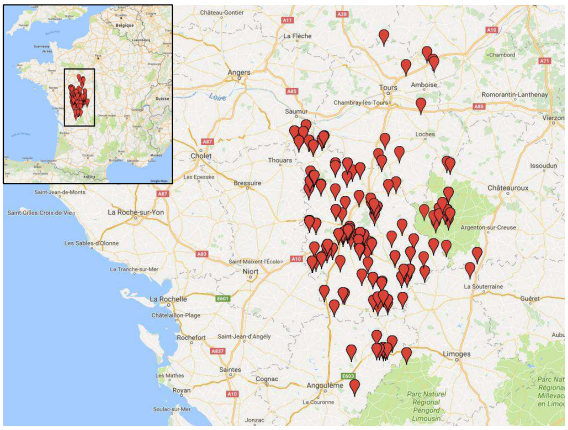


Fig. 1. The power plants of the test case *d*.

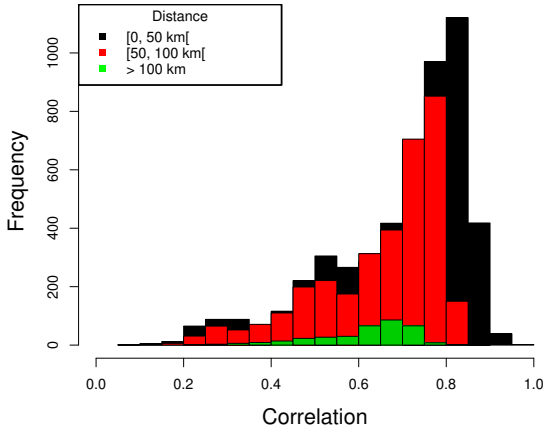


Fig. 2. Histogram by class of distance of the cross-correlation between the lagged production series after stationarization.

6 hours. The same evaluation is made by switching s_1 and s_2. The value retained is the maximum value. The correlation values are quite high, concentrated in the interval $[0.4 - 0.8]$ and, decrease with distance. Therefore, we can assume that these significant correlation values describe a spatial transfer of information between the power plants. This transfer, mainly due to cloud movements, can be used to anticipate meteorological perturbations and to improve the forecasts using production data only.

From a temporal point of view, Figure 3 presents the time lags for which the cross correlations described above reach their maximum. The idea behind this analysis is to determine the temporal limit of the propagation of correlation between the power plants. The figure shows that in the majority of cases, the maximum correlation is obtained for 15 minutes, but for some power plant couples it can be reached for 4 hours or more. The models proposed in the next section aim to exploit this spatial and temporal information.

III. Probabilistic Models for Spatial-Temporal PV Forecasting

We consider here two models chosen from the most efficient probabilistic forecasting methods in the literature as reference

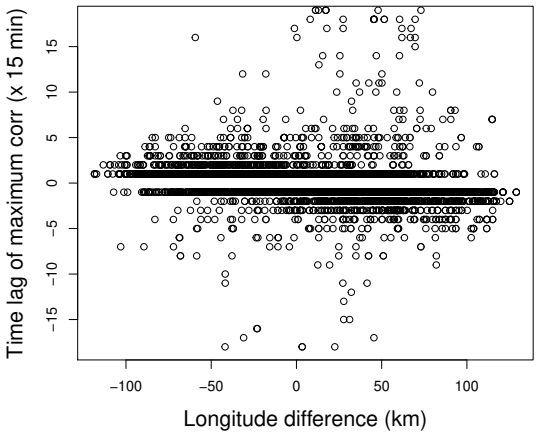


Fig. 3. Time lags for which the correlation between each power plant couple is at a maximum. One point per couple.

models: Kernel Density Estimation (KDE) and the Extreme Learning Machine (ELM) method proposed in [47]. The KDE model is a well-known method used to evaluate the future density of PV power [48], [49]. An information-based criterion is proposed to select the input variables. These two models respectively use only "local" on-site information from the PV installation the power of which is predicted. These models can thus be used to evaluate the contribution of the proposed spatial-temporal probabilistic approach. The advanced method we propose is based on the quantile regression approach, which we adapted for high dimensionality problems. We describe below the principle of these methods, the conditions of their implementation and the modifications proposed to improve their efficiency.

A. Kernel Density Estimation

KDE is a non-parametric approach for density estimation [50]–[52]. This method reduces the level of estimation errors compared to parametric approaches as there is no hypothesis on the underlying distribution. The minimization problem consists in providing an estimation of the density f of a random variable X. The n-dimensional multivariate kernel-estimator expression is:

$$\hat{f}(x) = \frac{1}{N|\mathbf{H}|} \sum_{i=1}^{N} K\left(\mathbf{H^{-1}}(x - x_i)\right) \qquad (1)$$

where x is the point where the evaluation is made, $x_i, i = 1 \dots N$ the data. $\mathbf{H}$ is a $n \times n$ matrix called the bandwidth matrix, which controls the smoothing and $|\mathbf{H}|$ its determinant. K is the kernel function. The kernel function K and the bandwidth matrix $\mathbf{H}$ are the two parameters which have to be determined to apply the KDE. The impact of the kernel function on the estimation quality is low and a standard Gaussian kernel implies a high computation cost [53]. In this paper, we have chosen to use the Epanechnikov kernel function:

$$K(u) = \frac{3}{4\sqrt{5}}(1 - u^2/5) \qquad u \in [-1, 1]. \qquad (2)$$

The multivariate version of this kernel function is the product $K(u) = \prod_{j=1}^{n} K(u_j)$. The smoothing matrix $\mathbf{H}$ is the most important parameter as it has a strong influence on the quality of the estimation [52]. Several approaches are possible for selecting the appropriate bandwidth matrix [54], [55]. In this paper, we use the smoothed cross validation technique [56]. Moreover, as the PV production data are positive and bounded, we adapt the general method to our needs by using boundary correction as proposed in the literature [52], [53].

