Comparaison de méthodes d'apprentissage automatique de prévision de la ressource solaire pour une application à une gestion optimisée des réseaux intelligents

Alexis Fouilloy

UNIVERSITE DE CORSE-PASCAL PAOLI
ECOLE DOCTORALE ENVIRONNEMENT ET SOCIETE
UMR CNRS 6134
Sciences Pour l'Environnement

Thèse présentée pour l'obtention du grade de
DOCTEUR EN MECANIQUE DES FLUIDES, ENERGETIQUE, THERMIQUE, COMBUSTION, ACOUSTIQUE

Mention : Energétique, Génie des procédés

Soutenue publiquement par
Alexis FOUILLOY

Le 24 Septembre 2019

Comparaison de méthodes d'apprentissage automatique de prévision de la ressource solaire pour une application à une gestion optimisée des réseaux intelligents

---

Directeurs :

M. Gilles NOTTON : MCF-HDR, Université de Corse
M. Cyril VOYANT : Dr-HDR, Université de Corse

Jury :

M. Cédric JOIN : PR, Université de Lorraine	Examinateur
M. Philippe LAURET : PR, Université de La Réunion	Rapporteur
M. Fawaz MASSOUH : MCF-HDR, Arts et Métiers ParisTech	Rapporteur
M. Jean-François MUZY : DR-CNRS, Université de Corse	Examinateur
Mme. Marie-Laure NIVET : MCF, Université de Corse	Invité
M. Gilles NOTTON : MCF-HDR, Université de Corse	Directeur de thèse
M. Giuseppe Marco TINA : PR, Université de Catane	Examinateur
M. Cyril VOYANT : Dr-HDR, Université de Corse	Co-Directeur de thèse

Résumé

Les enjeux relatifs à la production énergétique future, notamment en termes d'utilisation de ressources locales et « propres », conduisent les producteurs d'électricité à se tourner de plus en plus vers les sources renouvelables d'énergie et particulièrement les sources intermittentes que sont le vent et le soleil. Le problème est que leur caractère intermittent et aléatoire oblige les gestionnaires de réseau à limiter leur intégration au mix énergétique. Il est alors nécessaire de coupler différents systèmes de production pour garantir la stabilité du réseau et la sécurité des moyens de production. Afin de faciliter ces opérations de gestion et d'optimiser l'intégration des énergies renouvelables intermittentes, le solaire dans notre cas, il est nécessaire de s'intéresser à la prévision de la ressource. Dans le but de connaître à l'avance l'énergie disponible et de permettre une gestion optimale du couplage entre systèmes de production conventionnels et intermittents. Au cours de cette étude, nous avons développé et étudié un large panel de modèles de prévision du rayonnement solaire (persistance, persistance intelligente, filtre de Kalman, ARMA, réseau de neurones, processus Gaussien, machine à vecteurs de support, arbres de régressions simples, élagués, renforcés, ensachés et forêts aléatoires), pour des horizons de prévision utiles aux gestionnaires de réseaux, et appliqués à des données en provenance de différents sites. Ces travaux ont été réalisés dans le cadre d'un projet de recherche Horizon 2020, le projet TILOS pour « Technology Innovation for the Local Scale Optimum Integration of Battery Energy Storage » qui consiste en l'installation d'une centrale hybride solaire, éolienne et stockage par batteries $NaNiCl_2$ sur une petite île de l'archipel du Dodécanèse. Les horizons de prévision testés sont de 1 à 6 heures par pas de temps horaire (6h/1h) pour les 4 sites de mesures pour la prévision du rayonnement global horizontal. Les sites sont répartis en Europe dans des zones géographiques qui possèdent des climats différents : Ajaccio (Corse, France), Tilos (Dodécanèse, Grèce), Nancy (Grand Est, France) et Odeillo (Languedoc Roussillon, France). Nous avons caractérisé chaque site par variabilité des données, on entend par là leur tendance à varier fortement ou non avec le temps. Les principaux résultats de ces travaux sont que les prévisions sur des données sur des sites à faible variabilité peuvent être réalisés par des modèles simples. Plus la variabilité est élevée, plus la prévision et difficile à réaliser et des modèles plus complexes doivent être utilisés (basés sur l'apprentissage automatique et l'apprentissage d'ensemble) pour obtenir de meilleurs résultats. Nous avons par ailleurs utilisé une prévision probabiliste pour donner une plage de confiance de la prévision au gestionnaire de réseau. Etant donné que nous disposions de mesures des composantes directe normale et diffuse horizontale pour un des quatre sites, nous avons confronté nos modèles à ces prévisions. Il apparait que le rayonnement direct normal est difficile à prévoir, notamment à cause de sa forte variabilité, et que les forêts aléatoires sont les plus probants. Enfin des modèles ont été développés pour être insérés dans le système automatique de gestion de l'énergie, appliqués au rayonnement global incliné, avec un horizon de 10 minutes par pas de temps de 1 minute et un horizon de 2 heures avec des pas de temps de 10 et 15 minutes. Il apparait que les modèles d'apprentissage automatique donnent tous sensiblement de bons résultats et que le choix de l'un ou l'autre sera plutôt réalisé en fonction des contraintes techniques et pratiques des outils.

Mots clés : prévision, rayonnement global, intelligence artificielle, séries temporelles, apprentissage supervisé, stationnarité.

Abstract

The stakes relating to future energy production, particularly in terms of the use of local and "clean" resources, are leading electricity producers to turn more and more towards renewable sources of energy and particularly intermittent sources. the wind and the sun. The problem is that their intermittent and random nature forces network operators to limit their integration into the energy mix. It is then necessary to couple different production systems to guarantee the stability of the network and the security of the means of production. In order to facilitate these management operations and to optimize the integration of intermittent renewable energies, solar energy in our case, it is necessary to focus on the forecast of the resource. In order to know in advance, the available energy and to allow an optimal management of the coupling between conventional and intermittent production systems. In this study, we have developed and studied a wide range of solar radiation prediction models (persistence, smart persistence, Kalman filter, ARMA, neural network, Gaussian process, support vector machine, simple regression trees, pruned, boosted, bagged and random forests), for forecast horizons useful to network managers, and applied to data from different sites. This work was carried out as part of a Horizon 2020 research project, the TILOS project for "Technology Innovation for the Local Scale Optimum Integration of Battery Energy Storage" which consists of the installation of a solar, wind and solar hybrid power station. NaNiCl2 battery storage on a small island in the Dodecanese archipelago. The forecast horizons tested are from 1 to 6 hours per hour time step (6h / 1h) for the 4 measurement sites for horizontal global radiation prediction. The sites are spread across Europe in geographical areas with different climates: Ajaccio (Corsica, France), Tilos (Dodecanese, Greece), Nancy (Grand Est, France) and Odeillo (Languedoc Roussillon, France). We have characterized each site by data variability, which means their tendency to vary strongly or not with time. The main results of this work are that the forecasts on data on sites with low variability can be realized by simple models. The higher the variability, the more predictive and difficult to achieve, and more complex models must be used (based on machine learning and overall learning) for better results. We also used a probabilistic forecast to give a confidence range of the forecast to the network manager. Since we have measurements of the normal and diffuse horizontal direct components for one of the four sites, we compared our models to these predictions. It appears that normal direct radiation is difficult to predict, in particular because of its high variability, and random forests are the most convincing. Finally, models have been developed to be inserted in the automatic energy management system, applied to inclined global radiation, with a horizon of 10 minutes per time step of 1 minute and a horizon of 2 hours with no time steps 10 and 15 minutes. It appears that the machine learning models all give significantly good results and that the choice of one or the other will rather be made according to the technical and practical constraints of the tools.

Key words: forecasting, global radiation, artificial intelligence, time series, supervised learning, stationarity.

Table des matières

Introduction

L'intégration de plus en plus importante de moyens de production d'électricité à base de sources renouvelable d'énergie intermittente nécessite le développement de nouvelles structures de réseaux et de nouvelles méthodes de gestion des flux électriques. Pour ce faire, il convient de développer en parallèle :

- De nouveaux moyens de stockage d'énergie qui permettront de gérer les surplus de production et de décaler dans le temps leur apport augmentant ainsi leur disponibilité ;
- Des réseaux intelligents plus connus sous le nom de « smart grid » qui ont plusieurs avantages :
    - D'associer générateurs et charge dans des micro-réseaux et ainsi décentraliser la production en minimisant les pertes dans le réseau et en augmentant l'efficacité de la production ;
    - D'utiliser de nouveaux moyens de communication dans le but d'optimiser la production totale d'électricité et d'augmenter la qualité et la fiabilité des systèmes de puissance.
- Des méthodes de prévision de la production et de la consommation qui permettront d'anticiper les actions à réaliser par le gestionnaire du réseau à différentes échelles de temps par l'intermédiaire d'un Système de Gestion de l'Energie.

L'objet de cette thèse est de contribuer au développement de ce dernier point par une étude de différents modèles de prévision appliquée à quatre sites météorologiques.

Les stratégies de prévisions sont différentes selon les besoins, le choix de techniques se fait notamment en fonction de l'horizon temporel souhaité. La Figure 0-1 présente les différentes possibilités de prévision en considérant l'erreur de prévision en fonction de l'horizon de prévision.

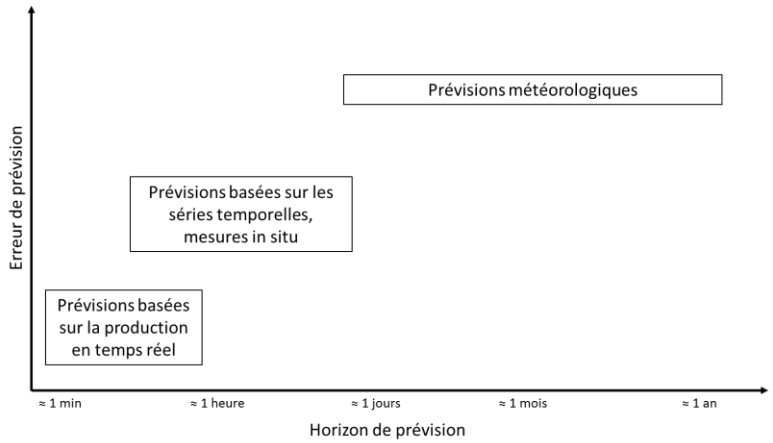


Figure 0-1 : Techniques de prévision, erreur de prévision en fonction de l'horizon temporel

Ce travail ayant été réalisé dans le cadre de notre participation au projet Horizon 2020 TILOS, les nécessités opérationnelles du gestionnaire du réseau et du système hybride nous ont tout d'abord conduit à développer des modèles de prévision pour des horizons temporels allant de 1 à 6 heures par des pas de temps horaire, puis des modifications du cahier des charges nous ont ensuite amené à mettre en œuvre des modèles pour des horizons allant de 1 à 10 minutes et de 15 minutes à 2 heures avec des pas de temps compris entre 1 minute et 15 minutes respectivement.

Sur la Figure 0-1, on peut ainsi voir que, pour ces horizons de prévision, les techniques les plus adaptées sont basées sur les séries temporelles de mesures in situ. Nous avons choisi de nous intéresser

à des modèles de prévision utilisant notamment des méthodes d'intelligence artificielles que nous comparons à des méthodes plus simples ne nécessitant pas d'apprentissage et d'historique de données.

En ce qui concerne l'organisation du manuscrit, nous présenterons dans le premier chapitre le contexte de ce travail et la problématique de l'intégration des énergies renouvelables intermittentes et aléatoires dans les réseaux électriques

Le second chapitre sera consacré à la présentation des données de rayonnement solaire et aux différentes étapes de leur préparation : collecte, contrôle qualité, filtration, formalisme des séries temporelles et la stationnarisation. Nous détaillerons, également, la méthode utilisée pour classer ces données en fonction de leur variabilité et l'utilisation de l'auto-information mutuelle pour déterminer la dimension de l'historique des données nécessaire à la réalisation de la prévision.

Le troisième chapitre concerne la recherche bibliographique sur la prévision univariée du rayonnement solaire. Il concentrera tous les modèles que nous avons mis en œuvre pour réaliser cette phase de prévision. Ces modèles se déclinent essentiellement en deux familles, les modèles sans apprentissage et les modèles avec apprentissage. Nous consacrerons une partie à l'encadrement de la prévision et nous aborderons le cas opérationnel du projet TILOS avec toutes les demandes et les difficultés liées au fonctionnement en temps réel.

Le quatrième chapitre sera consacré aux résultats des expérimentations réalisées. Il se décline en quatre grandes parties :

- La prévision du rayonnement global horizontal sur les quatre sites de mesure pour des horizons allant de 1 à 6 heures par pas de temps horaire.
- La méthodologie développée pour la prévision probabiliste ;
- Les expérimentations menées sur les autres composantes du rayonnement, à savoir le rayonnement diffus horizontal et le rayonnement direct normal ;
- Le développement des modèles pour la prévision du rayonnement global incliné liée au projet TILOS.

Le cinquième chapitre conclura ces travaux et présentera quelques perspectives ouvertes de recherches futures.

I. Contexte général et problématique

1. Introduction

Depuis l'aube de l'humanité, le développement énergétique rime avec évolution. En effet, les besoins grandissants en énergie ont connu une augmentation exponentielle. Cela implique que l'homme a dû, au fil de l'évolution, trouver de nouveaux moyens de production de l'énergie. C'est à la révolution industrielle que les avancées ont été les plus importantes ; la machine à vapeur et l'exploitation des énergies fossiles marque un tournant dans la façon de produire et d'utiliser l'énergie. Jusque dans le milieu du XXème siècle les énergies fossiles sont largement utilisées, les moteurs à combustion interne et les turbines à vapeur sont largement développés ainsi que les turbines hydrauliques.

Ce n'est qu'à partir des années 90 que le Monde commence à prendre conscience des répercussions et des conséquences liées à la façon de produire de l'énergie. Le premier acte fondateur de la prise en compte des effets de la production énergétique sur l'environnement est la Convention-Cadre des Nations Unies sur les Changements Climatiques (CCNUCC) qui a été adoptée au cours du Sommet de la Terre de Rio de Janeiro en 1992. La CCNUCC est la première tentative, dans le cadre de l'ONU, de mieux cerner ce qu'est le changement climatique et comment y remédier. Il s'agit du premier acte qui prend véritablement en compte les mécanismes qui lient la production d'énergie et la présence croissante des gaz à effet de serre dans l'atmosphère tout en différenciant la part des pays industrialisés et celle des pays en voie de développement.

Après cinq années d'attente le protocole de Kyoto est signé le 11 décembre 1997 lors de la 3ème Conférence des parties à la convention (COP3) à Kyoto, au Japon, il est entré en vigueur le 16 février 2005. Ce protocole visait à réduire, entre 2008 et 2012, d'au moins 5 % par rapport au niveau de 1990 les émissions de six gaz à effet de serre : dioxyde de carbone, méthane, protoxyde d'azote et trois substituts des chlorofluorocarbones.

Plus tard, en décembre 2009, 192 pays se sont réunis à la conférence de Copenhague et négocient un nouvel accord sur le climat qui va remplacer le protocole de Kyoto. D'après le secrétaire général de l'ONU, cette conférence donne naissance au « premier accord réellement mondial » dans le but de réduire de moitié les émissions de gaz à effet de serre d'ici 2050 en comparaison de celles de 1990 (Multon *et al.*, 2011).

Le 12 Décembre 2015, dans la lignée du travail effectué jusque-là, l'accord de Paris est signé. Les 195 pays réunis adoptent un accord historique pour lutter contre le changement climatique et entreprennent des mesures et investissent pour un avenir résilient, durable et bas carbone. L'Accord de Paris rassemble, pour la première fois, toutes les nations dans une cause commune en fonction de leurs responsabilités historiques, actuelles et futures. L'objectif principal de cet accord est de maintenir l'augmentation de la température mondiale bien en dessous de 2 degrés Celsius et de mener des efforts encore plus poussés pour limiter l'augmentation de la température à 1,5 degré Celsius au-dessus des niveaux préindustriels. La limite de 1,5 degré Celsius constitue une ligne de défense considérablement plus sûre contre les pires impacts du changement climatique.

Il aura fallu près de 30 ans pour qu'un réel positionnement politique soit pris pour la production d'énergie et l'impact environnemental. De plus, les difficultés posées par l'interaction entre développement économique et émission de gaz à effet de serre rendent l'équation encore plus difficile à résoudre.