Let $Y \in \mathbb{R}^p$ be the random variable, the realizations of which are the production of a power plant and $X \in \mathbb{R}^p$ the explanatory variables. The objective is to compute the probability density function of the conditioned random variable $Y_{t+k}|X_t$ where t is the time the prediction is made and k the horizon. This density function is obtained by:

$$f_{Y_{t+k}|X_t} = \frac{f_{Y_{t+k},X_t}}{f_{X_t}} \tag{3}$$

The forecast probability density function is :

$$\hat{f}_{Y_{t+k}|X_t} = \frac{1}{|\mathbf{H}|} \sum_{i=1}^{N} w(x, x_i) K(\mathbf{H^{-1}}(y - y_i)) \tag{4}$$

where

$$w(x, xi) = \frac{K\left(\mathbf{H^{-1}}(\mathbf{x} - \mathbf{x_i})\right)}{\sum_{j=1}^{N} K\left(\mathbf{H^{-1}}(\mathbf{x} - \mathbf{x_j})\right)}. \tag{5}$$

The mutual information criterion [57] has been used for choosing the inputs of the KDE model. It is a measure of "distance" (Kullbac-Leiber) between two probability density functions. Here this criterion is used to evaluate the distance between the probability density functions of the PV production and those of the meteorological variables. This evaluation is carried out for each meteorological variable. The values of the information criterion are used to classify the meteorological variables according to their proximity to the probability density of the production, following which the most relevant inputs are selected for the KDE model. Let X be the random variable representing a meteorological variable and Y that of the PV production. If f_X and f_Y denote their respective probability density functions, the average mutual information between Y and X is:

$$I(X, Y) = \int_Y \int_X f_{X,Y} log\left(\frac{f_{X,Y}}{f_X . f_Y}\right) dxdy. \tag{6}$$

A null value of this criterion indicates that the variables X and Y are independent.

B. ELM-Based Estimation

The Extreme Learning Machine method is a non-parametric density forecasting method presented and developed in [47]. It is based on feed forward neural networks but does not require tuning for the hidden layer parameters (the parameters are randomly selected making it fast). The proposed configuration of the ELM-based estimation model for this paper is the following:

- one model is tuned for each power plant;
- the predictors are the past solar observations of the site of interest with respective time lags and the NWP variables selected by the mutual information criterion presented in subsection III-A;
- the results are generated as predicted quantiles for a 6-hour horizon with 15-min time-steps; the forecasting model is updated monthly
- the non-crossing quantile constraint has been applied.

C. The proposed model: the QR-Lasso

The aim of quantile regression [58] is to provide an estimation of the following cumulative distribution function:

$$F(y|X = x) = \mathbb{P}(Y \leq y|X = x) = \mathbb{E}(\mathbf{1}_{Y \leq y}|X = x). \tag{7}$$

This estimation is made by providing information about a significant number of points describing the distribution known as the quantiles. For a continuous probability distribution function, the quantile of level $\alpha \in (0, 1)$, $\hat{q}_\alpha(Y|X)$ is

$$q_\alpha(Y|X) = \mathbb{F}^{-1}(\alpha) = \inf\{y, \mathbb{F}(y|X = x) \geq \alpha\}. \tag{8}$$

This quantile is a solution of the minimization problem:

$$q_\alpha(Y) = \arg\min_g \ \mathbb{E}[\rho_\alpha(Y - g(X))] \tag{9}$$

where $\rho_\alpha(u) = u\ (\alpha - \mathbf{1}_{\{u<0\}})$ is the loss function called the pinball loss. The problem (9) can be generalized to the conditional quantile of level α as :

$$q_\alpha(Y|X) = \arg\min_g \ \mathbb{E}[\rho_\alpha(Y - g(X))|X = x]. \tag{10}$$

When assuming a linear relation between Y and X, the estimation can be made on the observation set using:

$$\hat{q}_\alpha = \arg\min_\beta \ \sum_{i=1}^{n} \rho_\alpha(Y_i - X_i'\beta). \tag{11}$$

The equation (11) does not have an explicit solution; the solution has been found numerically, despite the fact that the pinball loss is neither differentiable in 0 nor strictly convex. The problem is then transformed into a linear optimization problem solved by the simplex algorithm (for small samples) or interior point method (for large samples) [59]. The condition of non-crossing quantiles has been taken into account. In practice, this condition is implemented by adding some constraints to the problem that specify that each quantile value is positive, lower than the next value, and lower than the maximum.

D. Adaptation for the High Dimensional Input

Spatio-temporal forecasting, when considering numerous PV plants, raises the question of choosing the appropriate input variables for the models involved. Possible inputs include the production series of the predicted site and the neighboring sites, and the lagged version of these series. The NWP meteorological data can also be integrated into the models. Given that the PV plants may cover a large area of tens of kilometers, several points of the NWP grid can be considered. The above information represents a high amount of potential inputs for the forecasting models and raises the issue of their

dimensionality and over-fitting. We propose here a variable selection process, which we have adapted to the quantile regression.

The method is based on a direct L1-penalization called the Least Absolute Shrinkage and Selection Operator (LASSO) [60]. This penalization is applied directly to the estimation process of the quantile regression model. The estimator is defined for the quantile of level α as:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \mathbb{E}[\rho_\alpha(Y - X'\beta)] + \lambda \|\beta\|_1 \right\}. \qquad (12)$$

This penalization also helps to avoid over-fitting in the model. For values of the penalization parameter λ that are high enough, some coefficients β are set to zero, thus allowing a selection to be made between the input variables.

The problem (12) can be reformulated as

$$\underset{\beta}{\operatorname{argmin}} \left\{ \mathbb{E}[\rho_\alpha(Y - X'\beta)] \right\} \quad \text{st} \quad \|\beta\|_1 \leq \hat{R}_\lambda. \qquad (13)$$

The equations (12) and (13) are equivalent. The latter is the most common form of optimization problem. The Lasso is also more suitable for operational models as its run-time is lower than those of other methods, like L2-penalization, gradient boosting or the stepwise process. The L1 loss ensures a variable selection that makes the model perform well with a reasonable number of variables and is stricter than the L2 loss. More details about the regularization paths can be found in the literature [60].

This modification of the standard quantile regression proposed here is essential in the spatial-temporal paradigm due to the high number of potential inputs. The new quantile regression model including this LASSO variable selection procedure is hereafter called QR-Lasso. The performance of this model is evaluated against the reference KDE model.

IV. Evaluation of the Forecasts

The performance of the proposed model is evaluated for horizons up to 6 hours ahead with a 15-min time step. It is considered that forecasts are updated every 15 minutes. This performance is compared to the reference model. The NWP data were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF). These forecasts are updated 4 times a day and are given with a 3-hourly temporal resolution. 15-min values were obtained through interpolation. The models were developed using the software R [61] with the packages *quantreg, glmnet, elmNN* and some specific functions developed for this study (matrice estimations for KDE, penalized quantile, etc). The probabilistic (distribution) forecasts will be characterized by their quantiles. The last third of the data set (around 5 months) was used as a testing set. The QR-Lasso model is updated for each time step for each decile. A single update step took approximately 5 seconds (on a 12 GB RAM computer).

A. Variables Selection Through Mutual Information

The variable selection procedure allows us to choose the relevant variables from numerous available inputs, measurements and NWPs, for the case where a large number of PV plants are involved in the spatio-temporal modeling.

TABLE I
Normalized mutual information values for 4 PV plants of the data set computed between PV production and NWP variables.

NWP Variables	Mutual information (%)			
	P_1	P_2	P_3	P_4
Temperature (T)	72.38	70.74	80.79	78.22
Relative Humidity (RH)	70.56	74.56	74.83	71.27
Wind Direction (WD)	25.11	21.88	26.11	21.86
Wind Speed (WS)	6.18	5.78	6.57	5.46
Precipitation	0.18	0.48	0.26	0.27

In Table I, we present the normalized mutual information between the PV production of four PV plants and the NWP variables. The mutual information has not been computed for the top net solar radiation variable since this variable is the main driver of the PV production process. The table shows that the two most important variables are the temperature and the humidity followed by the wind direction. The importance of the precipitation level is the lowest.

Using this process, we select the following input NWP variables for the reference KDE model, which will then be multivariate:

- the top net solar radiation (TSR),
- the temperature (T),
- the relative humidity (RH),
- the wind direction (WD).

B. Analysis of the Lasso Variable Selection

The potential input of the proposed spatio-temporal QR-Lasso model for each PV plant includes the measured power of the other 135 plants, the lagged values of these measurements and finally the NWP variables corresponding to the 4 surrounding NWP grid points. The subset of NWP variables selected by the mutual information criterion is considered. Then for each PV plant $P_i, i = 1, \ldots, 136$, the number of total potential inputs is $819 = 4$ (NWP forecasts) $+ (136 \times 6) - 1$ (6 lags). Figure 4 presents the number of non-null coefficients related to PV power inputs selected by the QR-Lasso model for each decile. The maximum number of variables selected is 320, showing that the selection process is efficient. The performance of the variable selection process is also evaluated by examining the geographical position of the selected PV plants with respect to the position of the PV plant of interest. Figure 5 presents the positions of the selected PV plants when forecasting the median of the production of two PV plants for a 6-hour horizon. The numbers of PV plants selected from the 136 initial plants are 33 (left) and 43 (right). For the two power plants of interest presented, the choice of neighboring plants is quite homogeneous. Some power plants close to the plant of interest do not provide more knowledge to the model (in terms of error reduction) than the selected ones. To ensure the parsimony, these plants are thus not selected by the model.

C. The Performance of the QR-Lasso Spatial-Temporal Approach

The quality of the probabilistic forecasts produced with the selected variables can be evaluated with several criteria [62].

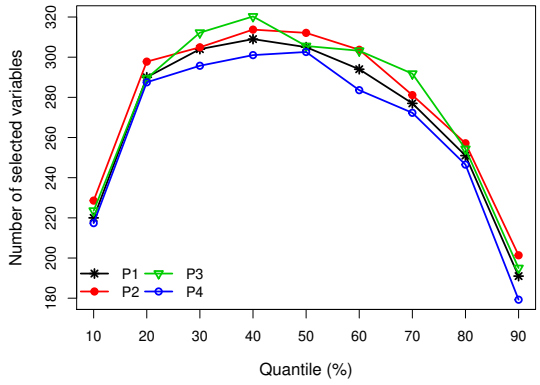


Fig. 4. Number of non-null coefficients computed by the QR-Lasso model by decile level (the meteorological coefficients are not included). Four PV plants $(P_1 - P_4)$ from the test case are represented and each line corresponds to one power plant. The horizon is 6 hours.

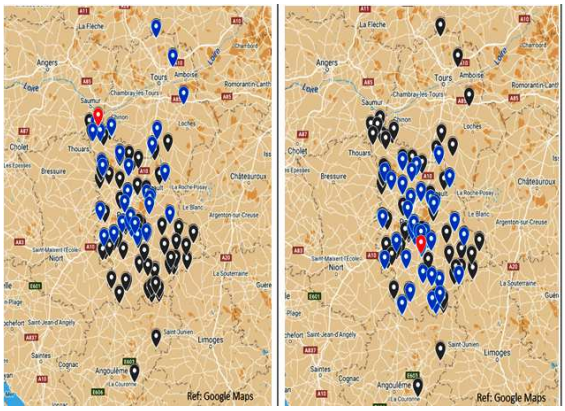


Fig. 5. The power plants selected by the QR-Lasso model (in blue) when forecasting the median of the future production for a 6-hour horizon for two PV plants (in red).

Here we limit our study to the two most common criteria, i.e. reliability and sharpness, and a combined criterion that represents overall skill.

Reliability is a criterion that describes the ability of the probabilistic forecast model to match the expected observation frequencies. For $\hat{q}^{\alpha}$ a predicted quantile of the level α, the reliability criterion verifies whether the associated proportion of data observed under this quantile is equal to the expected α. To evaluate the reliability of the full predictive density for each of our models, we evaluate the reliability for quantile q^{10} to q^{90} with an increasing step of 10. The quantiles can be interpreted as estimated coverage rates, which then allows us to evaluate the reliability through increasing coverage rate analysis. Formally, reliability can be defined for the quantile level (or coverage rate) α and horizon h as:

$$Rel_h^{\alpha} = \alpha - \frac{1}{N}\sum_{t=1}^{N} \mathbf{1}\{y_{t+h|t} \leq q_{t+h|t}^{\alpha}\} \qquad (14)$$

where t is the time when the prediction is made.

Figure 6 presents for these quantiles (predictive intervals) the observed deviation from the nominal coverage for 3h horizons. Perfect reliability means that there is no deviation, and this is represented by the horizontal black line on the figure. The deviation of the ELM model has also been plotted. Two deviations are plotted for the KDE representing two different smoothing matrices $H1, H2$. These smoothing matrices are respectively the results of the unbiased cross validation and the smooth cross-validation method (see [56] for more details on the functions and minimization process). The figure shows that, depending on the values of the smoothing matrices, the predictive densities can be either over-estimated (red line for KDE/H1) or under-estimated (green line for KDE/H2). The values of the deviation for the QR-Lasso model do not exceed $\pm 0.5\%$ and are lower than those of the KDE model. Moreover, the deviations of the nominal reliability values of the QR-Lasso model are lower than those of the ELM model (purple line).

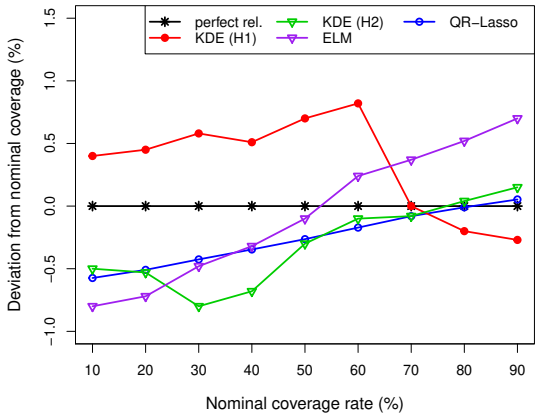


Fig. 6. Reliability of the predictive densities for 3-hour horizons. The red and green lines represent the KDE model with two cases of smoothing matrix. The purple and blue lines represent the ELM and QR-Lasso models. The values are given as average over the forecasting length on the testing set.