1.1. Les énergies renouvelables au niveau mondial

Dans le monde entier, l'association de la volatilité des marchés des énergies fossiles et la nécessité de la protection de l'environnement et de la réduction des émissions de gaz à effet de serre imposent une révision des stratégies énergétiques. Les énergies renouvelables disposent d'atouts essentiels et de caractéristiques remarquables pour prendre une place prépondérante dans les mix énergétiques des états. Les énergies renouvelables sont réparties en plusieurs sources :

- L'énergie solaire : issue du rayonnement solaire et qui peut être utilisée pour de la production soit thermique, soit électrique ;
- L'énergie éolienne : issue de la force générée par les vents et qui génère de l'énergie mécanique, aujourd'hui principalement convertie en électricité ;
- L'énergie hydraulique : issue de la pression générée par un stockage d'eau, la différence d'altitude convertie en vitesse et en énergie mécanique, aujourd'hui aussi convertie principalement en électricité ;
- L'énergie géothermique : issue de la chaleur générée par la croûte terrestre, convertie en chaleur ou en électricité ;
- L'énergie issue de la biomasse : issue du bois, du méthane, principalement convertie par combustion en chaleur puis parfois en électricité ;
- Les énergies marines : issues des mouvements au sein des océans, différence de hauteur due à la houle ou puissance des courants marins, principalement convertie en électricité

Naturellement, tous les types d'énergies renouvelables n'ont pas nécessairement la même utilité. En quelques chiffres, les énergies renouvelables au niveau mondial représentent fin 2015, plus de 19,3 % de la capacité énergétique mondiale. Elles fournissent 24,5 % de l'électricité mondiale à la fin de l'année 2016.

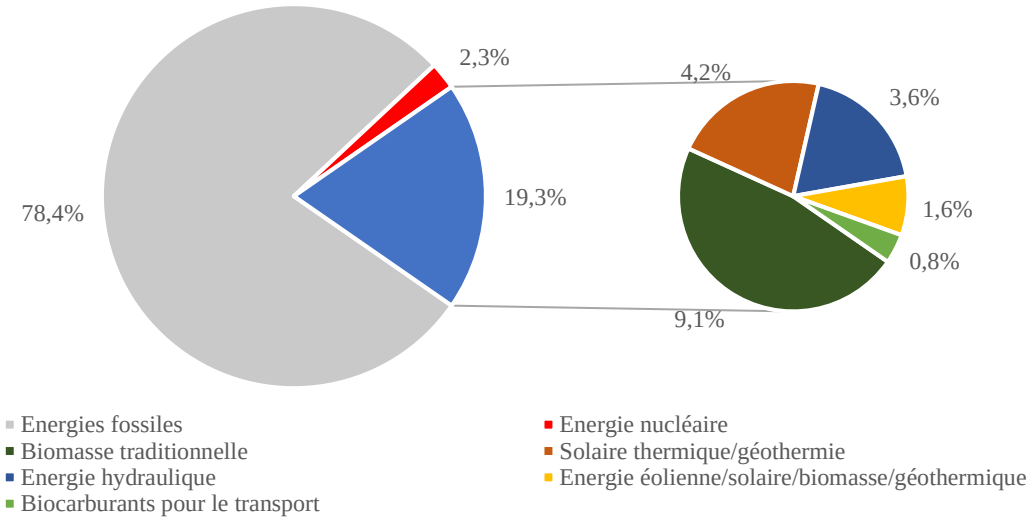


Figure I-1: Répartition des différentes sources d'énergie en 2016 (REN21 Renewable Energy Policy Network for the 21th Century, 2017)

La Figure I-1 représente les proportions des différentes sources d'énergie au niveau mondial en 2016, elle ne différencie pas la production d'électricité de la production de toutes les autres énergies. En effet, les sources d'énergies renouvelables sont utilisées dans la production de plusieurs types d'énergies (mécanique, chaleur…). Lorsque l'on s'intéresse de plus près à la production d'électricité, la répartition des différents moyens de production se présente comme indiquée sur la Figure I-2.

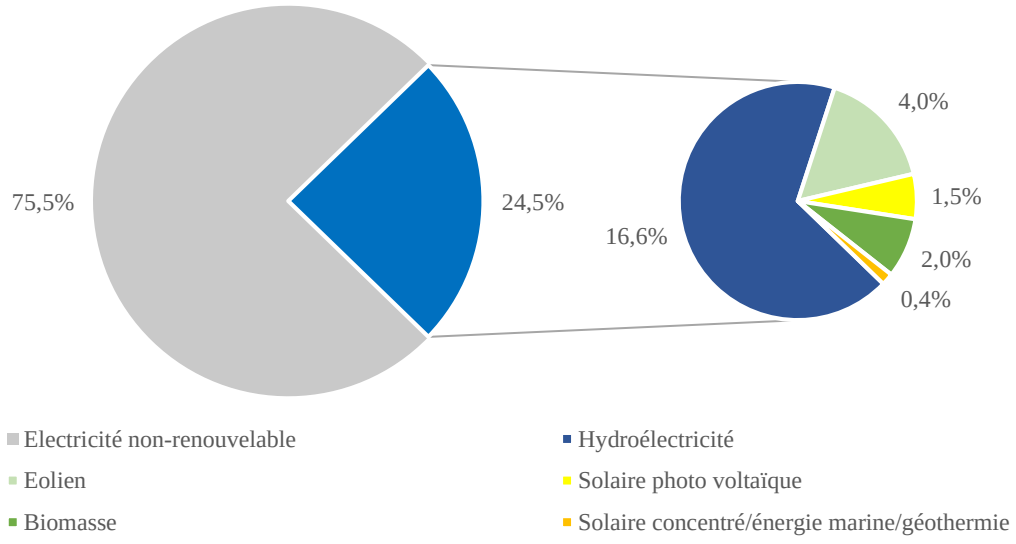


Figure I-2: Répartition des énergies renouvelables dans la production mondiale d'électricité (*REN21 Renewable Energy Policy Network for the 21th Century*, 2017)

Près d'un quart de la production mondiale d'électricité est assuré par des dispositifs utilisant les énergies renouvelables. Cette proportion aura tendance, dans les années à venir, à augmenter. D'après l'Agence Internationale de l'énergie (IEA : International Energy Agency), il y aura, notamment pour les pays émergents, une forte augmentation de la puissance installée pour toutes les sources d'énergies renouvelables confondues. Par exemple, à elle seule, la Chine devrait installer près de 1300 GW d'énergies renouvelables d'ici 2025. De plus, si ces formes d'énergies se développent fortement pour la production électrique, elles croissent aussi considérablement dans les secteurs de la production de chaleur et des transports. Cette croissance pose alors de nouveaux problèmes aux niveaux de leur intégration et de leur gestion et il convient d'entrevoir de nouvelles solutions notamment en termes de gestion de flux d'énergie.

Toutes ces sources d'énergie renouvelables n'ont évidemment pas les mêmes caractéristiques, notamment lorsque l'on parle de stabilité de production.

- L'énergie hydraulique est utilisable à volonté à n'importe quel moment.
- L'énergie géothermique est stable et fournit une source viable en quantité et en durée.
- Les énergies marines sont elles aussi stables, en effet les courants marins et les marées sont des phénomènes assez bien connus.
- En revanche, les énergies solaires et éoliennes ne disposent pas de cette caractéristique de stabilité, elles sont dites intermittentes, et qui plus est, stochastiques.

1.2. Les énergies renouvelables intermittentes

Lorsque l'on parle d'intermittence des sources d'énergie renouvelables, c'est dans le but de souligner leurs caractères variables et aléatoires, et ce, à différentes échelles de temps. Ces sources d'énergie ont des caractéristiques qui les différencient des moyens conventionnels de production d'énergie ; contrairement à ces derniers, il n'est pas possible d'adapter la production d'énergie en temps réel avec la consommation. Il n'est donc pas possible d'assurer la présence de la ressource à l'instant t, en absence de vent ou de soleil, la production est alors impossible à assurer ou à planifier.

Pour l'énergie éolienne, la variabilité provient des mouvements atmosphériques qui sont à l'origine du vent et dépend d'une multitude de facteurs : météorologiques (anticyclones, dépressions, gradients thermiques, etc…), ou encore des caractéristiques géographiques des lieux choisis pour implanter des moyens de production (orographie, désert, relief, proximité de la mer etc…) ce qui la rend particulièrement complexe à appréhender. Pour l'énergie solaire, domaine auquel nous allons nous intéresser dans ce manuscrit, les variations dans la quantité d'énergie reçue et la puissance produite dépendent de l'endroit où l'on se trouve sur Terre et se produisent à différentes échelles de temps :

- Saisonnière : à cause de la position de la Terre par rapport au Soleil ;
- Journalière : à cause du changement de l'angle entre le rayonnement et le sol ;
- Infra horaire : à cause des conditions météorologiques à l'échelle locale (état de l'atmosphère).

Ce comportement naturel de ces sources renouvelables ne facilite pas son intégration dans le mix électrique, et rend difficile le maintien de la stabilité du réseau et la garantie d'une fourniture énergétique de qualité. Pour ces différentes raisons, la part de la puissance électrique d'origine renouvelable intermittente est limitée par décret à 30% de la puissance appelée et parfois légèrement plus selon la situation locale.

En effet, en dessous de cette limite les problèmes liés à l'intermittence peuvent être absorbés par le réseau (par le biais des systèmes de production conventionnels, phénomènes de foisonnement, etc…), au-delà de cette limite le maintien de l'équilibre production/consommation serait très difficile à assurer et le risque d'effondrement du réseau serait alors accru.

Cet arrêté autorise alors, temporairement, la déconnexion des systèmes de production d'électricité d'origine renouvelable intermittente (à l'exception des systèmes de très faibles puissance crête) pour assurer le bon fonctionnement du réseau.

1.3. Le réseau électrique et sa gestion : le cas insulaire

La gestion du réseau électrique répond à une problématique complexe, à chaque instant il est nécessaire de respecter l'équilibre entre la production et la consommation. Or, l'électricité doit être utilisée dès lors qu'elle est produite, il faut donc intervenir soit au niveau des lieux de production soit au niveau du centre d'expédition (dispatching).

Si la production diminue soudainement en raison de la perte d'un moyen de production (ou d'un passage nuageux au-dessus d'une installation PV), cet équilibre est affaibli et la fréquence tombe en dessous de la fréquence de référence. Dans ce cas, une augmentation rapide de l'énergie électrique délivrée par un autre moyen de production connecté au réseau doit être réalisée (par exemple, une augmentation de la puissance produite d'un moteur qui fonctionne à charge partielle ou par le démarrage d'un nouveau moyen de production).

Dans le cas où une augmentation de la consommation, et, par conséquent, le démarrage d'un nouveau moyen de production, n'auraient pas été anticipés, le déséquilibre s'accentue et il est alors nécessaire de procéder au délestage d'une partie des usagers. Si cette opération n'est pas assez rapide, le risque est alors de subir un effondrement complet du réseau, en effet, la mise en production d'une centrale peut prendre beaucoup de temps (plus d'une trentaine de minutes pour un moteur diesel). De même, l'augmentation de la puissance produite, pour un système de production déjà en marche, n'est pas instantanée et dépend de sa vitesse de rampe (en anglais « ramp rate ») exprimée en % de puissance par unité de temps. Dans le cas, où la production devient supérieure à la consommation, la fréquence du réseau augmente et il faut alors stopper un moyen de production avant que celui-ci ne soit endommagé. Cette opération conduit aussi à une panne de courant sur le réseau (blackout).

Dans un réseau continental interconnecté, qui est de grande envergure et où les pics de consommation ne sont pas simultanés, la gestion de la balance est moins complexe. En effet, si l'on prend le cas de l'Europe pour lequel le réseau s'étend de Gibraltar à la Turquie et de l'Italie aux pays Scandinaves, les pics de production et de consommation n'ont pas lieu en même temps dans tous les pays. Plus le réseau est étendu, plus sa capacité à absorber les variations est importante. De plus, les gestionnaires de réseaux disposent de nombreux moyens de production répartis sur le territoire pour maintenir en permanence l'équilibre.

Ces problèmes de gestion sont d'autant plus ressentis sur les réseaux de petite taille, comme cela est le cas pour les réseaux insulaires, qui sont partiellement ou non interconnectés avec les grands réseaux continentaux.

Les réseaux insulaires ont une fragilité liée à leurs tailles : un court-circuit dans le système électrique va générer une chute de tension dans toute l'île (*Bilan électrique SEI 2016*, 2016; Humberto Marin, 2011), la faible inertie implique une variabilité rapide et importante avec des conséquences sur la tension ; ces problèmes sont aggravés par la proportion élevée de la capacité d'un générateur électrique en comparaison avec la puissance crête du réseau lui-même. Les probabilités de défaillance dans un réseau insulaire sont, de ce fait, bien plus élevées que dans un réseau interconnecté. Les chutes de tension et de fréquence sont plus nombreuses et plus profondes dans les îles que sur le continent (plusieurs dizaines chaque année dans les îles françaises) (*Bilan électrique SEI 2016*, 2016). Par exemple, en Corse avant l'interconnexion partielle avec la Sardaigne, il a été dénombré plus de 200 pannes par an sur le réseau de transport de l'électricité avec des creux de tension et de fréquence (moins de 46 Hz) (Barlier, 2000). La plupart du temps les îles sont faiblement peuplées, ce qui induit une consommation énergétique qui limite la puissance des systèmes de production. Cette contrainte impose l'utilisation de systèmes de production de petite taille, les plus adaptés étant les moteurs le plus souvent utilisant du fuel et, dans une moindre mesure, du gaz. L'avantage d'utiliser 3 moteurs de 20 MW, par exemple, plutôt qu'un seul de 60 MW est de disposer d'une plage de fonctionnement plus étendue, entre 65 et 100 % pour chaque moteur, soit une plage de fonctionnement de 13 à 60 MW dans le cas de 3 générateurs de 20 MW contre 42 à 60 MW dans le cas d'un seul moteur à 60 MW. Le second avantage, et pas le moindre, est, qu'en cas de perte d'un moyen de production de pouvoir limiter les dégâts qui en résulte.

La Figure I-3 présente la répartition de la production d'électricité pour les îles françaises (Corse, Réunion, Guadeloupe, Martinique) et la Guyane (gérée aussi comme réseau isolé).

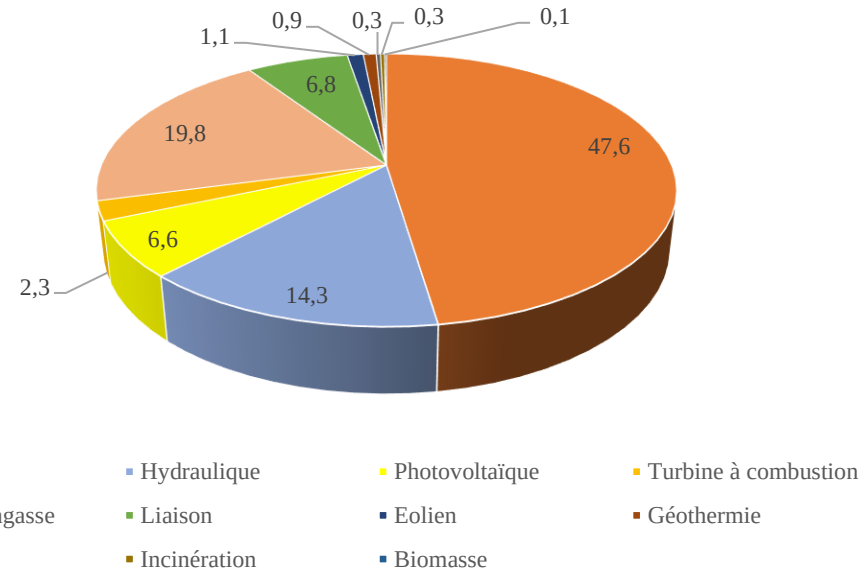


Figure I-3: Répartition de la production électrique par sources pour les réseaux insulaires et isolés (Bilan électrique SEI 2016, 2016)

La majeure partie de la production est réalisée par des sources thermiques (Diesel, charbon, bagasse). Ces sources, qui ont l'avantage d'être complétement contrôlables, ont l'inconvénient d'être fortement émettrices de CO_2 et de particules fines ce qui, ajouté au fait que l'on limite la pénétration des systèmes de production par énergies renouvelables, a pour effet d'une part :

- D'augmenter fortement les prix de revient de l'électricité ;
- D'autre part d'avoir une production énergétique avec un impact environnemental fort.

La section suivante traite de la nécessité de la prévision dans l'intégration des énergies renouvelables dans la production énergétique actuelle. Il sera alors question des volets liés au coût de la production mais aussi aux contraintes techniques imposées par l'utilisation de ces moyens de production.

2. Pourquoi prévoir ? Le coût de l'intermittence et les bénéfices de la prévision

Les enjeux énergétiques et écologiques des prochaines années conduisent naturellement vers une production d'énergie plus respectueuse de l'environnement tout en maintenant un niveau d'approvisionnement suffisant en quantité et en qualité.