Sharpness is a complementary analysis tool to reliability and can be used to evaluate the concentration of the predictive distributions. Sharpness is often evaluated by taking the average width of centered prediction intervals. After defining a set of prediction intervals, the distances between the two boundaries (length of interval) are computed and all of these distance values are averaged over the evaluation set [63]. The more concentrated the predictive distributions are, the sharper the forecasts, which is preferable. The sharpness is defined for a quantile level α for the horizon h as:

$$sharp_h^{\alpha} = \frac{1}{N}\sum_{t=1}^{N}\left(\hat{q}_{t+h|t}^{1-\frac{\alpha}{2}} - \hat{q}_{t+h|t}^{\frac{\alpha}{2}}\right). \qquad (15)$$

The evaluation of sharpness (as inter-quantile interval lengths) is presented in Figure 7 for the same prediction intervals as on the reliability plot. The maximum observed production is used to normalize the interval lengths. The figure shows that the average interval length increases with the nominal quantile (coverage rate). The values range from 3% to 58% and are lower than those observed for the KDE model for quantiles up to q^{60}. The inter-quantile interval lengths in the QR-Lasso

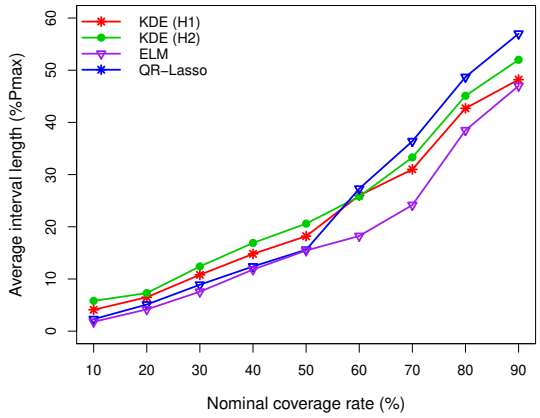


Fig. 7. Sharpness evaluation of the predictive densities for 3-hour horizons. The red and green lines are the interval lengths of the KDE model for two cases of smoothing matrix. The purple and blue lines are those of the ELM and QR-Lasso models.

TABLE II
CRPS OF THE SPATIO-TEMPORAL QR-LASSO MODEL AND THE REFERENCE KDE MODEL FOR FIVE POWER PLANTS.

Power plant	CRPS (% Pmax)		
	KDE	ELM	QR-Lasso
P1	7.20	7.00	5.20
P2	8.32	7.42	4.14
P3	7.65	7.57	5.23
P4	8.67	7.68	5.14
P5	8.32	7.08	4.38

model are higher than those in the KDE model for upper quantiles.

D. Overall Probabilistic Forecasting Skill

In addition to reliability and sharpness, an overall evaluation criterion for probabilistic forecasts is the Continuous Rank Probability Score (CRPS). The CRPS evaluates the entire predictive distribution and can be seen as a criterion that combines reliability with sharpness. It is defined for a cumulative distribution function F and its observation y as

$$crps(F, y) = \int_{-\infty}^{\infty} \left(F(x) - H(x - y)\right)^2 dx \qquad (16)$$

where $H(x)$ is the Heaviside function, whose value is 0 for strictly negative x and 1 for positive x. The CRPS can be interpreted as a measure of distance between the observed cumulative distribution function and the forecasted one [64].

The average CRPS values over the 6-hour forecast horizon are presented in Table II for the reference KDE model with matrix $H1$ (the one with the best performances), the ELM model and the QR-Lasso model for five power plants in the test case. The CRPS values are lowest for the QR-Lasso model, confirming the improved forecasting performance compared to the KDE and ELM models. This analysis provides one more illustration of how well the QR-Lasso model performs. Moreover, the CRPS values are lower than those of the set of models presented in the review [11].

Figure 8 represents the predictive densities as a set of prediction intervals for six days of the testing set for the power plant P_1. The QR-Lasso model performs quite well to anticipate the variations in production. The length of the quantile intervals is low compared to what can be observed in the literature [47]. Overall, most of the variations for days characterized by high production variability are predicted quite efficiently.

V. CONCLUSION

In this paper we proposed a probabilistic spatio-temporal model that exploits off-site information to improve short-term forecasting of photovoltaic production. The model is based on quantile regression that has been adapted to integrate a LASSO variable selection process. This makes it particularly suitable for situations where data from a high number of PV installations is available since it is able to handle directly the resulting high dimensionality and over-fitting issues that characterize the use of such large amounts of data.

A real world test case with a high number of PV plants was used to illustrate the full potential of the probabilistic spatio-temporal approach. The evaluation of the predicted densities was carried out employing evaluation criteria that are widely used for probabilistic forecasts. The model shows significant performance improvement compared to the reference kernel density model and to models in the literature. The probabilistic spatio-temporal model appears to be an interesting technique that performs well to predict the future densities of PV generation.

The perspectives of this work include the possibility of extending the forecast horizons to daily forecasts. This would emphasize the trade-off significance between NWP forecasts and historic data measurements, and help to define the horizon limit where the spatial-temporal approach does not permit improvement. Another improvement could be to consider a time-varying smoothing matrix to estimate the kernel density, which is a key parameter of this model.

ACKNOWLEDGMENT

The authors would like to thank the French industrial Hespul for providing the PV data as well as the European Center for Medium-Range Weather Forecasts (ECMWF) for providing the NWP data.

REFERENCES

[1] C. W. Potter, A. Archambault, and K. Westrick, "Building a smarter smart grid through better renewable energy information," in *2009 IEEE/PES Power Systems Conference and Exposition*, Mar. 2009, pp. 1–5.

[2] P. Pinson, C. Chevallier, and G. N. Kariniotakis, "Trading Wind Generation From Short-Term Probabilistic Forecasts of Wind Power," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1148–1156, Aug. 2007.

[3] G. Kariniotakis, *Renewable Energy Forecasting: From Models to Applications*. Woodhead Publishing, Jun. 2017.

[4] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, pp. 78–111, Oct. 2016.

[5] S. Sobri, S. Koohi-Kamali, and N. A. Rahim, "Solar photovoltaic generation forecasting methods: A review," *Energy Conversion and Management*, vol. 156, pp. 459–497, Jan. 2018.