Dans cette optique, il devient donc naturel de penser à allier les systèmes de production conventionnels (nucléaire, énergie fossile, hydroélectricité…) avec les moyens de production qui utilisent les sources renouvelables et intermittentes (solaire et éolien), tout en considérant l'aspect économique de la production d'énergie. Une des conséquences directes de cette nécessité est la croissance rapide des capacités de production d'énergie par des systèmes qui utilisent les énergies solaires et éoliennes. La croissance du marché mondial des systèmes photovoltaïques et éoliens sur ces dernières années se poursuit toujours avec 50 GW crête d'installations photovoltaïques et 62,7 GW d'éoliennes installées en 2015 (+ 25% pour le photovoltaïque et + 22% pour l'énergie éolienne par rapport à 2014). Les chiffres parlent d'eux même, à la fin de l'année 2017, les systèmes photovoltaïques représentent 106,6 GW en Europe et 405 GW dans le monde (« Baromètre Photovoltaïque 2018 | EurObserv'ER », 2018), et les systèmes de production éoliens représentent 168,9 GW en Europe et

539,3 GW dans le monde (« Baromètre éolien 2018 | EurObserv'ER », 2018). L'augmentation de la part de l'électricité produite par les systèmes photovoltaïques et éoliens impose de développer des méthodes et des moyens pour intégrer pleinement ces deux sources d'énergie renouvelables intermittentes et stochastiques (SERIS) dans les réseaux électriques. Ainsi, l'un des principaux défis pour l'approvisionnement énergétique mondial dans un proche avenir est la forte intégration des sources d'énergies renouvelables (Heinemann *et al.*, 2006a). Le caractère aléatoire et « fatal » (terme utilisé par les gestionnaires de réseaux) de ces sources d'énergies pose de nombreux problèmes aux gestionnaires de réseaux avec un impact important sur le coût de production.

Le coût supplémentaire induit par cette intermittence est lié à une sous-production ou une surproduction due au caractère aléatoire fluctuant des ressources solaires et éoliennes. La conséquence directe est une diminution de la stabilité de la production et de la sureté d'approvisionnement.

Afin de diminuer ou lisser ces variations, il est nécessaire d'utiliser des dispositifs de stockage de l'énergie ainsi que des systèmes de production d'énergie de « secours » qui seront capables de compenser immédiatement les variations de puissance. Ces moyens de production, dits de sauvegardes, doivent rester allumés de manière à pouvoir rapidement agir pour le maintien de l'équilibre production/consommation.

Dans le cas d'une surproduction du PV et de l'éolien, par rapport à un seuil de sécurité, ces systèmes sont déconnectés induisant encore un coût supplémentaire de production puisque l'investissement n'est pas utilisé pendant cette déconnexion.

Il est évident que de telles difficultés, induites par l'intermittence de la vitesse du vent et du rayonnement solaire, entraîneront un coût de production supplémentaire par rapport à une production conventionnelle. La mise en relief des coûts de production est une tâche délicate car elle dépend de plusieurs paramètres, tels que le pays considéré et la politique énergétique qui y est menée, la situation du réseau électrique (réseau connecté, partiellement connecté ou bien distant), les caractéristiques météorologiques des sites de mise en œuvre, etc.

Dans cette partie nous allons montrer les avantages liés à une prévision de qualité lorsque l'on souhaite diversifier les moyens de production d'électricité et augmenter l'intégration des énergies renouvelables intermittentes. Cette question sera traitée d'un point de vue technique mais aussi économique. Le fait d'aborder ces deux aspects qui, de prime abord, semblent éloignés, nous permettra d'étayer notre argumentation sur la nécessité de la prévision.

2.1. L'intégration des SERIS au réseau électrique

L'incertitude et la variabilité des ressources éoliennes et solaires imposent des actions supplémentaires et complexes pour maintenir constamment l'équilibre du système et sa stabilité électrique. Une plus grande flexibilité du système est nécessaire pour prendre en compte la variabilité de l'offre et la relation entre les niveaux de production et de charge.

L'opérateur électrique a déjà des difficultés à maintenir l'équilibre production/consommation avec des moyens de production d'énergie classiques et gérables et en particulier dans des réseaux électriques de petite taille et/ou non interconnectés (insulaires par exemple). La fiabilité du système électrique dépend alors de sa capacité à s'adapter aux changements attendus et imprévus (que ce soit du point de vue de la production ou de la consommation) et aux perturbations, tout en maintenant la qualité et la continuité de la distribution de l'électricité au client (Notton, 2015). Même si aucune SERIS n'est intégrée au réseau électrique, des réserves d'énergie sont nécessaires, elles peuvent être divisées en deux catégories: réserve en cas d'imprévu, utilisée en cas d'événement spécifique (comme la mise sous

tension d'une centrale) et réserve non évènementielle, utilisée en permanence (en raison, par exemple, d'une prévision de charge peu fiable) (Sjoerd Brouwer *et al.*, 2014).

Les différents types de réserves ont des temps de mises en œuvre différents qui ont chacune leur application. Les réserves, dont la mise en œuvre est la plus rapide, sont destinées à une réponse en fréquence, elles répondent à des évènements qui apparaissent en quelques secondes. Pour les évènements dont l'échelle temporelle est plus longue, de l'ordre de la minute à l'heure, l'opérateur a plus de temps pour agir et se sert de réserves dont la montée en puissance est plus longue dans un but de maintien de la charge. Enfin pour des évènements plus longs (de l'ordre de plusieurs heures) l'opérateur utilise des réserves dites complémentaires (remplacement d'un moyen de production ou augmentation importante de la demande) (Milligan *et al.*, 2010). L'introduction des SERIS dans un réseau électrique n'affecte que la réserve non événementielle, notamment en raison de la prévision imparfaite de leur production (Sjoerd Brouwer *et al.*, 2014). Pour le gestionnaire de réseaux, un événement prévu et anticipé est plus facile à gérer. Le gestionnaire du réseau électrique a besoin de connaître la production et la consommation d'électricité à venir, et ce pour différents horizons temporels (Figure I-4) (Diagne *et al.*, 2013; Lara-Fanego *et al.*, 2012).

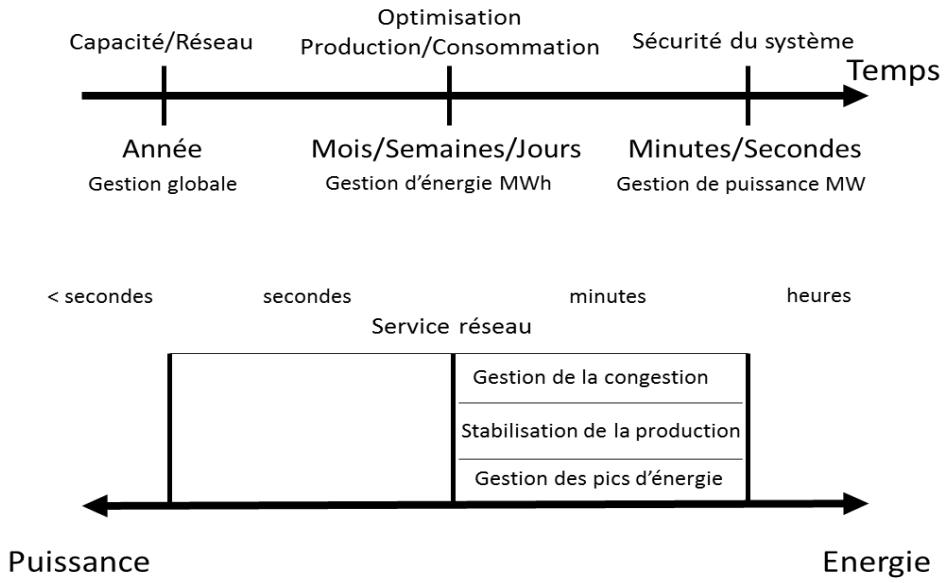


Figure I-4: Echelle de prévision pour la gestion énergétique du réseau électrique

L'intermittence et le caractère incontrôlable de la production des SERIS posent également des problèmes de fluctuations de tension, ces propriétés ont un impact non négligeable sur la qualité de l'alimentation locale et accentuent les problèmes de stabilité du réseau (Anderson et Leach, 2004; Moreno-Munoz *et al.*, 2008).

Par exemple, lorsque la production par les SERIS est disponible et que le niveau de charge est faible, les générateurs à moteurs thermiques doivent revenir à un niveau de production minimal, s'accompagnant d'un rendement faible et de coûts de production élevés. Le haut niveau de pénétration des SERIS dans le mix énergétique génère alors un surcoût important de la production électrique.

L'intégration des SERIS dans les réseaux électriques oblige toutes les centrales thermiques à énergie fossile à s'allumer et à s'éteindre plus souvent et à modifier plus fréquemment leurs niveaux de production pour s'adapter en permanence à la charge, ce qui a deux conséquences majeures :

- Une augmentation de l'usure des unités de production ;

- Une diminution de rendement d'environ 4% (de 0 à 9% (Sjoerd Brouwer *et al.*, 2014)), avec des contraintes thermiques supplémentaires sur les équipements.

Toutes ces généralités sur les SERIS révèlent que leurs propriétés intrinsèques complexifient grandement leur utilisation au sein des réseaux de production électrique conventionnels, cependant il est difficilement envisageable de se passer de ces moyens de production à l'heure où l'on se soucie de plus en plus de l'impact environnemental de la production d'électricité et de la raréfaction des ressources fossiles.

2.2. Prévoir la production des SERIS : une nécessité pour une meilleure intégration

Comme nous l'avons souligné précédemment, les propriétés des SERIS ne permettent pas leur intégration de manière forte et optimale au sein de la production énergétique. La prévision de la production de ces différents systèmes devient alors indispensable (Paulescu *et al.*, 2013). Cette étape prévisionnelle est nécessaire pour ensuite estimer les réserves, planifier la production électrique, gérer de manière optimale le stockage et négocier les prix sur le marché international de l'électricité (Diagne *et al.*, 2012; Hammer *et al.*, 1999; Lauret *et al.*, 2015; McCandless, 2015).

Une petite erreur dans la prévision induit deux effets négatifs: l'opérateur du réseau peut recevoir de lourdes pénalités car une prévision imprécise ne permet pas d'atteindre le profil de production prévu et le recours à l'utilisation de générateurs auxiliaires est plus important pour compenser l'écart entre production prévue et production réelle (Koeppel et Korpås, 2008; Masa-Bote *et al.*, 2014). Une solution consiste également à utiliser un stockage local en combinaison avec les SERIS pour compenser ces écarts et/ou de combiner plusieurs SERIS réparties sur une vaste zone (foisonnement) afin de lisser la production et garantir une puissance moyenne « plus garantie » ; les erreurs de prévisions individuelles de chaque SERIS peuvent alors se compenser et l'erreur de prévision globale se réduire (effet global) (da Silva Fonseca Junior *et al.*, 2014).

Différents systèmes de stockage ont été développés et constituent une solution viable pour absorber les excès de puissance et d'énergie produits par les SERIS (et les libérer en période de pointe), pour compenser et lisser des fluctuations très courtes et pour maintenir une qualité de la production d'énergie. Ces moyens de stockage sont généralement classés en 3 catégories, rassemblées dans le Tableau I-1 (Butler *et al.*, 2002) :

- Stockage d'énergie de masse ou stockage de gestion de l'énergie utilisé pour découpler le calendrier de production et de consommation ;
- Génération distribuée ou puissance de comblement pour le lissage des pics ; le stockage est utilisé pendant quelques secondes à quelques minutes et assure la continuité du service lors du passage d'une source d'énergie à une autre ;
- Qualité de l'alimentation ou fiabilité de distribution. L'énergie stockée (plutôt la puissance stockée) est utilisée quelques secondes ou moins pour assurer la continuité de la qualité de l'alimentation et permettre un meilleur lissage de la production.

Tableau I-1: Les différentes catégories de stockage et leurs applications (Butler *et al.*, 2002)

Catégorie	*Puissance de décharge*	*Temps de décharge*	*Energie stockée*	*Application représentative*
Stockage de masse	10-1000 MW	1-8 h	10-8000 MWh	Mise à niveau de la charge
Génération distribuée	0,1-2 MW	0,5-4 h	50-8000 MWh	Ecrêtage, report
Qualité d'alimentation	0,1-2MW	1-30 s	0,03-16,7 MWh	Qualité/fiabilité de distribution

Le Tableau I-1 montre que les moyens de stockage d'énergie agissent à différentes échelles en temps et que leur gestion nécessite de connaître la puissance ou l'énergie produite par les SERIS à différents horizons temporels et granularités temporelles : de très court ou court pour la catégorie de qualité de l'alimentation à horaire ou quotidien pour le stockage de masse.

Une bonne prévision est nécessaire pour une gestion optimale du stockage : elle permet de diminuer les réserves de flexibilité (Bhatt *et al.*, 2016; Black et Strbac, 2006; Koeppel et Korpås, 2008; Masa-Bote *et al.*, 2014) et d'optimiser la gestion du stockage d'énergie en anticipant les phases de charge et de décharge.

De même, l'opérateur électrique a besoin de connaître la production future à différents horizons, de un à trois jours pour la préparation des moyens de production (ainsi que pour planifier des phases de maintenance préventive ou corrective), de quelques minutes à quelques heures pour planifier la mise en route de centrales de réserve (entre 5 min et 40 h selon les moyens de production d'énergie, de quelques secondes à plusieurs minutes pour prévoir la montée en puissance des moyens de production (Saguan, 2007).

Le Tableau I-2 rassemble les différentes caractéristiques de mise en œuvre des systèmes conventionnels de production d'énergie.

Tableau I-2: Caractéristiques des moyens de production d'électricité

Moyen de production	**Taille (MW)**	**Capacité minimale**	**Montée en puissance (%/min)**	**Temps de démarrage**
Centrale nucléaire	400-1300 /réacteur	20%	1%	18-40 h
Centrale thermique à vapeur	200-800 /turbine	50%	0,5-5%	5-20 h
Centrale à combustion fossile	1-200	50-80%	10%	10 min-1 h
Centrale à cogénération	100-400		7%	1-4 h
Centrale hydroélectrique	50-1300			5 min
Turbine à combustion (fuel ou gaz)	25			15-20 min
Moteur thermique (fuel ou gaz)	20	65%		45-60 min

Par conséquent, les horizons pertinents de prévision peuvent et doivent aller de 5 minutes à plusieurs jours, comme le confirment Diagne *et al.* (2013) et Elliston et MacGill (2010). Il est donc évident que le pas de temps des données prédites varie également (énergie journalière ou horaire, énergie sur 10 ou 20 minutes, puissance…) en fonction des objectifs et de l'horizon de prévision. La Figure I-5 (Diagne *et al.*, 2013, 2012) résume les méthodes de prévision existantes en fonction de l'horizon de prévision, de l'objectif et du pas de temps.

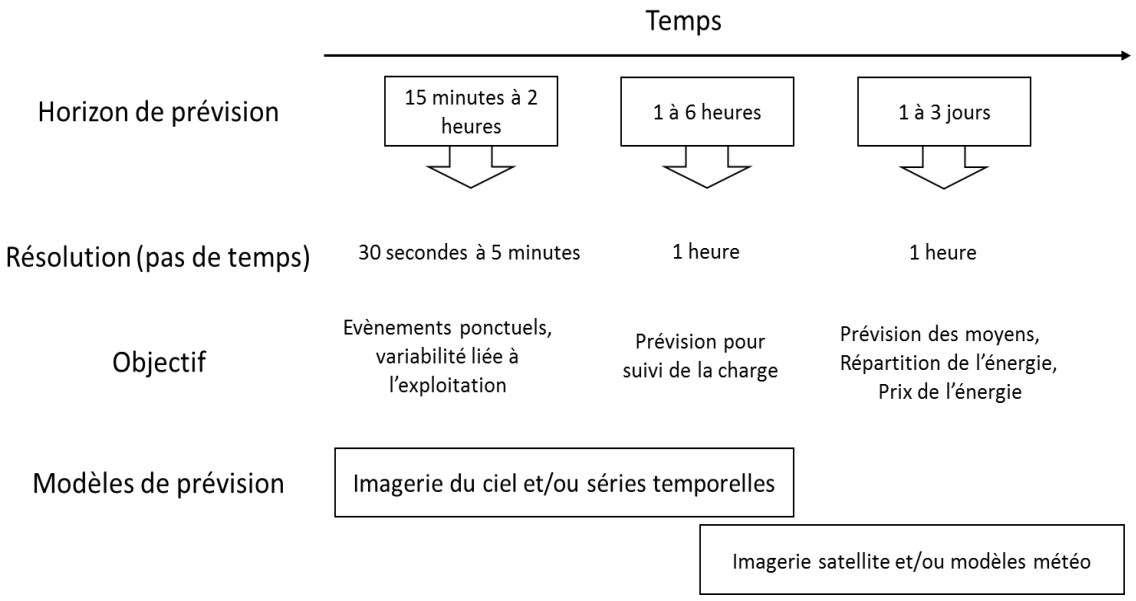


Figure I-5: Relation entre horizon de prévision, modèles de prévision et activités associées (H. M. Diagne et al., 2012 ; M. Diagne et al., 2013)

Dans la section qui suit nous allons aborder la notion de fluctuation des tarifs liée à l'intégration des SERIS.

2.3. Variation du prix de l'électricité en raison des contraintes techniques

Hirth a écrit (2014) : « Si l'électricité était un bien économique comme un autre, la variabilité des énergies renouvelables n'aurait pratiquement aucune incidence. Toutefois, l'électricité présente des caractéristiques particulières, dont la plupart découlent du fait qu'elle ne peut être stockée qu'à un coût élevé. En conséquence, de simples analyses microéconomiques, telles que l'optimisation du bien-être en combinant différentes technologies de production, nécessitent un soin particulier et des outils spécifiques ». La variation temporelle de la production d'électricité, plus importante pour les SERIS, et le travail du gestionnaire de réseau électrique visant à équilibrer cette variation ont une incidence sur le coût énergétique (Borenstein, 2012). Comme souligné précédemment, l'utilisation de l'électricité impose de fortes contraintes. Ces aspects nécessitent un traitement approprié de l'électricité dans les analyses économiques (Joskow, 2008), en particulier pour la production d'électricité intermittente (Joskow, 2011).

Le prix de l'électricité varie dans le temps, l'espace et les délais entre le contrat et la livraison :

- Comme la production et la consommation varient de manière significative, le prix de l'électricité varie dans le temps, parfois de plus d'un facteur 2 (Joskow, 2011), voir même d'un facteur 10 (Borenstein, 2012) en une journée; cette variation quotidienne des prix est rarement observée pour les autres biens ;
- La capacité du réseau électrique limite la quantité d'électricité pouvant être transportée et conduit à des écarts de prix parfois élevés entre des emplacements très proches,
- L'ajustement rapide de la production de la centrale électrique pour assurer l'équilibre entre la production et la consommation, est coûteux et le prix de l'électricité fournie peut être très différent du prix contracté.

Sur les trois dimensions « temps, espace et délai », les écarts de prix se produisent de manière aléatoire et saisonnière. Ainsi, même sur un marché de l'énergie classique, utilisant uniquement des

moyens énergétiques contrôlables, le prix du kWh varie considérablement. Il apparaît que savoir parfaitement quelle sera la consommation électrique (charge) et la production à différents horizons amélioreraient la gestion des différentes sources d'énergie et réduiraient le prix de l'énergie correspondant.

2.4. Le cout de l'intermittence

Toutes ces variations induisent fatalement des surcoûts (Hirth, 2014; Joskow, 2008; Milligan *et al.*, 2011): la production des SERIS ne suit pas la charge et, comme le stockage d'électricité n'est pas illimité et très coûteux, cette variabilité est coûteuse ; la production des SERIS est incertaine jusqu'au dernier moment et, la négociation sur les prix de l'électricité ayant lieu la veille de la livraison, les écarts entre la production prévue et la production réelle, doivent être compensés rapidement.

Le centre de recherches UKERC au Royaume Uni a réalisé une vaste étude sur les impacts de l'intermittence sur la gestion du réseau électrique et les coûts supplémentaires, sur la base de plus de 200 articles internationaux. Un coût est associé à chacune de ces caractéristiques, afin de les comparer économiquement (Figure I-6).

Hirth *et al.* (2015), sur la base d'une revue de la littérature portant sur plus de 100 articles, ont estimé les coûts d'intégration des SERIS et a suggéré de les diviser en trois sous-coûts, en fonction des particularités des SERIS (Hirth *et al.*, 2013) : variabilité temporelle, incertitude et contraintes de localisation; ces trois effets « négatifs » peuvent être réduits par une prévision fiable.

Ces surcoûts peuvent également être divisés en coûts dus aux « impacts d'équilibrage du système » et aux « impacts de fiabilité », le premier relatif aux ajustements rapides à court terme, pour la gestion des fluctuations du pas de temps de la minute à l'heure et le second lié aux incertitudes de la production (Katzenstein et Apt, 2012).

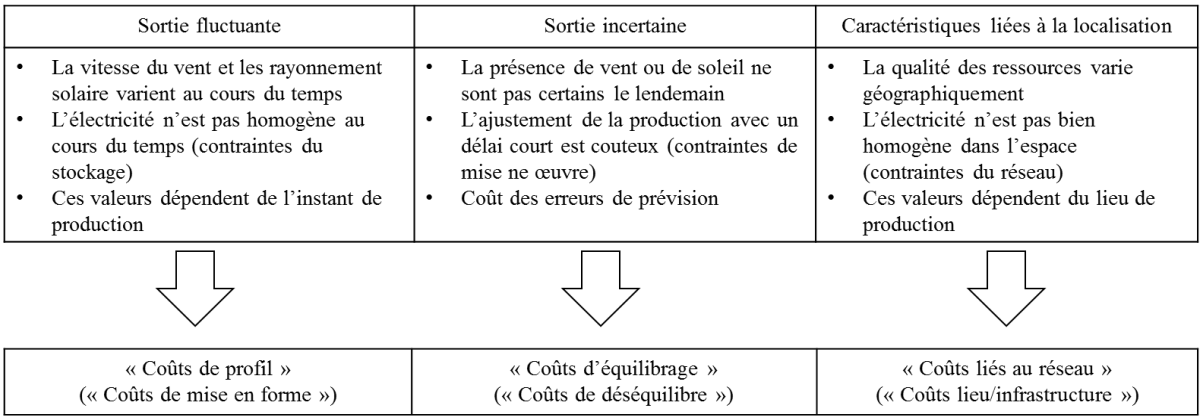


Figure I-6: Caractéristiques des énergies renouvelables et différents coûts induits

Le Tableau I-3 présente les coûts de l'intermittence par MWh pour différentes installations de SERIS, ces coûts sont différents du fait de la différence de localisation (différents pays, etc…) du potentiel en énergies renouvelables, des caractéristiques du réseau électrique etc… Ces résultats sont issus d'une étude bibliographique (Notton *et al.*, 2018) et tous les montants ont été convertis en euros avec le taux de conversion du 1er janvier de l'année de publication du document correspondant.

Tableau I-3: Cout de l'intermittence pour différentes technologies

Type de SERIS	Coût de l'intermittence par MWh	Détails supplémentaires	Référence
SERIS	1-6€	Revue étendue, seule l'augmentation des stockages influence le coût	(Sjoerd Brouwer *et al.*, 2014)
	0-6€	Taux de pénétration modéré, calculé par des modèles	(Hirth et Ziegenhagen, 2015)
	0-13€	Coûts observés, pas d'influence du taux de pénétration, écarts dus aux différents marchés nationaux	(Hirth et Ziegenhagen, 2015)
	25-35€ jusqu'à 50% du coût de production	Taux de pénétration élevé : 30-40%	(Hirth *et al.*, 2015)
	Cout cyclique : 0,36-0,97€ par générateur au fioul	Dans l'ouest des USA, plus de 100 générateurs étudiés (fioul, gaz naturel ou charbon) considérant les démarrages (à chaud ou à froid), fonctionnement à bas rendement et temps de monté en puissance	(Lew *et al.*, 2013)
Eolien	1-4€ < 10% au coût de production	Pour un taux de pénétration >20% basé sur plusieurs études et pays	(Holttinen *et al.*, 2011)
	1,4-2,6€	Dans l'ouest du Danemark, pour le marché Nordique « un jour avant ». Taux de pénétration de 24%	(Holttinen *et al.*, 2011)
	3,08€	Ajout d'un abattement par tonne de CO_2 19,25€/tonne	(Lueken, 2012)
	1,57-4,22€		(Logan *et al.*, 2008; Smith *et al.*, 2006)
	7,3-11,7€ (16% du coût de production)	2,92-4,38€ pour de l'équilibrage court, 4,38-7,30€ pour le maintien d'une marge de disponibilité plus importante. En Grande Bretagne, étude basée sur des opérateurs indépendants	(Gross *et al.*, 2006)
	0,34-6,46€	Basée sur des opérateurs indépendants	(Wiser *et al.*, 2008)
	5.93 ± 0.86€ en 2008, 2.81 ± 0.37 € en 2009	Coût des variations intra horaire pour 20 fermes éoliennes	(Katzenstein et Apt, 2012)
	1,94€	En accord avec l'administration de l'énergie de Bonneville, un tarif de 4,8€ est ajouté pour couvrir les frais d'intégration	(Kinsey Hill, 2008; Logan *et al.*, 2008)
Photovoltaïque	6,16-8,47€	Ajout d'un abattement 25,4-30,8€/tonne CO_2	(Lueken, 2012)
	10€		(Horin *et al.*, 2014)
	5€	Pour une grande centrale PV à Tucson, Arizona, taux de pénétration de 20%	(Gowrisankaran *et al.*, 2015)
	1,43€	Pour une centrale PV de 11 GW en Grande Bretagne	(AER, 2016)
	7,48€	Projection sur 2030 pour une puissance installée de 40 GW	(AER, 2016)
Solaire thermique	3,85€	Avec ajout d'un abattement de 11,55€/tonne CO_2	(Lueken, 2012)

Les coûts de l'intermittence sont très variables au regard des différentes références. On peut cependant considérer que globalement ils varient entre 1 et 12 € par MWh.

2.5. Prévoir pour augmenter les bénéfices des SERIS dans la production d'énergie

Les centrales thermiques au charbon et les turbines à combustion ont les coûts de cycle les plus élevés. Les turbines hydroélectriques et les moteurs à combustion interne présentent les coûts de cyclage les plus bas (Bird *et al.*, 2013). Les turbines à combustion sont bien adaptées aux pics de production elles peuvent, en effet, être démarrées rapidement (Notton, 2015). Ainsi, en utilisant plusieurs prévisions de production ou une prévision la plus fiable possible, les gestionnaires de réseaux peuvent planifier et exploiter efficacement les différents moyens de production, réduisant ainsi la consommation de carburant, les coûts d'exploitation et de maintenance, ainsi que les émissions, par rapport à une simple production «générée» (Porter et Rogers, 2008).

Une action COST (coopération européenne dans le domaine de la science et de la technologie) (*COST action ES1002 Weather intelligence for Renewable Energies (WIRE)*, 2012) sur la prédiction des ressources (WIRE, ES1002) a réalisé une étude bibliographique sur la prévision du vent ; le document final souligne « même si la nécessité et les avantages de la prévision de l'énergie éolienne sont généralement acceptés, peu d'analyses ont examiné en détail les avantages de la prévision pour un service public ». Cependant, certains impacts positifs et importants ont été trouvés dans la littérature.

L'incertitude sur la prévision autrement dit l'erreur de prévision est un paramètre important dans les coûts d'intégration (DeMeo *et al.*, 2007). L'absence de prévisions fiables implique l'utilisation de réserves d'énergie plus importantes, qui ne peuvent pas être utilisées à d'autres fins (Luickx *et al.*, 2010).

Il est plus facile de réaliser une prévision de production sur un ensemble sites de production en même temps, que sur une centrale PV unique, car l'effet de foisonnement a tendance à lisser la production (Figure I-7).

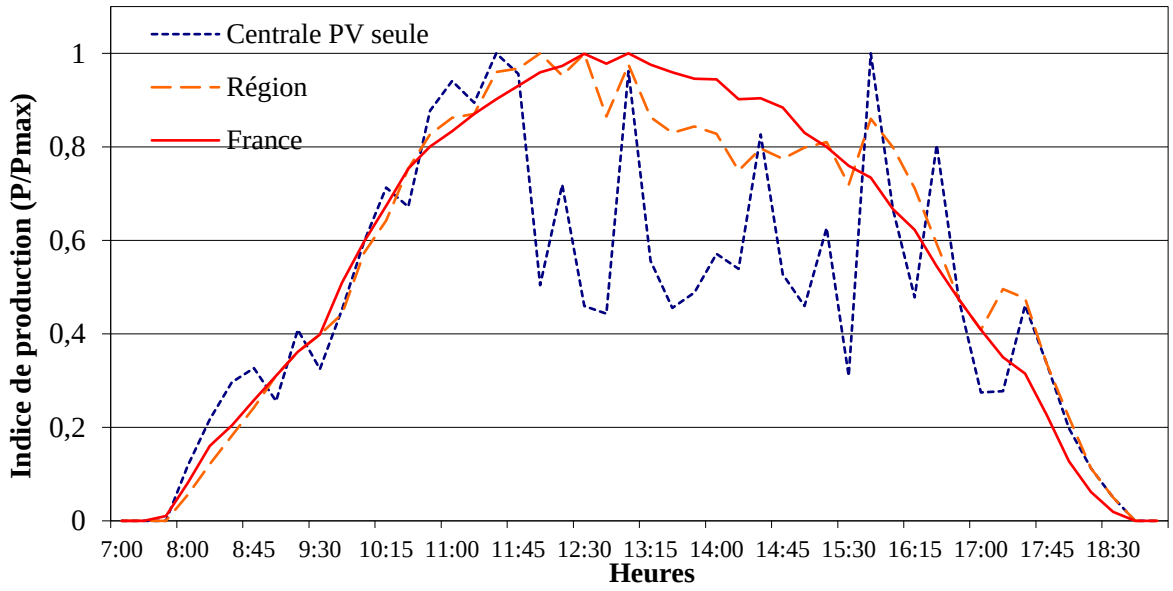


Figure I-7: Effet du foisonnement sur la production

Une étude de (Brancucci Martínez-Anido *et al.*, 2014) sur l'impact de la prévision sur le coût a été réalisée en considérant 7 scénarios: (1) pas d'énergie solaire, (2) pas de prévision d'énergie solaire, (3)

avec prévision de la puissance solaire, (4) avec une amélioration de 25% de cette prévision, (5) avec une amélioration de 50% de cette prévision, (6) avec une amélioration de 75% de cette prévision et (7) avec une prévision parfaite de la puissance solaire.

Les principales conclusions tirées sont :

- Avec un taux d'intégration de l'énergie solaire de 25% chez Independent System Operator New England (ISO-NE) et l'utilisation de méthodes de prévision, les coûts de production nets sont réduits de 22,9% ; Coûts de production nets = coûts de carburant + coûts d'exploitation et de maintenance variables + coûts de démarrage et d'arrêt + coûts d'importation - revenus d'exportation ;
- Avec une prévision améliorée de 25%, les coûts de production nets ne sont encore réduits que de 1,56% et aucune économie significative n'est réalisée pour une amélioration supplémentaire.
- De meilleures prévisions de l'énergie solaire ou une échelle de temps inférieure à une heure pourraient tout de même générer des économies supplémentaires.

L'utilisation d'une méthode de prévision sur un horizon temporel allant jusqu'à 75 minutes pour une centrale PV de 1 MW a permis de réduire les réserves d'énergie flexibles de 21% (5 minutes) et de 16% (15 minutes) par rapport au modèle de persistance et de réduire la probabilité de déséquilibre de 19,65% et 15,12% (Kaur *et al.*, 2016).

La prévision du rayonnement direct normal pour le solaire à concentration reste peu développée par rapport au rayonnement global. Cependant une étude (Kraas *et al.*, 2013) réalisée pour le système CSP de 50 MW, Andasol 3 en Espagne, a conclu que l'utilisation d'un modèle de prévision statistique réduisait le montant des pénalités (dues au marché « day-ahead ») de 47,6% par rapport à l'utilisation d'un simple modèle de persistance.

2.6. Synthèse

La prévision devrait être la première réponse à la gestion de la nature variable de la production basée sur les énergies renouvelables intermittentes, avant la mise en place de stratégies plus coûteuses de stockage de l'énergie et de systèmes de réponse à la demande. En outre, une fois mis en place, le système de prévision offre des avantages supplémentaires grâce à une utilisation optimisée de ces ressources côté demande.

Il apparait que les coûts d'intégration varient énormément lorsqu'on s'y intéresse de plus près. En effet, les méthodes de calculs et la définition des coûts sont déjà très variables d'une étude à l'autre, mais de plus, les applications des systèmes de production, ainsi que les différents systèmes de stockage qui y sont couplés et le taux de pénétration donnent eux aussi un grand éventail de possibilités. Enfin la diversité des conditions météorologiques et les conditions générales d'utilisation des centrales complexifient grandement l'étude objective des coûts d'intégration des systèmes de production aléatoires. Malgré tout il est possible de tirer certaines conclusions qui semblent être le dénominateur commun à la plupart des dispositifs étudiés :

- Les coûts d'intégration dus à l'intermittence et à la variabilité de la production résultent de la production non garantie des SERIS imposée au gestionnaire de réseau électrique qui doit prendre des mesures spécifiques en vue de maintenir l'équilibre production / consommation. Certaines de ces mesures ont un impact négatif sur le fonctionnement des autres moyens de production d'énergie ;

- Ces coûts d'intégration comprennent divers « sous-coûts » pour lesquels une bonne prévision de la production ou de la ressource n'a pas la même influence ;
- Ces coûts d'intégration dépendent du taux d'intégration des SERIS dans le réseau électrique : plus le taux d'intégration est élevé, plus le coût d'intégration est important et plus l'influence d'une bonne prévision sera significative.

Une méthode de prévision fiable pour la production aura une influence très positive sur :

- La réduction des coûts d'intégration ;
- La diminution des coûts d'exploitation annuels moyens ;
- La diminution des réserves manquantes ;
- La réduction des déconnexions d'installations photovoltaïques ou éoliennes.