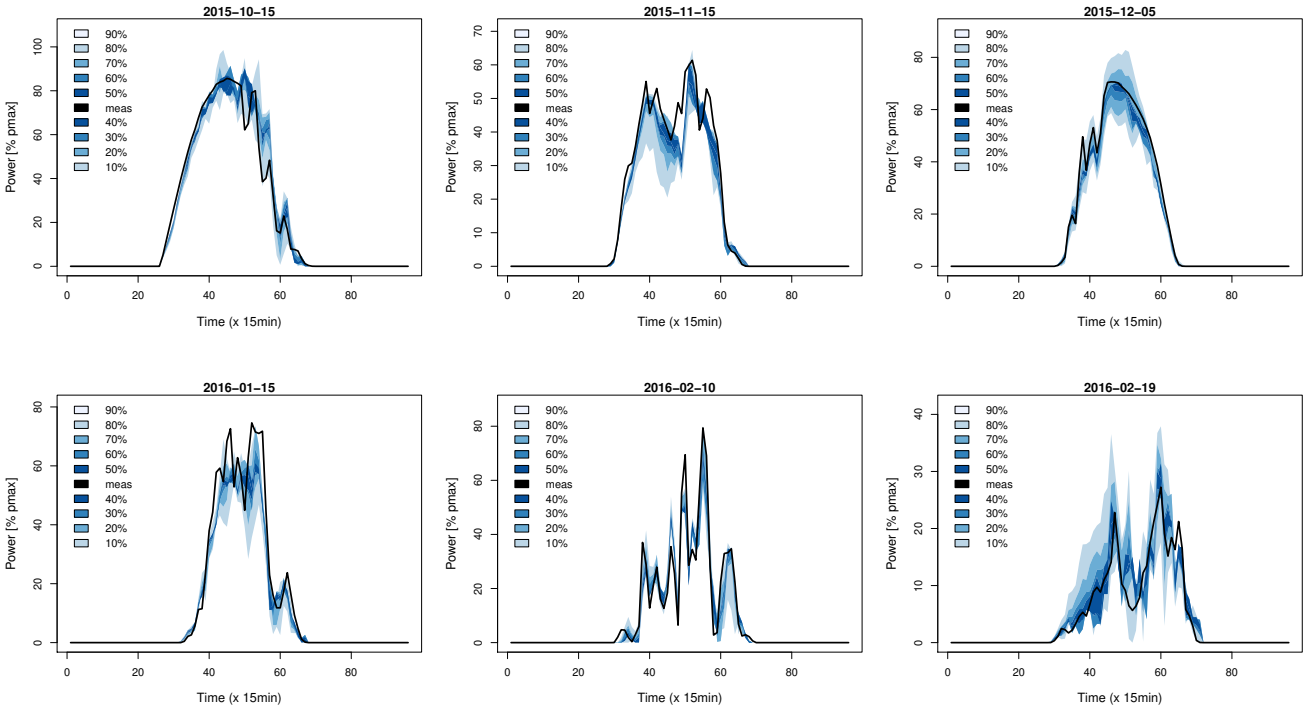


Fig. 8. Examples of predictive densities forecasted with the QR-Lasso model for 6 days in different months of the year. The quantiles are plotted from 10% to 90% with a 10% step. The time step is 15 min and the model is updated every 15 min.

[6] M. Zamo, O. Mestre, P. Arbogast, and O. Pannekoucke, "A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: Deterministic forecast of hourly production," *Solar Energy*, vol. 105, pp. 792–803, Jul. 2014.

[7] S. Sperati, S. Alessandrini, and L. Delle Monache, "An application of the ecmwf ensemble prediction system for short-term solar power forecasting," *Solar Energy*, vol. 133, pp. 437–450, Aug. 2016.

[8] M. Pierro, F. Bucci, M. De Felice, E. Maggioni, D. Moser, A. Perotto, F. Spada, and C. Cornaro, "Multi-Model Ensemble for day ahead prediction of photovoltaic power generation," *Solar Energy*, vol. 134, pp. 132–146, Sep. 2016.

[9] Y. Liu, S. Shimada, J. Yoshino, T. Kobayashi, Y. Miwa, and K. Furuta, "Ensemble forecasting of solar irradiance by applying a mesoscale meteorological model," *Solar Energy*, vol. 136, pp. 597–605, Oct. 2016.

[10] S. Alessandrini, L. Delle Monache, S. Sperati, and G. Cervone, "An analog ensemble for short-term probabilistic solar power forecast," *Applied Energy*, vol. 157, pp. 95–110, Nov. 2015.

[11] M. Zamo, O. Mestre, P. Arbogast, and O. Pannekoucke, "A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production," *Solar Energy*, vol. 105, pp. 804–816, Jul. 2014.

[12] G. I. Nagy, G. Barta, S. Kazi, G. Borbély, and G. Simon, "GEFCom2014: Probabilistic solar and wind power forecasting using a generalized additive tree ensemble approach," *International Journal of Forecasting*, vol. 32, no. 3, pp. 1087–1093, Jul. 2016.

[13] J. Huang and M. Perry, "A semi-empirical approach using gradient boosting and -nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting," *International Journal of Forecasting*, vol. 32, no. 3, pp. 1081–1086, Jul. 2016.

[14] E. Scolari, F. Sossan, and M. Paolone, "Irradiance prediction intervals for PV stochastic generation in microgrid applications," *Solar Energy*, vol. 139, pp. 116–129, Dec. 2016.

[15] Y. Chu and C. F. M. Coimbra, "Short-term probabilistic forecasts for Direct Normal Irradiance," *Renewable Energy*, vol. 101, pp. 526–536, Feb. 2017.

[16] A. Yona, T. Senjyu, A. Y. Saber, T. Funabashi, H. Sekine, and C. H. Kim, "Application of Neural Network to One-Day-Ahead 24 hours Generating Power Forecasting for Photovoltaic System," in *2007 International Conference on Intelligent Systems Applications to Power Systems*, Nov. 2007, pp. 1–6.

[17] Y. Huang, J. Lu, C. Liu, X. Xu, W. Wang, and X. Zhou, "Comparative study of power forecasting methods for PV stations," in *2010 International Conference on Power System Technology*, Oct. 2010, pp. 1–6.

[18] A. Mellit and A. Pavan, "A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy," *Solar Energy*, vol. 84, no. 5, pp. 807–821, 2010.

[19] A. Dolara, F. Grimaccia, S. Leva, M. Mussetta, and E. Ogliari, "A Physical Hybrid Artificial Neural Network for Short Term Forecasting of PV Plant Power Output," *Energies*, vol. 8, no. 2, pp. 1138–1153, Feb. 2015.

[20] H. Hassani and E. S. Silva, "Forecasting with Big Data: A Review," *Annals of Data Science*, vol. 2, no. 1, pp. 5–19, Mar. 2015.

[21] S. I. Vagropoulos, E. G. Kardakos, C. K. Simoglou, A. G. Bakirtzis, and J. P. S. Catalão, "ANN-based scenario generation methodology for stochastic variables of electric power systems," *Electric Power Systems Research*, vol. 134, pp. 9–18, May 2016.