L'amélioration de la qualité de la prévision a également été étudiée (de l'absence de précision à une fiabilité théorique de 100%) : au-delà d'un pourcentage donné d'amélioration du modèle de prévision, son influence est réduite.

Cet examen montre également que les prévisions sont susceptibles de générer un intérêt économique indéniable pour des centrales de petites ou moyennes puissances. Le développement du stockage de l'énergie nécessite des stratégies d'exploitation spécifiques pour une gestion optimale qui ne peuvent être développées sans une bonne connaissance des flux d'entrée et de sortie à venir.

3. Contexte de la thèse : le projet TILOS

Ces travaux de thèse ont été réalisés dans le cadre d'un projet horizon 2020 auquel participe l'Université de Corse. L'acronyme TILOS signifie « Technology Innovation for the Local scale Optimum integration of battery energy Storage », le projet porte sur l'intégration optimale d'un stockage d'énergie à une échelle locale au sein d'un micro réseau intelligent sur l'île de Tilos dans l'archipel du Dodécanèse, en Grèce. Le montant total du projet, d'une durée de 4 ans (02/2015 à 01/2019) s'élève à 14 M€, dont 11 M€ sont financés par le financement H2020 et 3 M€ par les partenaires ; à cela s'ajoutent l'achat et l'installation des systèmes PV et éolien à la charge d'Eunice. Dans cette partie, nous allons décrire brièvement les caractéristiques du projet et notre participation à celui-ci.

3.1. Problématique

Tilos est une petite île dons la situation énergétique est représentative des régions isolées et peu ou non interconnectées à un réseau électrique conséquent. Cette île se trouve en bout de ligne d'un réseau qui alimente neuf îles interconnectées les unes aux autres et alimentées par les centrales thermiques de Kos et Kalymnos pour une puissance installée de 120 MW et près de 24 MW de systèmes renouvelables intermittents (15,2 MW d'éolien et 8,8 MW de PV). Elle se situe à mi-distance entre les îles de Kos et de Rhodes. La Figure I-8 présente la situation géographique de l'île de Tilos.

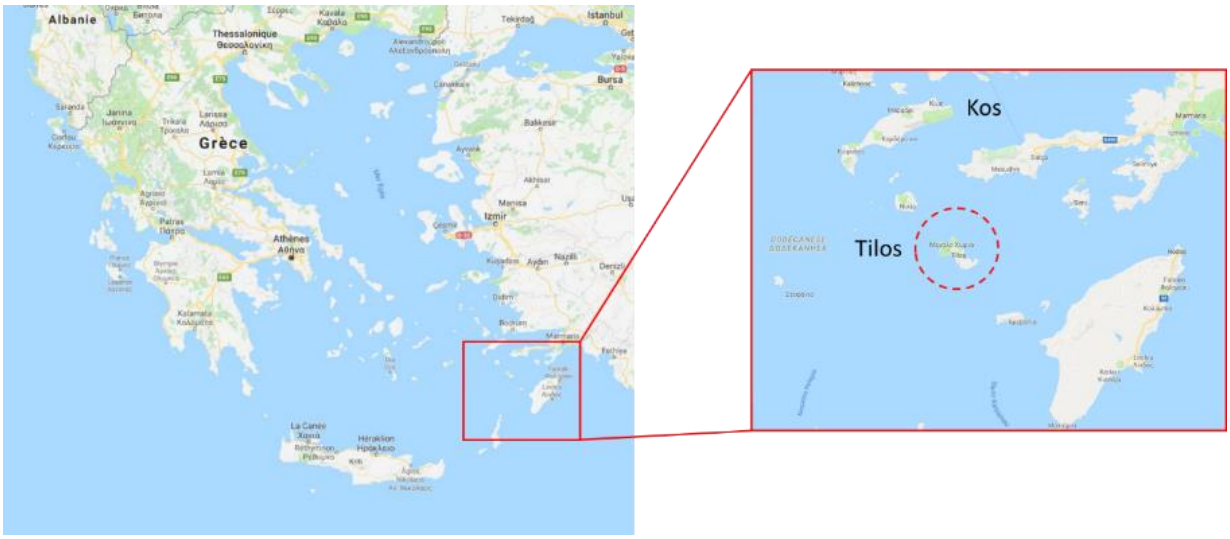
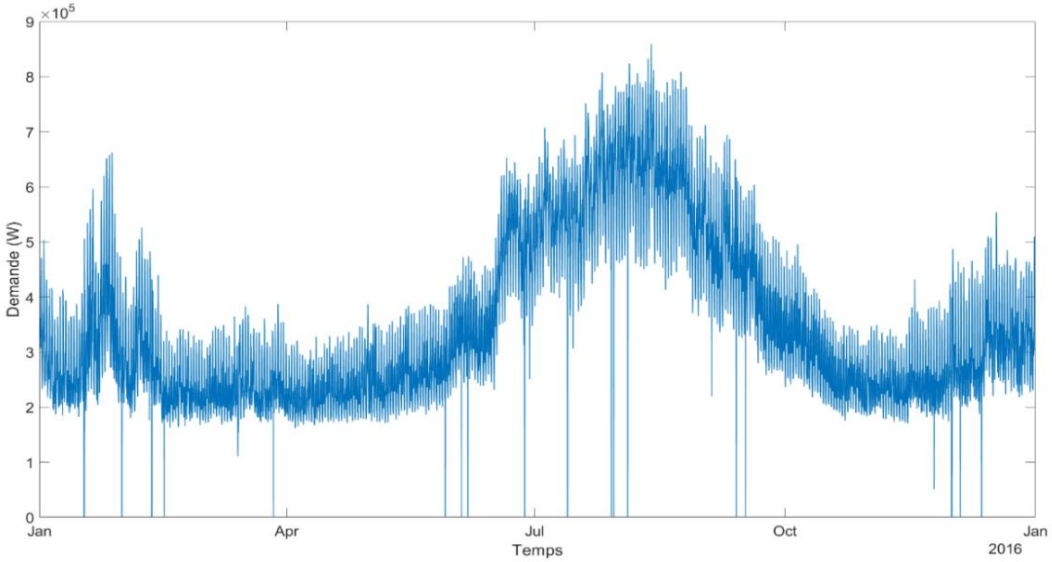


Figure I-8: Situation géographique de l'île de Tilos

Sur Tilos, la population se trouve essentiellement concentrée dans deux villages, Livadia et Megalo Chorio. La population vivant à l'année est d'environ 500 habitants, et l'été celle-ci et doublée voir triplée. Les caractéristiques qui sont propres aux îles de cette taille sont :

- Leur fourniture énergétique dépend en grande partie d'un câble sous-marin qui les relie à des îles de plus grande taille, ce qui tend à diminuer la qualité de l'électricité fournie ;
- Les infrastructures énergétiques, dont elles disposent présentent de sérieux problèmes de fonctionnement.
- De nombreuses déconnexions de systèmes intermittents, dues à la limite de pourcentage de la charge atteinte (30 % du potentiel éolien perdu), impliquent une limite de l'intégration des énergies renouvelables atteinte.
- La forte saisonnalité de la demande énergétique, représentée par les pics estivaux dus à la fréquentation touristique. La Figure I-9 représente la demande en électricité sur l'île de Tilos pour l'année 2016, cette représentation montre bien la forte saisonnalité de la demande.



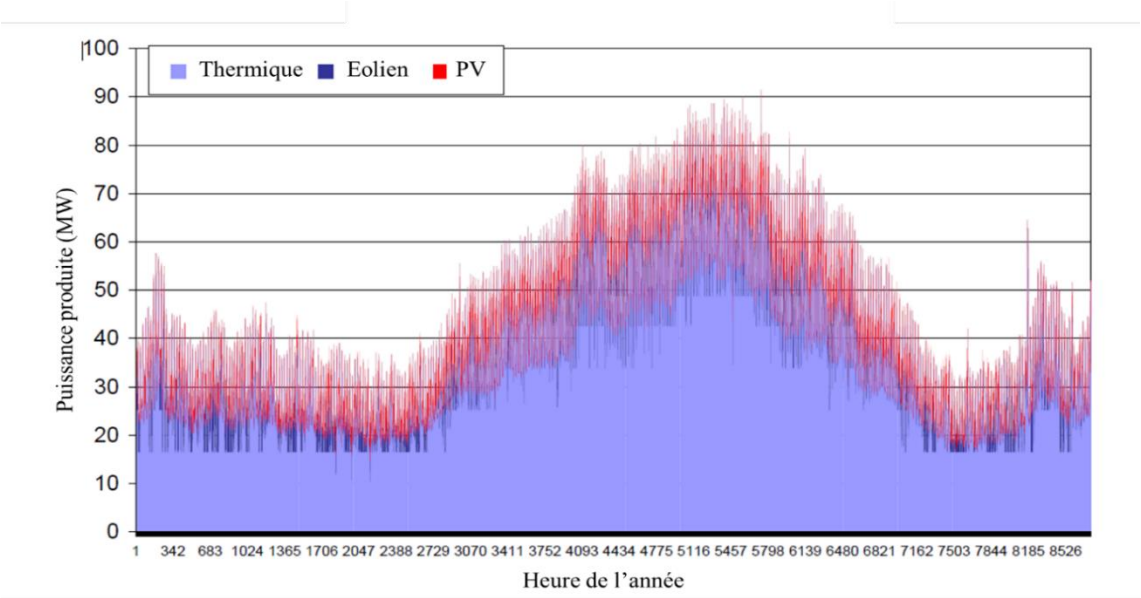


Figure I-9: En haut : Mesure de la demande en fonction du temps sur l'île de Tilos pour l'année 2016. En bas, part des différentes sources

Le réseau électrique de Kos Kalymnos est approvisionné à 85% par le fuel et 15% par les énergies renouvelables.

Ces particularités d'îles isolées et partiellement interconnectées en font d'excellents bancs d'essais pour le développement et l'expérimentation de solutions innovantes pour la production et la gestion des micros réseaux intelligents. L'objectif est de remplacer la production d'électricité à base d'énergie fossile par des systèmes de production d'énergie basés sur les énergies renouvelables et augmenter l'autonomie en utilisant un stockage d'énergie, qui permettrait également d'augmenter le taux d'intégration des énergies renouvelables intermittentes.

D'un point de vue du potentiel en énergies renouvelables, Tilos et aussi un lieu idéal pour l'implantation de systèmes de production solaires et éoliens. La Figure I-10 présente les potentiels solaires et éoliens de Tilos (1750 kWh/m²/an de rayonnement global et 6-7 m/s de moyenne annuelle de vent).

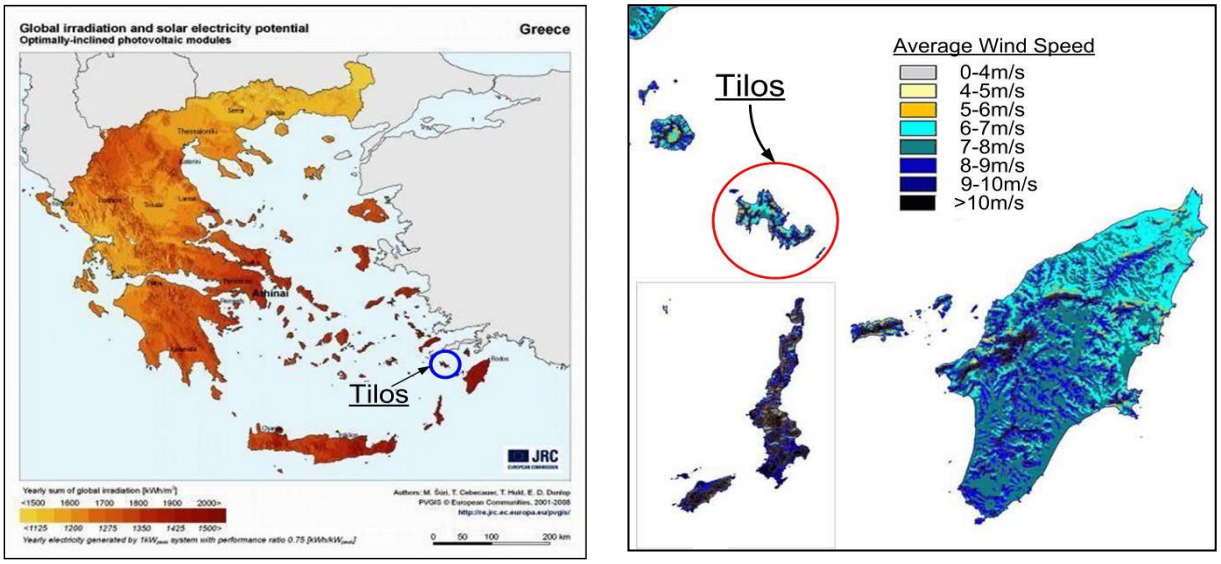


Figure I-10: Cartographie des potentiels énergétiques de l'île de Tilos, à gauche potentiel solaire, à droite potentiel éolien

Le projet TILOS a pour objectif de développer un micro-réseau intelligent sur l'île de Tilos, basé sur un système hybride de production d'énergie, une infrastructure de mesure avancée (mesures de demande électrique, station météo, etc…) et une plate-forme logicielle avancée capable de prévoir la production et la demande d'électricité, avec dans un second temps, un objectif de de planifier l'exportation ou non d'énergie à travers l'interconnexion.

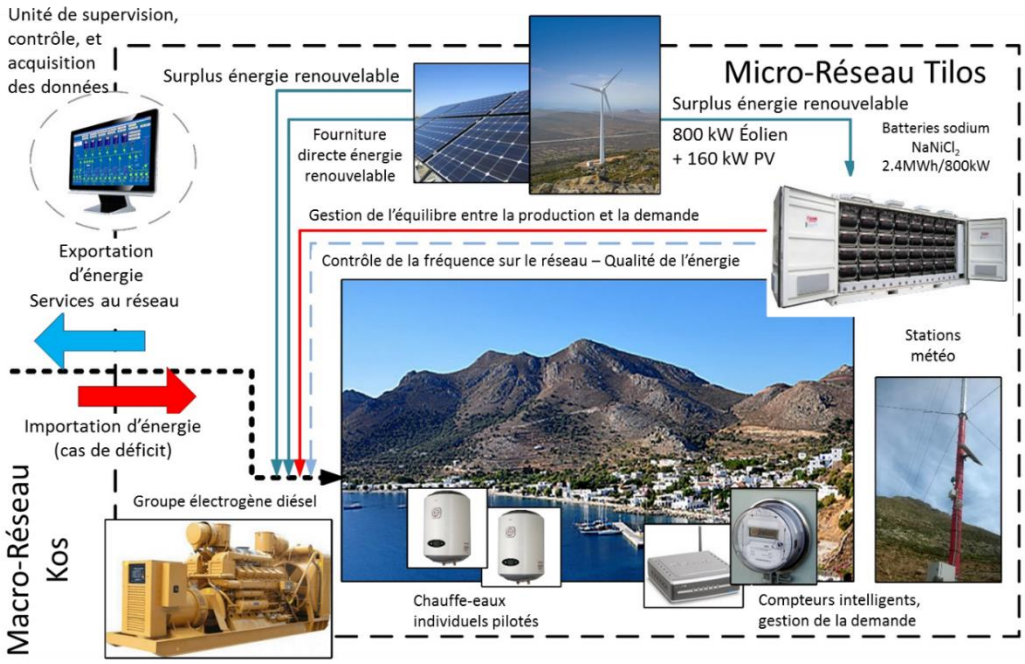
Nous allons maintenant décrire les éléments constitutifs du projet en lui-même et faire ressortir l'implication de notre travail au sein de l'organisation du projet.

3.2. Le micro-réseau intelligent

Le micro réseau de Tilos est divisé en plusieurs parties (représentée sur la Figure I-11) :

- Le système hybride de production, composé d'une éolienne Enercon E53 de 800 kW, d'un champ PV de 160 kW et d'un système de stockage par batteries $NaNiCl_2$ de 1,44 MWh/400kW ;
- Le système de mesures et le dispositif de protection, qui sont constitués par un réseau de dispositifs de mesures (compteurs basse tension intelligents, compteurs de charge moyenne tension, station météorologique) et par des interrupteurs de charge à réenclenchement automatique ;
- La plateforme logicielle, qui contient tous les outils logiciels qui sont utilisés par le micro réseau et qui supporte le système de gestion de l'énergie, le simulateur de micro réseau, la plateforme de prévision, l'analyseur de performance du système de stockage et le système de gestion intelligente de la demande.

La gestion du micro réseau est réalisée par le centre de supervision et d'acquisition des données (SCADA : Supervisory Control and Data Acquisition). Un système de production d'énergie « sauvegarde » alimenté en diesel est présent en cas d'urgence.



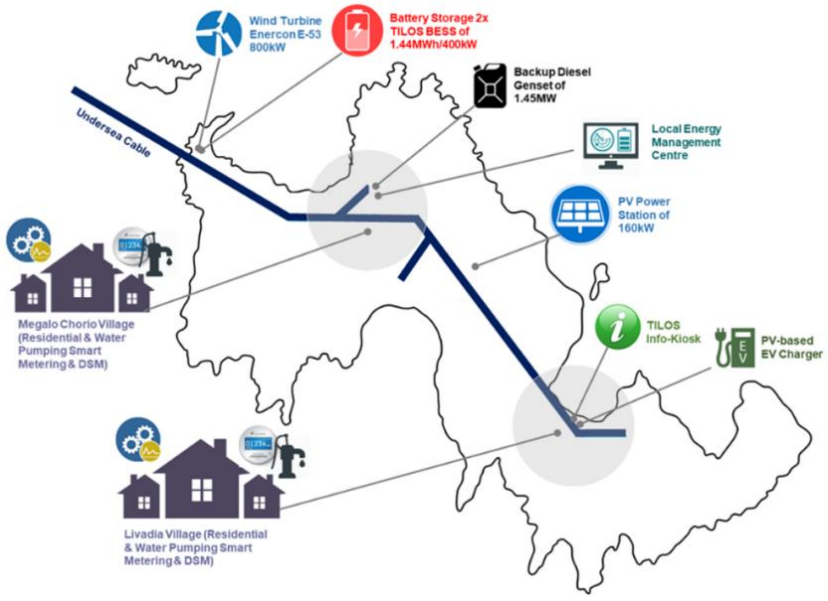


Figure I-11: Constitution du micro réseau intelligent de Tilos et répartition géographique des différents composants

Le micro réseau de Tilos pourra fonctionner selon 3 modes (Figure I-12) :

- Système complétement autonome, mode isolé avec 100% d'énergies renouvelables ;
- Système interconnecté, avec une autonomie de l'ordre de 75% et un complément par l'interconnexion avec le réseau de Kos ;
- Interaction selon les lois du marché, vente et achat d'énergie.

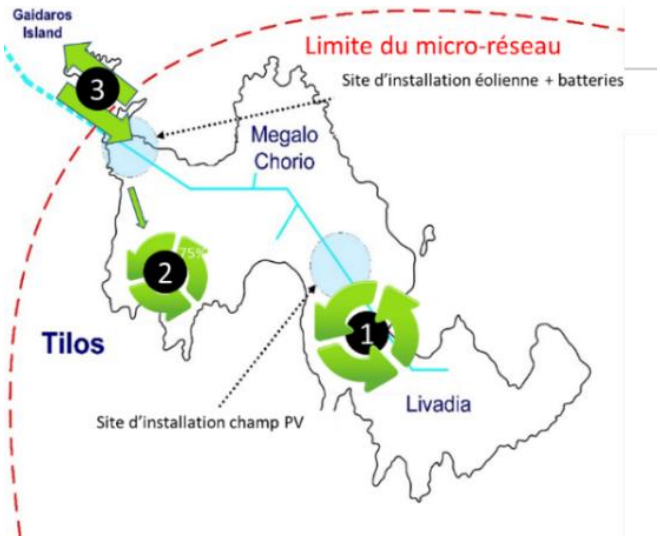


Figure I-12: Modes de fonctionnement du micro-réseau

3.2.1. Le système hybride de production

La centrale hybride de Tilos est une centrale qui combine productions photovoltaïque, éolienne et stockage par batteries de nouvelle technologie $NaNiCl_2$. L'utilisation du stockage d'énergie au sein des systèmes de productions basés sur les énergies renouvelables a été largement étudié (Notton, 2015; Notton *et al.*, 2018), et les effets bénéfiques de ces derniers, notamment pour les réseaux de petite taille, sont maintenant prouvés. Le dimensionnement du réseau de Tilos a été réalisé d'après les résultats des simulations basées sur les travaux de Zhang (2016), Hailong (2016) et Kavadias (2017).

L'installation comprend une éolienne d'une puissance nominale est de 800 kW, installée en Juillet 2017. La production annuelle de cette éolienne est d'approximativement 2,1 GWh, ce qui représente environ 70% des besoins de l'île de Tilos. L'énergie produite par cette éolienne, est à la fois utilisée pour les besoins de l'île mais aussi pour fournir de l'énergie au macro réseau Kos-Kalimnos. Cette production permet d'éviter le rejet de 1350 tonnes de dioxyde de carbone (CO_2), 2 tonnes d'oxydes d'azote (NOx) et 21 tonnes de dioxyde de souffre (SO_2) par an.

Figure I-13: Eolienne sur son emplacement sur l'île de Tilos (source : Facebook "Tilos horizon")

Le champ photovoltaïque (PV), est un champ de petite taille d'une puissance de 160 kW crête, il est constitué de 592 panneaux solaires de 270 W crête chacun. Cette unité de production contribue à hauteur de 265 MWh de production par an, cela représente près de 9% des besoins de l'ile. Cette partie de la production permet, annuellement, d'éviter l'émission de 170 tonnes de dioxyde de carbone (CO_2), 0,25 tonnes d'oxydes d'azote (NOx) et 2,7 tonnes de dioxyde de souffre (SO_2). La Figure I-14 présente le champ de panneaux solaires implanté sur l'île de Tilos.

Figure I-14: Vue du champ photovoltaïque de Tilos (source : Facebook Tilos horizon)

La troisième partie de la centrale hybride de Tilos est le système de stockage de l'énergie par batteries. Elle comprend deux blocs de batteries $NaNiCl_2$ de technologie Zebra, situés à proximité de l'éolienne. Chaque bloc de batteries est d'une capacité de puissance de 400 kW et de 1440 kWh d'énergie. Ces blocs sont desservis par un onduleur d'une puissance de 450 kW. La technologie Zebra est caractérisée par une haute densité énergétique, une disponibilité d'une grande capacité et une grande puissance de sortie tout en ayant une durée de vie assez longue. De plus, leur résistance accrue au stress mécanique et électrique ainsi que leur faible besoin de maintenance et leur recyclabilité les rendent idéales pour l'utilisation pour des systèmes de production d'énergie isolés basés sur les sources renouvelables d'énergie.

Figure I-15: Bloc de batterie avant la livraison sur Tilos

Le système de batteries installé sur Tilos doit assurer différents aspects du fonctionnement du micro réseau :

- Nivellement de la charge, en optimisant la production et la distribution de l'énergie ;
- Augmentation de la qualité de l'énergie, en compensant les fluctuations de puissance avec l'utilisation de la puissance active des batteries combinées à l'onduleur ;
- Optimisation de la production issue des énergies renouvelables, en atténuant les effets de l'intermittence des énergies renouvelables et en réduisant les coupures dues à la limitation ;
- Rendre le micro réseau opérationnel, capacité de fonctionner en mode isolé en maintenant une haute efficacité du réseau ;
- Capacité de démarrer de façon autonome le micro réseau de Tilos et de l'alimenter.

Les actions combinées de ces différents composants de la centrale sont coordonnées par le système de gestion de l'énergie qui est implémenté dans l'infrastructure SCADA au sein de la centrale hybride et dans le contrôle central du micro réseau.

3.2.2. Le système de mesures avancé et les dispositifs de protection

Le système de mesures avancé est une plateforme matérielle et logicielle qui permet de contrôler, en temps réel, par la mesure et la commande, les niveaux de charge sur le réseau (du côté du consommateur). Les échanges d'informations se font au niveau de la SCADA. En recueillant directement les informations au niveau du consommateur, la plateforme peut prendre en charge la gestion de la demande au niveau local. Cette stratégie de gestion permet une pénétration accrue des

sources d'énergies renouvelables, ainsi que l'amélioration du fonctionnement des installations de stockage et de la qualité de la fourniture de services au réseau.

La Figure I-16 présente le prototype de dispositif de mesure avancé et de communication au réseau installé chez les consommateurs. Ce dispositif agit en temps réel et maintient la connexion entre la production et les consommateurs.

Figure I-16: Installation chez le client du système de mesures avancé et du dispositif de communication avec la SCADA (source : Facebook Tilos Horizon)

Les dispositifs de sécurité couplés sur le réseau sont de différentes sortes, en ce qui concerne le réseau de distribution aérien (moyenne tension), deux commutateurs de charge (associés à des compteurs pour mesurer les échanges) accessibles et contrôlables à distance ont été installés à proximité de la jonction entre le micro réseau Tilos et le macro réseau Kos-Kalimnos. En ce qui concerne l'interconnexion, un réenclencheur automatique est en place au niveau de la connexion aux câbles sous-marins qui maintiennent l'interconnexion avec le macro réseau.

Ces deux équipements de protection ont pour vocation d'assurer le fonctionnement du réseau en cas de défaillance et de maintenir la qualité de la fourniture en réalisant :

- La mise en œuvre de la protection contre les défaillances éventuelles, avec une possibilité de détection de ces dernières, dès qu'une défaillance sur l'interconnexion entre Kos et Tilos est détectée, le micro réseau fonctionne de manière isolée avec la prise en charge de la fourniture d'électricité par la centrale hybride ;
- L'application de la stratégie d'exploitation visant à une transition en douceur du mode interconnecté au mode insulaire ;
- Prise en charge d'une fonction de synchronisation pour repasser du mode en îlot au mode interconnecté sans impact perceptible sur la distribution.

Globalement, tous les dispositifs de protection renverront automatiquement les informations et seront contrôlés par le centre SCADA local, en coordination avec le gestionnaire de Kos.

3.2.3. Plateforme logicielle

Le fonctionnement du micro réseau tout entier est basé sur la plateforme logicielle. Elle comprend les applications relatives aux différents travaux de recherche qui ont été implémentés pour le fonctionnement opérationnel, cela va de la simulation du micro réseau, à la gestion de l'énergie, aux taches d'optimisation et à la prévision.

Les éléments de cette plateforme sont les suivants :

- Le système de gestion de l'énergie, qui est une application qui n'est pas uniquement destinée aux besoins de Tilos, mais qui peut fonctionner pour différentes configurations de micro réseaux. Ce système intègre tous les éléments du micro réseau :
  - Il est en relation avec les autres dispositifs de la plateforme
  - Il optimise les interactions entre les composants du micro réseau.

  Ses principales missions sont le « dispatch » du courant selon la prévision de production pour maintenir l'équilibre entre la production, le stockage et le réseau.

- Le simulateur de micro réseau est une boîte à outils logicielle dont le but principal est la conception optimale de micro-réseaux communautaires et insulaires. En utilisant différents éléments énergétiques tels que : les systèmes basés sur les énergies renouvelables, le stockage, les connexions au consommateur, etc... L'analyse et la simulation micro réseau prennent en compte les divers modes de fonctionnement du système dans l'intérêt des différents protagonistes, ainsi que différents critères d'optimisation tels que la sécurité d'approvisionnement, les coûts globaux du système et les performances environnementales. Ce simulateur peut être ajusté pour se conformer à n'importe quel cadre juridique.
- La plateforme de prévision est un système complet qui réalise automatiquement l'exécution des différents modèles pour toutes les variables dont la prévision est nécessaire, comme la demande, la production solaire et la production éolienne. L'objectif, au travers de cette plateforme, est de faciliter la gestion intelligente du réseau. En parallèle, la plateforme stocke automatiquement les résultats des prévisions et les distribue aux différents acteurs du micro réseau. L'accès en temps réel aux bases de données, ou les résultats des prévisions sont stockés, ainsi que la gestion des différents modèles de prévision implémentés sont pris en charge par un service online (plateforme web). Cette plateforme permet l'amélioration à tout moment des modèles qui y sont implémentés par des mises à jour ou l'ajout de nouveaux modèles.
- L'outil d'analyse de performance du système de stockage :
  - Réalise les diagnostics des performances du système de batteries ;
  - Rassemble toutes les informations sur le fonctionnement des batteries ;
  - Analyse les données du système lui-même ;
  - Permet d'améliorer l'utilisation du stockage, de détecter d'éventuels dysfonctionnements
  - Prévoit l'état du système ainsi que la disponibilité des composants.

  Ce travail en amont permet de planifier les interventions de maintenance pour améliorer le fonctionnement global du système.

- L'interface de gestion du réseau intelligent du point de vue du consommateur (SM-DSM pour Smart Microgrid-Demand Side Management) :
  - Rassemble les informations collectées par les dispositifs de mesures de la demande ;
  - Permet leur contrôle et leur gestion via une interface homme machine.

Tous les consommateurs qui disposent d'un dispositif de gestion ont ainsi accès à ces données.

Cette plateforme logicielle est le centre névralgique du micro réseau et concentre tous les outils qui servent durant l'exploitation.

3.3. Fonctionnement global

La globalité du micro réseau intelligent fonctionne en temps réel et tous les composants sont en interaction. La Figure I-17 présente la globalité du projet et les différentes interactions au sein du système.

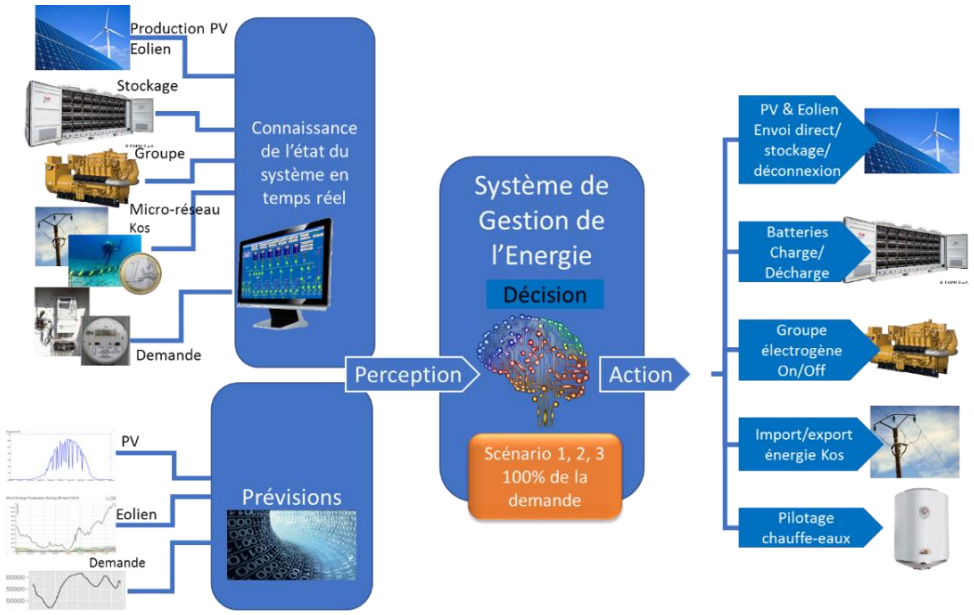
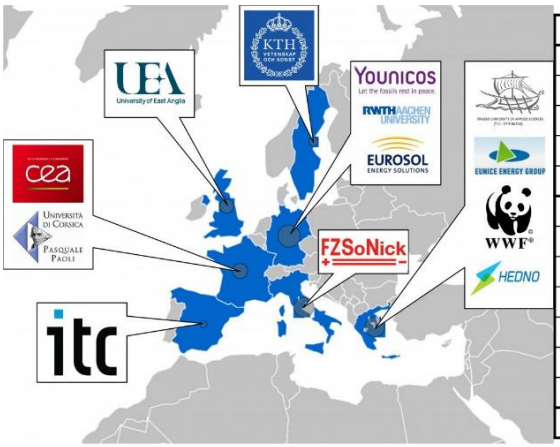


Figure I-17: Vue d'ensemble du micro réseau et de son fonctionnement

Notre participation au projet se divise en trois missions : la principale concerne le développement de modèles à court terme de prédiction de la production PV implémentés dans la plateforme de prévision, la seconde consiste à étudier la possibilité d'utiliser de tels systèmes en Corse avec un dimensionnement différents des composants et enfin, le développement d'un code de simulation d'une chaine de stockage hydrogène introduit dans un logiciel de simulation de systèmes hybrides.

Les prévisions, qui sont réalisées en permanence, alimentent la gestion du micro réseau et permettent ainsi l'optimisation de cette gestion. Le projet arrive dans sa phase de finalisation et la plupart des dispositifs sont installés ou en cours d'installation.

La réalisation de ce projet implique la participation de plusieurs partenaires (Figure I-18) qui interviennent dans leur cadre de compétence. Les partenaires sont répartis en Europe et sont académiques, industriels ou organismes non gouvernementaux. La totalité du projet est divisé en groupes de travail (work packages), toutes les tâches sont réparties en fonction des compétences des partenaires et la coordination est réalisée par le porteur du projet.