[22] C. Y. Zhang, C. L. P. Chen, M. Gan, and L. Chen, "Predictive Deep Boltzmann Machine for Multiperiod Wind Speed Forecasting," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 4, pp. 1416–1425, Oct. 2015.

[23] M. Khodayar, O. Kaynak, and M. E. Khodayar, "Rough Deep Neural Architecture for Short-Term Wind Speed Forecasting," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 6, pp. 2770–2779, Dec. 2017.

[24] C. A. Glasbey and D. J. Allcroft, "A spatiotemporal auto-regressive moving average model for solar radiation," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 57, no. 3, pp. 343–355, Jun. 2008.

[25] R. Dambreville, P. Blanc, J. Chanussot, and D. Boldo, "Very short term forecasting of the Global Horizontal Irradiance using a spatio-temporal autoregressive model," *Renewable Energy*, vol. 72, pp. 291–300, Dec. 2014.

[26] R. Amaro e Silva and M. C. Brito, "Impact of network layout and time resolution on spatio-temporal solar forecasting," *Solar Energy*, vol. 163, pp. 329–337, Mar. 2018.

[27] M. Journée and C. Bertrand, "Improving the spatio-temporal distribution of surface solar radiation data by merging ground and satellite

measurements," *Remote Sensing of Environment*, vol. 114, no. 11, pp. 2692–2704, Nov. 2010.
[28] D. Yang, C. Gu, Z. Dong, P. Jirutitijaroen, N. Chen, and W. M. Walsh, "Solar irradiance forecasting using spatial-temporal covariance structures and time-forward kriging," *Renewable Energy*, vol. 60, pp. 235–245, Dec. 2013.
[29] S. Quesada-Ruiz, Y. Chu, J. Tovar-Pescador, H. T. C. Pedro, and C. F. M. Coimbra, "Cloud-tracking methodology for intra-hour DNI forecasting," *Solar Energy*, vol. 102, pp. 267–275, Apr. 2014.
[30] M. Lazzaroni, S. Ferrari, V. Piuri, A. Salman, L. Cristaldi, and M. Faifer, "Models for solar radiation prediction based on different measurement sites," *Measurement*, vol. 63, pp. 346–363, Mar. 2015.
[31] E. Lorenz, D. Heinemann, and C. Kurz, "Local and regional photovoltaic power prediction for large scale grid integration: Assessment of a new algorithm for snow detection," *Progress in Photovoltaics: Research and Applications*, vol. 20, no. 6, pp. 760–769, Sep. 2012.
[32] E. Lorenz, J. Kühnert, D. Heinemann, K. P. Nielsen, J. Remund, and S. C. Müller, "Comparison of global horizontal irradiance forecasts based on numerical weather prediction models with different spatio-temporal resolutions," *Progress in Photovoltaics: Research and Applications*, vol. 24, no. 12, pp. 1626–1640, Dec. 2016.
[33] J. D. Patrick, J. L. Harvill, and C. W. Hansen, "A semiparametric spatio-temporal model for solar irradiance data," *Renewable Energy*, vol. 87, Part 1, pp. 15–30, Mar. 2016.
[34] F.-V. Gutierrez-Corea, M.-A. Manso-Callejo, M.-P. Moreno-Regidor, and M.-T. Manrique-Sancho, "Forecasting short-term solar irradiance based on artificial neural networks and data from neighboring meteorological stations," *Solar Energy*, vol. 134, pp. 119–131, Sep. 2016.
[35] Y. He, Q. Xu, J. Wan, and S. Yang, "Short-term power load probability density forecasting based on quantile regression neural network and triangle kernel function," *Energy*, vol. 114, pp. 498–512, Nov. 2016.
[36] J. Li, J. K. Ward, J. Tong, L. Collins, and G. Platt, "Machine learning for solar irradiance forecasting of photovoltaic system," *Renewable Energy*, vol. 90, pp. 542–553, May 2016.
[37] A. Tascikaraoglu, B. Sanandaji, G. Chicco, V. Cocina, F. Spertino, O. Erdinc, N. Paterakis, and J. P. Catalao, "Compressive Spatio-Temporal Forecasting of Meteorological Quantities and Photovoltaic Power," *IEEE Transactions on Sustainable Energy*, vol. PP, no. 99, pp. 1–1, 2016.
[38] C. Yang, A. A. Thatte, and L. Xie, "Multitime-Scale Data-Driven Spatio-Temporal Forecast of Photovoltaic Generation," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 1, pp. 104–112, Jan. 2015.
[39] C. Yang and L. Xie, "A novel ARX-based multi-scale spatio-temporal solar power forecast model," in *2012 North American Power Symposium (NAPS)*, Sep. 2012, pp. 1–6.
[40] V. Berdugo, C. Chaussin, L. Dubus, G. Hebrail, and V. Leboucher, "Analog method for collaborative very-short-term forecasting of power generation from photovoltaic systems," *Next Generation Data Mining Summit (NGDM'11)*, 2011.
[41] R. J. Bessa, A. Trindade, C. S. P. Silva, and V. Miranda, "Probabilistic solar power forecasting in smart grids using distributed information," *International Journal of Electrical Power & Energy Systems*, vol. 72, pp. 16–23, Nov. 2015.
[42] F. Golestaneh, H. B. Gooi, and P. Pinson, "Generation and evaluation of space–time trajectories of photovoltaic power," *Applied Energy*, vol. 176, pp. 80–91, Aug. 2016.
[43] B. Zhang, P. Dehghanian, and M. Kezunovic, "Spatial-temporal solar power forecast through use of Gaussian Conditional Random Fields," in *2016 IEEE Power and Energy Society General Meeting (PESGM)*, Jul. 2016, pp. 1–5.
[44] X. G. Agoua, R. Girard, and G. Kariniotakis, "Short-term spatio-temporal forecasting of photovoltaic power production," *IEEE Transactions on Sustainable Energy*, vol. PP, no. 99, pp. 1–1, 2017.
[45] J. Tastu, P. Pinson, P. J. Trombe, and H. Madsen, "Probabilistic Forecasts of Wind Power Generation Accounting for Geographically Dispersed Information," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 480–489, Jan. 2014.
[46] A. Tascikaraoglu, B. M. Sanandaji, G. Chicco, V. Cocina, F. Spertino, O. Erdinc, N. G. Paterakis, and J. P. S. Catalão, "A short-term spatio-temporal approach for Photovoltaic power forecasting," in *2016 Power Systems Computation Conference (PSCC)*, Jun. 2016, pp. 1–7.
[47] F. Golestaneh, P. Pinson, and H. B. Gooi, "Very Short-Term Non-parametric Probabilistic Forecasting of Renewable Energy Generation #x2014; With Application to Solar Energy," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3850–3863, Sep. 2016.
[48] Y. Zhang and J. Wang, "K-nearest neighbors and a kernel density estimator for GEFCom2014 probabilistic wind power forecasting," *International Journal of Forecasting*, vol. 32, no. 3, pp. 1074–1080, Jul. 2016.
[49] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," *Solar Energy*, vol. 83, no. 10, pp. 1772–1783, Oct. 2009.
[50] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. CRC Press, Apr. 1986.
[51] G. R. Terrell and D. W. Scott, "Variable Kernel Density Estimation," *The Annals of Statistics*, vol. 20, no. 3, pp. 1236–1265, 1992.
[52] M. P. Wand and M. C. Jones, *Kernel Smoothing*. CRC Press, Dec. 1994.
[53] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, Mar. 2015.
[54] A. R. Mugdadi and I. A. Ahmad, "A bandwidth selection for kernel density estimation of functions of random variables," *Computational Statistics & Data Analysis*, vol. 47, no. 1, pp. 49–62, Aug. 2004.
[55] T. A. O'Brien, K. Kashinath, N. R. Cavanaugh, W. D. Collins, and J. P. O'Brien, "A fast and objective multidimensional kernel density estimation method: fastKDE," *Computational Statistics & Data Analysis*, vol. 101, pp. 148–160, Sep. 2016.
[56] T. Duong and M. L. Hazelton, "Cross-validation Bandwidth Matrices for Multivariate Kernel Density Estimation," *Scandinavian Journal of Statistics*, vol. 32, no. 3, pp. 485–506, Sep. 2005.
[57] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley, Aug. 1991.
[58] R. Koenker, *Quantile Regression*. Cambridge University Press, 2005.
[59] G. B. Dantzig, *Linear Programming and Extensions*, 2007.
[60] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.
[61] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2014.
[62] P. Pinson, H. A. Nielsen, J. K. Møller, H. Madsen, and G. N. Kariniotakis, "Non-parametric probabilistic forecasts of wind power: required properties and evaluation," *Wind Energy*, vol. 10, no. 6, pp. 497–516, Nov. 2007.
[63] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, Mar. 2007.
[64] H. Hersbach, "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems," *Weather and Forecasting*, vol. 15, no. 5, pp. 559–570, Oct. 2000.