Partenaire				Rôle principal
FZSonick	FZSoNick	IT	Industrie	Système de batteries
Younicos	Younicos	DE		Système de gestion de l'énergie
EUNICE	EUNICE	GR		EPC - RES Assets Investor & Operator
EUROSOL	EUROSOL	DE		Système de gestion côté consommateur
HEDNO	HEDNO	GR		Distributeur
TEIP		GR	Recherche & Technologie	Coordination
CEA	cea	FR		Optimisation du système de batteries
ITC	itc	ES		Prévision
UEA	UEA	UK		Modèles économiques
UCCP		FR		Prévision solaire
RWTH	RWTH AACHEN UNIVERSITY	DE		Simulations
KTH		SE		Dissémination
WWF		GR	ONG	Engagement de la population

Figure I-18: Partenaires du projet TILOS et leurs principales missions (www.tiloshorizon.eu)

En Grèce, l'université du Pirée (http://www.unipi.gr/) (aujourd'hui université de West Attica) (~~qui~~) est l'organisme porteur du projet et spécialisé dans les énergies renouvelables, Eunice (http://eunice-group.com/) gère la mise en œuvre de la centrale ainsi que la station météorologique implantée sur l'île et le centre de contrôle. WWF en Grèce (http://www.wwf.gr/en/) gère toute la partie communication avec la population et la sauvegarde des espèces protégées sur le site. HEDNO (https://www.deddie.gr/en/i-etaireia/profil) gère du réseau électrique grec.

En Italie, l'entreprise FZSoNick (http://www.fzsonick.com/en/) est le fabricant du dispositif de stockage $NaNiCl_2$.

En Allemagne, Younicos (https://www.younicos.com/) est une entreprise spécialisée dans les énergies renouvelables pour le dimensionnement et les algorithmes de pilotages des installations. L'université de Aachen (https://www.rwth-aachen.de/) intervient sur le volet algorithme de pilotage. L'entreprise Eurosol (https://www.eurosol.eu/) propose des solutions de gestion de l'énergie électrique par l'utilisation des compteurs intelligents.

En Suède, l'université KTH (https://www.kth.se/en) est responsable de toute la partie dissémination des résultats et communication.

En Angleterre, l'université UEA (https://www.uea.ac.uk/) est chargée de la mise en place du business modèle.

En Espagne, l'institut technologique des Canaries (ITC) est responsable de la gestion des serveurs de données et de logiciels et développe des modèles de prévision de vent, de production photovoltaïque et de demande d'électricité.

En France, le Commissariat aux Energies Alternatives (CEA) réalise la modélisation numérique des batteries et une partie de la prévision de production photovoltaïque.

4. Synthèse

Aujourd'hui, les enjeux de la production d'électricité au niveau mondial sont multiples. Il est nécessaire qu'ils répondent à une logique à la fois environnementale et économique. Les besoins énergétiques en constante augmentation et la nécessité de diminuer l'impact environnemental conduisent, obligatoirement, au développement d'une intégration grandissante des sources renouvelables d'électricité. Cette augmentation de la part des énergies renouvelables au sein du mix énergétique pose néanmoins de nouveaux défis aux gestionnaires de réseaux. En effet, les

caractéristiques propres aux énergies renouvelables intermittentes compliquent considérablement la gestion de l'approvisionnement électrique.

D'un point de vue économique, ces incertitudes et la complexification de la gestion induisent fatalement des surcoûts qui ne peuvent pas être ignorés.

La réalisation de projets innovants, comme le projet TILOS, apporte une nouvelle façon d'approvisionner en électricité des sites isolés ou partiellement interconnectés. Avec une production essentiellement basée sur les sources renouvelables d'énergie et le développement de systèmes intelligents, il devient alors possible de produire et de distribuer une énergie de qualité, propre et d'assurer la stabilité du réseau.

Ces nouveaux modes de production et de gestion s'inscrivent parfaitement dans l'air du temps en ce qui concerne l'impact environnemental et économique de la production et de la gestion d'électricité. Dans ce contexte, il devient alors de plus en plus important de développer des modèles de prévision fiables afin de donner aux gestionnaires de réseaux les outils nécessaires pour accroître la part d'utilisation des énergies renouvelables intermittentes, c'est la problématique à laquelle nous nous proposons dans ce document de participer.

II. Collecte et prétraitement des données

Dans ce chapitre, nous allons présenter de manière détaillée les données que nous avons utilisées pour cette étude ainsi que l'ensemble des traitements que nous utiliserons lors de la modélisation du rayonnement solaire. Les hypothèses, les approximations ainsi que tout le formalisme nécessaire seront exposés.

1. Les séries temporelles

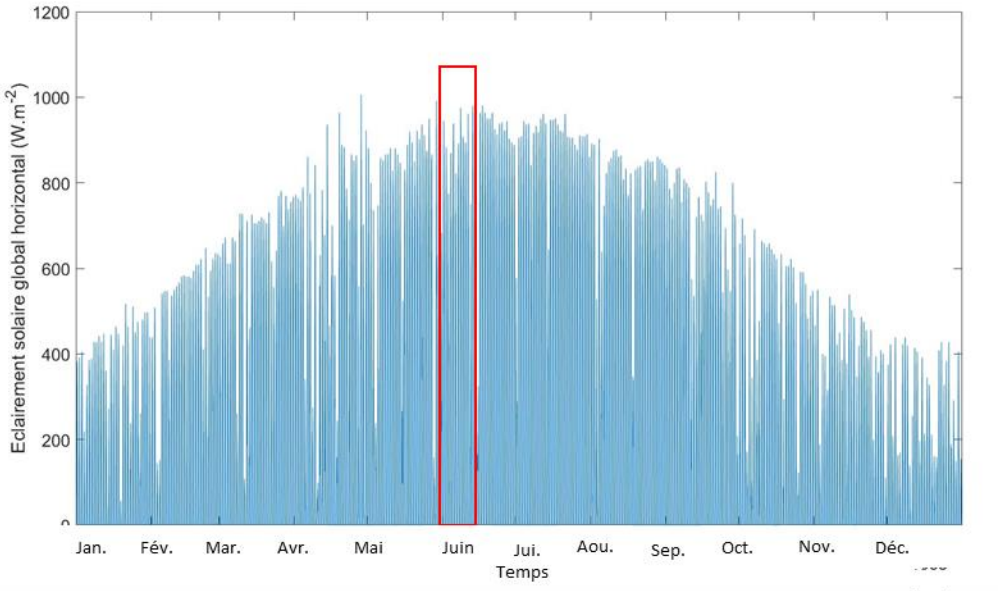
Une série temporelle (parfois appelée chronique), est une succession d'observations mesurées au cours du temps et qui représente un phénomène donné. Par hypothèse, on considère que l'intervalle temporel est régulier et qu'il n'y a pas de modification dans le protocole d'acquisition. En ce qui concerne les observations de phénomènes naturels, plus particulièrement du rayonnement solaire, il est difficile d'obtenir des séries de mesures fiables sur de grandes périodes ; en effet les critères d'acquisition des données doivent rester « théoriquement » inchangés durant toute la campagne de mesure : pas de temps constant, pas de dérive des capteurs, maintenance et suivi régulier des installations de mesures. Il est difficile de respecter scrupuleusement toutes ces obligations, en particulier en ce qui concerne le contrôle de la dérive des appareils ; cependant, nous sommes contraints à faire l'hypothèse que le protocole de mesure varie « peu » au cours du temps et, dans la suite du document nous allons ainsi parler de pseudo séries temporelles, afin de pouvoir utiliser un formalisme mathématique rigoureux.

Si on appelle x_t la valeur d'une grandeur x mesurée à l'instant t (compris entre 1 et n), alors on peut considérer qu'une série temporelle X (vecteur) de la grandeur x s'écrit :

$$X_t = (x_1, x_2 \dots x_t \dots x_n) \text{ aussi noté } X = \{x_t : t \in [1, n]\} \qquad \text{(II-1)}$$

Traditionnellement, en première analyse, il est très utile de représenter graphiquement l'évolution temporelle d'un phénomène en portant en ordonnée la valeur de l'observation et en abscisse le temps t. Dans le cas des données de rayonnement solaire, cette représentation est une première façon de s'assurer de la qualité des mesures.

La Figure II-1 présente deux années de mesures d'irradiation solaire globale horizontale à Ajaccio, il s'agit de données acquises avec une résolution horaire.



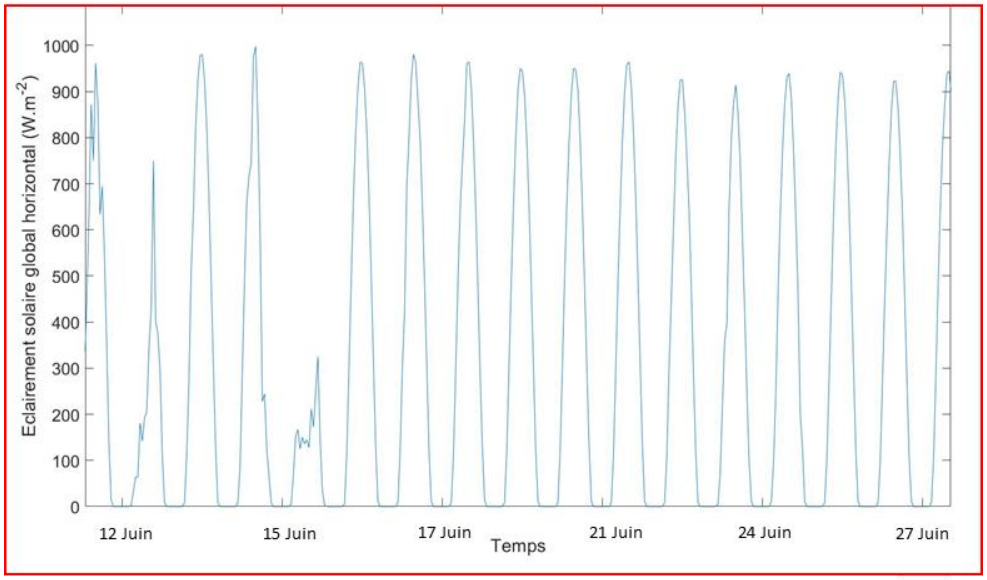


Figure II-1: Représentation graphique de la série temporelle de l'irradiation solaire globale horizontale mesurée à Ajaccio. Agrandissement pour illustrer la périodicité journalière

Cette vue d'ensemble des données mesurées permet une première évaluation qualitative du jeu de données et permet en particulier d'avoir un aperçu des ordres de grandeur, de juger de la présence de « trous » et d'évaluer la périodicité (double dans le cas du rayonnement solaire : 24 heures et 365 jours).

Ce profil de mesures (Figure II-1) montre une série de mesures « propres », on entend par propre le fait que la série temporelle contienne moins de 2% de trous ou de valeurs aberrantes (« outliers ») (David *et al.*, 2016). Cette analyse visuelle (et subjective) ne révèle pas d'anomalie majeure, cependant elle est loin d'être suffisante, les outils mathématiques de vérification objective plus poussés doivent être utilisés.

Les fondements de l'analyse des séries temporelles sont basés sur la décomposition, l'analyse des sous-séries puis la recomposition de la chronique. Cette approche, par décomposition, suppose que la série temporelle peut être séparée en éléments plus simples qui sont modélisables, afin d'être ensuite recomposée pour donner la prévision.

Les études historiques sur le sujet (Bourbonnais et Terraza, 2008) ont permis de standardiser la décomposition des séries temporelles en trois composantes :

- La tendance, ou composante extra-saisonnière, notée E_t, qui représente la variation sur le long terme i.e. son évolution fondamentale ;
- La composante saisonnière S_t qui est une composante cyclique, avec une période intra-annuelle, qui se reproduit de façon plus ou moins permanente d'une année sur l'autre ;
- La composante résiduelle, couramment notée R_t, qui rassemble toutes les variations inexprimées par les deux autres composantes.

Cette procédure de décomposition/recomposition repose sur une modélisation particulière qui l'autorise. On parle alors de schéma de décomposition. On le retrouve principalement sous trois formes :

- Le schéma additif, celui-ci suppose l'orthogonalité (l'indépendance) des différentes composantes de la série temporelle :

$$x_t = E_t + S_t + R_t \qquad \text{(II-2)}$$

- Le schéma multiplicatif, lorsque les composantes extra saisonnières et saisonnières sont liées entre elles mais pas avec la composante résiduelle :

$$x_t = E_t.S_t + R_t \quad \text{(II-3)}$$

- Le schéma multiplicatif complet, dans lequel les trois composantes interagissent :

$$x_t = E_t.S_t.R_t \quad \text{(II-4)}$$

Les interrogations qui subsistent et qui sont importantes, lorsqu'on s'intéresse aux séries temporelles, concernent la présence ou non d'une saisonnalité et le type de schéma qu'il convient de retenir.

Dans le cas du rayonnement solaire, la physique nous donne les réponses à ces interrogations, avec la présence de périodicités journalières et annuelles telles qu'illustrées sur la Figure II-1. Il semble qu'il y ait un consensus au niveau des chercheurs pour considérer le schéma multiplicatif simple (Equation II-4) dans le cadre de la modélisation de chroniques de rayonnement solaire.

L'analyse plus poussée des séries temporelles sera décrite de manière détaillée dans le paragraphe 3. La partie suivante est consacrée aux différents sites étudiés.

2. Les quatre sites expérimentaux

Dans le cadre de cette étude, nous avons porté une attention particulière à pouvoir comparer les modèles développés sur différents jeux de données.

En effet, de manière à pouvoir tirer des conclusions sur la robustesse et l'adaptabilité des modèles, il était nécessaire de pouvoir objectivement tester les méthodologies sur des données provenant de sites dont les caractéristiques géographiques, et par la même météorologiques, sont différentes. La collecte des données a permis de rassembler des mesures en provenance de quatre sites géographiquement éloignés avec d'importantes différences au niveau des caractéristiques météorologiques et en termes de variabilité des données solaires. Les quatre stations sont :

- Ajaccio (41°54'46"N/8°39'13"E, alt. 4 m, Corse, France) ;
- Tilos (36°25'04"N/27°22'59"E, alt. 96 m Dodécanèse, Grèce) ;
- Odeillo (42°29'36"N/2°01'47"E, alt. 1650 m, Languedoc Roussillon, France) ;
- Nancy (48°40'00"N/6°09'34"E, alt. 271 m, Grand Est, France).

La Figure II-2 présente la situation géographique des différents sites étudiés.

Figure II-2: Répartition des sites de mesures

Comme nous pouvons le voir les sites sont répartis en Europe et plus ou moins distants de la mer.

- Ajaccio : la station de mesures se situe sur le site de notre laboratoire lieu-dit Vignola. Les étés sont secs et chauds et les hivers doux et humides. La station se situe à proximité de la mer et les premières montagnes (altitude >1000 m) se trouvent à une vingtaine de kilomètres. La Figure II-3 présente la situation du laboratoire et de la station de mesures.

Figure II-3: Situation de la station de mesures d'Ajaccio

- Tilos : la station de mesures de Tilos est intégrée au champ photovoltaïque de la centrale hybride construite dans le cadre du projet Horizon 2020. Le climat sur l'île de Tilos est caractérisé par des étés secs et chauds et des hiver doux, avec peu de précipitations sur l'année. Le climat y est qualifié de semi-aride. La station de mesures se trouve au centre de l'île relativement proche de la mer et au milieu du relief, le point culminant de l'île à une altitude de 654 m et se trouve à moins d'une dizaine de kilomètres. La Figure II-4 présente une vue d'ensemble de l'ile de Tilos ainsi que la station de mesures de l'éclairement solaire.

Figure II-4: Vue d'ensemble de l'île de Tilos et d'un dispositif de mesures

- Odeillo : les données d'Odeillo proviennent du laboratoire Procédés Matériaux Energie Solaire (PROMES CNRS UPR 8521). La station de mesures se situe en montagne, à 1650 mètres d'altitude et est relativement éloignée de la mer (environ 80 kilomètres), le climat est montagnard mais avec une influence méditerranéenne. Les étés sont assez chauds mais perturbés (orages fréquents), les hivers sont secs et froids avec des précipitations neigeuses. Les montagnes environnantes dans un rayon de 10 km ont une altitude supérieure à 2500 m.

Figure II-5: Situation et station de mesures du four solaire d'Odeillo

- Nancy : la station de mesures de Nancy se trouve en zone urbaine (IUT de Nancy Brabois). Nancy est située à 650 km à l'Est de l'Océan Atlantique et à 600 km au Nord de la mer Méditerranée. De ce fait, le climat qui y règne est de type semi-continental. Les hivers sont froids et secs par temps de gel. Les étés ne sont pas toujours ensoleillés mais assez chauds. Les brouillards sont fréquents à l'automne et les vents rares et peu violents. La Figure II-6 présente une vue de l'IUT Nancy Brabois d'où proviennent les mesures d'ensoleillement.

Figure II-6: Campus de l'IUT de Nancy Brabois

Les quatre stations de mesures sont géographiquement éloignées et se trouvent dans des zones qui possèdent des dominantes climatiques différentes. L'intérêt pour notre étude est de tester la robustesse des modèles et de notre protocole de traitement et d'utilisation des données. Il sera alors intéressant de comparer les modèles et leur performance par rapport à des conditions climatiques différentes.

Dans le Tableau II-1, nous avons consigné l'ensemble des données collectées et utilisées au cours de cette étude.