Xwégnon Ghislain Agoua received a Master's degree in Statistics (2014) from ENSAI (Ecole nationale de la statistique et de l'analyse de l'information), France and a PhD degree (2017) in energetics from MINES ParisTech, PSL - Research University. His research interests include statistical modeling, forecasting techniques, time series analysis, spatio-temporal regression models, and their applications to solar and wind generation modeling.

Robin Girard received a Master's degree (2004) in Computer Science and Applied Mathematics from INPG in Grenoble, France and a PhD (2008) in applied Mathematics from Joseph Fourier University in Grenoble.
He is currently a Research Engineer at the Centre for Energy and Processes of the Ecole des Mines de Paris. His research interests include wind and solar power forecasting, optimization in planning of energy production and spatio-temporal patterns of renewable power production.

George Kariniotakis (S'95-M'02-SM'11) was born in Athens, Greece. He received his Eng. and M.Sc. degrees from Greece in 1990 and 1992 respectively, and his PhD degree from Ecole des Mines de Paris in 1996. He is currently with the Centre PERSEE of MINES ParisTech as a senior scientist and head of the Renewable Energies and Smartgrids Group. He has authored more than 220 scientific publications in journals and conferences. He has been involved as participant or coordinator in more than 40 R&D projects in the fields of renewable energies and distributed generation. Among them, he was the coordinator of some major EU projects in the field of wind power forecasting such as Anemos, Anemos.plus and SafeWind projects. His scientific interests include time series forecasting, decision-making under uncertainty, modelling, management and power systems planning.

Annexe B

Simplexe et point intérieur

L'algorithme du simplexe [156] a été découvert en 1947 par Georges B. Dantzig [157]. C'est un algorithme très efficace de résolution des programmes linéaires. Considérons un programme linéaire sous forme canonique :

$$\min \quad c^T x, x \in \mathcal{R}^n$$

avec $Ax \geq b$ et $x \geq 0$. Chacune des m contraintes définies par le couple (A, b) ampute le domaine de variation de x d'un demi espace bordé par un hyperplan. Le domaine de x résultant de l'ensemble des contraintes est un polyèdre convexe (de plus, si ce domaine est borné et non vide, c'est un polytope).
L'objectif à minimiser étant linéaire, les ensembles de points prenant la même valeur sont des hyperplans parallèles, et l'on conclut qu'on peut toujours trouver une solution optimale sur un des sommets du polyèdre.

L'idée principale de la méthode du simplexe est de caractériser les sommets du polyèdre de façon algébrique, en faisant correspondre à chacun des sommets une base admissible. On part d'un sommet initial (donc d'une base admissible) puis on passe successivement d'un sommet à un sommet adjacent, en améliorant constamment la valeur de l'objectif.
Lorsqu'il n'est plus possible de passer d'un sommet à un sommet adjacent sans diminuer la valeur de l'objectif, on stoppe l'algorithme. Dantzig a démontré que cela se produit après un nombre fini d'itérations, et qu'on a forcément atteint un optimum.
La méthode du simplexe revient donc à une exploration combinatoire du polyèdre admissible des solutions.

Contrairement à l'algorithme du simplexe, la méthode de point intérieur part d'un point situé *à l'intérieur* du polyèdre admissible.
Le principe de cette méthode consiste à faire évoluer de façon itérative ce point vers une solution optimale du problème, tout en restant à l'intérieur strict du polyèdre admissible.
La méthode du point intérieur est fondamentalement différente de la méthode du simplexe. En effet, la méthode du simplexe finit par donner, en un nombre fini d'étapes, la valeur exacte de l'optimum alors que les méthodes de point intérieur se contentent de faire converger x. Il n'est pas possible d'atteindre l'optimum puisqu'on veille à ce que les itérés restent à l'intérieur strict du polyèdre. De plus si la solution n'est pas unique (une face entière du polyèdre est solution), les méthodes de point intérieur fourniront une solution située à l'intérieur de cette face tandis que le simplexe donnera un de ses sommets.
En terme de complexité algorithmique, les deux méthodes sont de type polynomial. A

précision ϵ fixée, le nombre d'opérations à effectuer pour obtenir une solution approché à ϵ près est borné par un polynôme de la taille du problème.

Une fois le problème de minimisation de la régression quantile résolu, on peut définir un indicateur de la qualité de l'ajustement proposé par R. Koenker & J. Machado [158]. Il est défini par

$$R^1(\alpha) = 1 - \frac{\min\limits_{\beta \in \mathbb{R}^p} \rho_\alpha(Y_i - X_i'\beta)}{\min\limits_{\beta_0 \in \mathbb{R}^p} \rho_\alpha(Y_i - \beta_0)}. \tag{B.1}$$

Ce critère est compris entre 0 et 1. Ce coefficient d'ajustement est similaire a celui utilisé dans le cas de la régression linéaire classique mais l'évaluation ne porte pas sur les mêmes aspects. Le coefficient R^2 de la régression linéaire mesure les performances d'un modèle en terme d'estimation de l'espérance conditionnelle avec une faible variance résiduelle. Le coefficient $R^1(\alpha)$ quant à lui mesure la performance de la régression quantile pour un quantile spécifique avec pour critère la minimisation d'une somme pondérée de résidus absolus. Koenker & Machado montrent dans leurs travaux que ce coefficient augmente avec le nombre de covariables du modèle. Les outils pour la construction des modèles, et le choix du meilleur modèle ayant ainsi été définis, nous nous sommes intéressés aux tests et diagnostics qui peuvent être mis en place dans le cadre d'une régression quantile. C'est dans ce cadre que nous présentons par la suite les propriétés asymptotiques de l'estimateur de la régression quantile.

Théorème Supposons que $\varepsilon_\alpha = Y - X'\beta_\alpha$ admette, conditionnellement à X, une densité en 0, $f_{\varepsilon_\alpha}(0|X)$ et que $J_\alpha = \mathbb{E}[f_{\varepsilon_\alpha}(0|X)XX']$ soit inversible. Alors

$$\sqrt{n}\left(\hat{\beta}_\alpha - \beta_\alpha\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \alpha(1-\alpha)J_\alpha^{-1}\mathbb{E}[XX']J_\alpha^{-1}\right) \tag{B.2}$$

Pour compléter l'étude des performances de nos modèles, nous vérifions par la suite que cette propriété asymptotique est vérifiée pour nos estimateurs et pour différents niveau de quantiles. Avec cette propriété asymptotique, on peut aussi construire des tests ou des intervalles de confiance autour de β_α. Pour cela, il existe plusieurs méthodes d'inférence. Certaines d'entre elles s'appuient sur une estimation directe de la variance asymptotique en partant du théorème B. Cette démarche est complexifiée par la présence de la densité conditionnelle $f_{\epsilon_\alpha}(0|X)$. D'autres méthodes se basent soit sur les tests de rang [55] soit sur du bootstrap mais sont coûteuses en temps de calcul.

Enfin, la régression quantile permet de ne pas supposer *a priori* que les variables explicatives ont un effet homogène sur l'ensemble de la distribution de la variable d'intérêt Y. Cette hypothèse peut être testée à partir des estimations obtenues. Par exemple, on peut tester l'homogénéité de l'effet d'une variable X_k correspondant à l'égalité des coefficients $\beta_{k,\alpha_1}, \ldots, \beta_{k,\alpha_m}$ où $(\alpha_1, \ldots, \alpha_m)$ peuvent être par exemple l'ensemble des déciles. Pour cela on s'appuie sur la distribution asymptotique jointe $(\hat{\beta}_{\alpha_1}, \ldots, \hat{\beta}_{\alpha_m})$ donnée par une généralisation du théorème B à plusieurs quantiles.

Résumé

L'évolution du contexte énergétique mondial et la lutte contre le changement climatique ont conduit à l'accroissement des capacités de production d'énergie renouvelable. Les énergies renouvelables sont caractérisées par une forte variabilité due à leur dépendance aux conditions météorologiques. La maîtrise de cette variabilité constitue un enjeu important pour les opérateurs du système électrique, mais aussi pour l'atteinte des objectifs européens de réduction des émissions de gaz à effet de serre, d'amélioration de l'efficacité énergétique et de l'augmentation de la part des énergies renouvelables. Dans le cas du photovoltaïque (PV), la maitrise de la variabilité de la production passe par la mise en place d'outils qui permettent de prévoir la production future des centrales. Ces prévisions contribuent entre autres à l'augmentation du niveau de pénétration du PV, à l'intégration optimale dans le réseau électrique, à l'amélioration de la gestion des centrales PV et à la participation aux marchés de l'électricité.
L'objectif de cette thèse est de contribuer à l'amélioration de la prédictibilité à court-terme (moins de 6 heures) de la production PV. Dans un premier temps, nous analysons la variabilité spatio-temporelle de la production PV et proposons une méthode de réduction de la non-stationnarité des séries de production. Nous proposons ensuite un modèle spatio-temporel de prévision déterministe qui exploite les corrélations spatio-temporelles entre les centrales réparties sur une région. Les centrales sont utilisées comme un réseau de capteurs qui permettent d'anticiper les sources de variabilité. Nous proposons aussi une méthode automatique de sélection des variables qui permet de résoudre les problèmes de dimension et de parcimonie du modèle spatio-temporel. Un modèle spatio-temporel probabiliste a aussi été développé aux fins de produire des prévisions performantes non seulement du niveau moyen de la production future mais de toute sa distribution. Enfin nous proposons, un modèle qui exploite les observations d'images satellites pour améliorer la prévision court-terme de la production et une comparaison de l'apport de différentes sources de données sur les performances de prévision.

Mots Clés

Production photovoltaïque, séries temporelles, spatio-temporel, prévision court-terme, stationnarité, densité, quantile, énergies renouvelables, images satellites, NWP, sélection de variables

Abstract

The evolution of the global energy context and the challenges of climate change have led to an increase in the production capacity of renewable energy. Renewable energies are characterized by high variability due to their dependence on meteorological conditions. Controlling this variability is an important challenge for the operators of the electricity systems, but also for achieving the European objectives of reducing greenhouse gas emissions, improving energy efficiency and increasing the share of renewable energies in EU energy consumption. In the case of photovoltaics (PV), the control of the variability of the production requires to predict with minimum errors the future production of the power stations. These forecasts contribute to increasing the level of PV penetration and optimal integration in the power grid, improving PV plant management and participating in electricity markets.
The objective of this thesis is to contribute to the improvement of the short-term predictability (less than 6 hours) of PV production. First, we analyze the spatio-temporal variability of PV production and propose a method to reduce the non-stationarity of the production series. We then propose a deterministic prediction model that exploits the spatio-temporal correlations between the power plants of a spatial grid. The power stations are used as a network of sensors to anticipate sources of variability. We also propose an automatic method for selecting variables to solve the dimensionality and sparsity problems of the space-time model. A probabilistic spatio-temporal model has also been developed to produce efficient forecasts not only of the average level of future production but of its entire distribution. Finally, we propose a model that exploits observations of satellite images to improve short-term forecasting of PV production.

Keywords

Photovoltaic generation, time series analysis, short-term forecasting, stationarity, probabilistic, renewable energies, satellites images, numerical weather prediction, variable selection