Tableau II-1: Données disponibles et utilisées

Données / **Lieu**	**Type de données de rayonnement**	**Résolution temporelle**	**Années disponibles**
Ajaccio	Global horizontal	Minute	3
Nancy	Global horizontal	Minute	2
Odeillo	Global horizontal Direct normal Diffus horizontal	Minute	3
Tilos	Global horizontal Global incliné	Minute	2

La collecte des données est une tâche essentielle, les mesures in situ de bonne qualité sur de grands intervalles temporels sont difficiles à obtenir.

Dans le paragraphe suivant, les différentes étapes du prétraitement des données de rayonnement solaire seront exposées.

3. Préparation des données

Avant d'utiliser les données pour réaliser une prévision, il est nécessaire d'effectuer plusieurs prétraitements. Ces étapes ont pour but de rendre les données mesurées utilisables par les méthodes que nous mettons en œuvre lors de la modélisation. L'utilisation d'un prétraitement des variables d'entrées

implique l'ajout d'une étape supplémentaire de post-traitement des variables de sortie des modèles pour qu'elles correspondent à la variable initiale (on parlera alors de recoloration de la chronique), on pourra ainsi comparer les prévisions en sortie de modèle avec les données d'entrée (modélisation univariée, avec données endogènes).

Avant d'effectuer la phase de prétraitement, il convient de s'assurer que les données ne présentent pas de défaut : si elles sont conformes au contrôle de qualité elles pourront alors être envoyées au prétraitement afin d'alimenter les modèles et de réaliser une prévision. Dans le cas contraire, elles pourront, dans certains cas, subir des corrections avant utilisation.

Nous allons présenter le protocole de traitement des données étape par étape, en répondant aux problèmes rencontrés. Nous commencerons par le contrôle qualité, la gestion des données manquantes, puis la filtration. Il sera ensuite question des traitements plus mathématiques avec la stationnarisation et la recherche des corrélations entre et au sein des séries temporelles étudiées, et enfin le partitionnement, la construction des matrices et la validation croisée de l'apprentissage.

3.1. Contrôle qualité

Le contrôle qualité des observations radiométriques comprend divers tests visant à établir des limites à l'intérieur desquelles des données sont dites acceptables. Ces tests sont guidés par un raisonnement physique (pour détecter des événements physiquement impossibles). La plupart de ces tests peuvent être utilisés de manière totalement automatisée ; cependant, certains tests pour détecter des erreurs spécifiques (par exemple, des calibrages faussés des instruments ainsi que des impacts de l'ombre et de la neige) nécessitent une décision prise par un opérateur humain. La Figure II-7 rassemble quelques exemples d'évènements qui peuvent impacter une série de données mesurée.

Figure II-7: Exemples de perturbation de la mesure, de gauche à droite et de haut en bas : glace, condensation, présence d'oiseau

Parmi les données dont nous disposons, certaines sont brutes directement issues de l'acquisition. De ce fait, il est possible qu'elles contiennent des erreurs qui peuvent provenir de l'appareil de mesure lui-même (salissures ou autres) ou de la chaîne d'acquisition (coupure, problèmes de connexion…). Dans le cadre de la conférence de l'Union Européenne pour les Géosciences (European Geosciences Union), Espinar *et al.* (2012) ont réalisé un rapport basé notamment sur les travaux de Muneer et Fairooz (2002) ainsi que Geiger *et al.* (2002a) sur le contrôle qualité des données météorologiques. Ce rapport rassemble les différentes façons d'encadrer des données collectées afin de s'assurer de leur qualité en utilisant des critères objectifs de validation. Nous effectuons un contrôle qualité dit par les « extrêmes », les valeurs

mesurées d'éclairement solaire sont encadrées par une enveloppe représentative des limites physiques de l'éclairement. Dans le cas des données au pas de temps de la minute, les encadrements pour chaque composante du rayonnement sont donnés dans les équations suivantes :

- Pour l'éclairement global horizontal :

$$0 < \dot{I}_g < \min(1{,}2.\, I_0; 1{,}5.\, I_0.\cos(\theta_z)^{1,2} + 100) \quad \text{II-5}$$

- Pour l'éclairement direct normal :

$$0 < \dot{I}_b < I_0 \quad \text{II-6}$$

- Pour l'éclairement diffus horizontal :

$$0 < \dot{I}_{db} < \min(0{,}8.\, I_0; 1{,}5.\, I_0.\cos(\theta_z)^{1,2} + 50) \quad \text{II-7}$$

Au cours de l'étape de contrôle qualité, il convient également de gérer les données manquantes, point très critique. Une série temporelle étant basée sur le suivi d'une grandeur au cours du temps, toutes les mesures manquantes sont une perte importante d'information. Souvent, dans les logiciels comme Matlab ou R statistiques, ces données manquantes sont remplacées par la valeur NaN (« Not a Number » en anglais) et aucune opération mathématique n'est réalisable sur ce type de données. De ce fait, lorsqu'une variable est absente elle devient inutilisable. Cela pose alors le problème de la réponse à ces données manquantes. En première approche, on pourrait penser qu'il est judicieux de supprimer les lignes pour lesquelles des données sont absentes, cependant cette réponse drastique engendre la perte d'information ainsi que la rupture dans la dimension temporelle de la série de données. C'est pour cette raison qu'il est nécessaire d'avoir une approche plus nuancée.

Lorsque ces données manquantes sont trop étendues dans la série temporelle, celle-ci devient quasiment inutilisable. Il est possible de remplacer les données pour combler ces manques lorsque ces derniers ne sont pas trop étendus.

Dans le cadre du projet TILOS, ces problèmes étant récurrents en début de projet, nous avons mis en place un protocole de traitement des données manquantes. Sur les données qui ont une résolution « minute » nous avons déterminé deux cas de remplacement :

- Pour un manque de données sur un intervalle inférieur à 5 minutes les données sont remplacées par une interpolation entre les deux bornes de l'intervalle ;
- Pour un manque de données entre 5 minutes et 30 minutes, les données sont remplacées par des données prévues par un modèle de persistance intelligente.

Ces méthodes de remplacement ont été éprouvées dans le cadre des prévisions réalisées pour le projet TILOS et donnent une réponse assez intéressante pour ces intervalles de données manquantes peu étendus. D'autres possibilités ont été utilisées comme des remplacements par des modèles semblables à des modèles de prévision mais ces méthodes sont quelque peu contestables et nous avons pu remarquer que, contrairement au remplacement par des mesures, la présence de ces données estimées ou prévues sur un intervalle conséquent avait un effet néfaste sur la prévision proprement dite.

Une dernière vérification des données concerne leur concordance avec la base temporelle. En effet, il existe différentes échelles de temps : temps légal (celui de la montre), temps universel (UTC), temps solaire moyen, temps solaire vrai … nous avons décidé d'indexer les données suivant le temps UTC (Coordinated Universal Time), cela permet d'uniformiser les échelles des jeux de données et de s'affranchir des contraintes légales propres à chaque pays (comme le changement d'heure en été ou en hiver) (Iqbal, 1983).

Le

Tableau II-2 rassemble les résultats obtenus après le contrôle qualité des données.

Tableau II-2: Données et contrôle qualité

Données	Nombre de points	Données aberrantes	Données manquantes	Echelle temporelle
Global horizontal Ajaccio	1 578 240	<2%	Non	UTC+1
Global Horizontal Tilos	1 054 080	Non	<2%	UTC+2
Global horizontal Odeillo	1 576 800	Non	<2%	UTC+1
Direct normal Odeillo	1 576 800	Non	<2%	UTC+1
Diffus horizontal Odeillo	1 576 800	Non	<2%	UTC+1
Global horizontal Nancy	1 051 082	<2%	Non	Temps Légal => UTC+1

Comme le montre le Tableau II-2 récapitulant les contrôles qualité effectués sur les données, on constate que les données dont nous disposons sont de « bonne qualité » car elles proviennent de stations où le suivi des instruments est régulier.

La qualité des données ayant été vérifiée, nous pouvons débuter les étapes de prétraitement. La première d'entre elles consiste à stationnariser les données de rayonnement solaire après avoir réalisé une modélisation de l'éclairement solaire par conditions de ciel clair, c'est l'objet du paragraphe suivant.

3.2. Stationnarisation, modèle et indice de ciel clair

Une des grandes questions dans l'étude de séries temporelles est de savoir si celles-ci suivent un processus stationnaire. On entend par là le fait que la structure du processus sous-jacent évolue ou non avec le temps. Si la structure reste la même, le processus est dit alors stationnaire. Or il apparaît que l'évolution de l'éclairement solaire ne suit pas un processus stationnaire car elle présente une double périodicité (annuelle et journalière).

La plupart des modèles qui seront présentés dans ce document sont dits stationnaires et ne peuvent être appliqués (dans des conditions optimales) qu'à des séries stationnaires (Voyant *et al.*, 2018a). Il n'est donc pas possible de les mettre en œuvre directement sur les données à notre disposition ; il convient, au préalable, de les stationnariser afin de supprimer les périodicités connues. Cette étape est nécessaire car il est prouvé que l'utilisation de séries temporelles non stationnaires dans la modélisation donne des résultats erronés, la modélisation n'est alors pas valide (Phillips, 1986).

Dans ce paragraphe, nous présenterons, dans un premier temps, le modèle de rayonnement solaire par ciel clair qui sera utilisé pour définir les indices de ciel clair (second temps) et ainsi réaliser la stationnarisation (3ème étape).

3.2.1. Le modèle de ciel clair

Le modèle dit « ciel clair » consiste à calculer le rayonnement solaire qui parvient au sol en l'absence de couverture nuageuse. Il prend en compte l'atténuation du rayonnement solaire hors atmosphère lorsque celui-ci la traverse pour atteindre le sol. C'est donc le rayonnement maximal théorique qu'il est

possible de recevoir en un lieu donné. L'avantage de cette modélisation réside dans le fait qu'elle est réalisable en n'importe quel point du globe à condition de connaître l'état moyen de l'atmosphère. Dans la littérature, on constate que le développement de ce type de modèle remonte au début des années 1980.

Le premier modèle développé est celui de Kasten, il est basé sur le calcul de l'épaisseur optique (Kasten, 1980). Les différentes valeurs de rayonnement sont calculées en prenant en compte les absorptions et diffusions à deux altitudes différentes (2500 m et 8000 m) (Kasten, 1984). Le modèle a notamment besoin de la turbidité de Linke comme paramètre d'entrée.

A la même période, Bird (1980) mettent en équation les différents phénomènes d'atténuation de l'atmosphère en basant leur travail sur les modèles de transferts radiatifs. Le modèle a besoin de trois paramètres en entrée :

- La hauteur de la colonne d'eau (cm) ;
- La profondeur optique (calculée à partir de l'atténuation spectrale sur deux longueurs d'ondes : 380 et 500 nm) ;
- L'épaisseur d'ozone (cm).

En 1989, le modèle CPCR2 est développé par Gueymard (1989). Il est basé sur la transmittance de l'atmosphère sur deux bandes spectrales. Les rayonnements directs et diffus sont alors définis comme des fonctions des transmittances des bandes. Ce modèle sera amélioré par Gueymard lui-même (2004). Les fonctions de transmittance sont améliorées en ajoutant la distribution spectrale hors atmosphère ainsi que la valeur de la constante solaire.

Entre 2000 et 2002, dans le cadre du programme ESRA (European Solar Radiation Atlas), le modèle Esra est développé et utilisé dans Heliosat-2 (programme d'estimation du rayonnement par satellite) (Geiger *et al.*, 2002b; Rigollier *et al.*, 2000). Ce modèle est basé sur des travaux de paramétrisation de la profondeur optique de Rayleigh par Kasten (1996) et sur la turbidité de Linke.

Le modèle SOLIS, développé dans le cadre du programme Heliosat-3 par Mueller *et al.* (2004), est basé sur les calculs des modèles de transferts radiatifs en utilisant une loi d'atténuation de type de Beer-Lambert modifiée. Les composantes du rayonnement solaire sont obtenues par intégration sur le spectre solaire.

Les travaux de comparaison des modèles dans la littérature permettent de choisir de manière objective les modèles qui répondent le mieux aux besoins. Ineichen (2006) a comparé les modèles parmi les plus performants répertoriés. Cette étude, qui porte sur 16 sites de mesures avec des altitudes et des climats différents, conclut sur la supériorité du modèle SOLIS bien que très gourmand en ressources.

Le modèle SOLIS simplifié est alors développé par Ineichen (2008). Il parvient à diminuer la quantité de ressources utilisées pour le calcul. La détermination des paramètres qui sont nécessaires à l'utilisation des modèles de transferts radiatifs est réalisée en amont pour un large éventail d'altitudes ; ces paramètres sont alors utilisés en entrées du modèle pour déterminer analytiquement la meilleure combinaison.

Nous avons choisi d'utiliser le modèle SOLIS simplifié dans le cadre de notre étude. Décrivons le modèle plus en détail ainsi que tous les paramètres dont nous avons besoin pour l'utiliser.

On considère que l'éclairement solaire direct normal par ciel clair résulte de l'atténuation de l'éclairement hors atmosphère sous la forme d'une loi de Beer-Lambert :

$$\dot{I}_{b,\lambda}^{cs} = \dot{I}_0 . \exp(-M_\lambda . \tau) \qquad \text{II-8}$$

Il s'agit du rayonnement direct monochromatique reçu au niveau du sol sur une surface normale.

- I_0 est le rayonnement hors atmosphère ;
- M_λ la masse optique de l'air pour une longueur d'onde λ ;
- τ la profondeur optique.

Cette expression étant valide pour une radiation monochromatique, on considère que la profondeur optique est constante dans toute la masse d'air. Pour pallier cette approximation, il est nécessaire de modifier l'expression pour des intervalles de longueur d'onde, l'expression devient alors :

$$\dot{I}_b^{cs} = I_0 . \exp\left(-\frac{\tau}{sin^b\left(\frac{\pi}{2}-\theta_z\right)}\right) \qquad \text{II-9}$$

Dans cette expression, b est le paramètre d'ajustement obtenu en calculant les coefficients des modèles de transferts radiatifs entre deux hauteurs solaires différentes.

Lorsque l'on désire exprimer le rayonnement global horizontal par la relation de Beer-Lambert, cette dernière n'est plus valable et doit être modifiée ainsi :

$$\dot{I}_g^{cs} = I_o . \exp\left(-\frac{\tau}{sin^g\left(\frac{\pi}{2}-\theta_z\right)}\right) . \sin(\frac{\pi}{2}-\theta_z) \qquad \text{II-10}$$

Ici, g est un autre paramètre d'ajustement obtenu de la même manière que précédemment pour b (calcul des paramètres des modèles de transferts radiatifs). Cette expression est une bonne approximation en ce qui concerne la modélisation du rayonnement global (Mueller *et al.*, 2004).

Concernant la composante diffuse du rayonnement, une bonne approximation est définie de manière analogue aux deux expressions précédentes :

$$\dot{I}_d^{cs} = I_o . \exp\left(-\frac{\tau}{sin^d\left(\frac{\pi}{2}-\theta_z\right)}\right) \qquad \text{II-11}$$

De la même manière, le paramètre d est issu des calculs de modélisation des transferts radiatifs.

Ces expressions sont valables dans le cas d'une atmosphère pure, ou seules les caractéristiques des gaz qui la composent sont prises en compte. En réalité, ce n'est pas le cas, la présence d'aérosols et de particules en suspension dans l'atmosphère a des effets non négligeables sur le rayonnement qui parvient jusqu'au sol. Il faut alors établir une nouvelle façon d'exprimer l'éclairement qui traverse l'atmosphère en prenant en compte ces effets.

L'équation (II-12) donne l'expression de l'éclairement modifiée avec la prise en compte des aérosols :

$$\dot{I}_0' = I_0 . \left(0{,}12 . w^{0{,}56} . aod_{700}{}^2 + 0{,}97 . w^{0{,}032} . aod_{700} + 1{,}08 . w^{0{,}0051} . ln\left(\frac{p}{p_0}\right)\right) \qquad \text{II-12}$$

Dans cette expression, nous prenons en compte la hauteur de colonne d'eau *w*, aussi appelée hauteur d'eau condensable, exprimée en cm et la profondeur optique à 700 nm, aod_{700}. p est la pression à l'altitude considérée et p_0 la pression au niveau de la mer.

En remplaçant alors dans les formules du rayonnement et en prenant en compte les profondeurs optiques respectives, les expressions deviennent :

$$\dot{I}_b^{cs} = I_0' . \exp\left(-\frac{\tau_b}{sin^b\left(\frac{\pi}{2}-\theta_z\right)}\right) \qquad \text{II-13}$$

$$\dot{I}_g^{cs} = I'_0 . \exp\left(-\frac{\tau_g}{sin^g\left(\frac{\pi}{2}-\theta_z\right)}\right) . \sin(\frac{\pi}{2}-\theta_z) \quad \text{II-14}$$

$$\dot{I}_d^{cs} = I'_0 . \exp\left(-\frac{\tau_d}{sin^d\left(\frac{\pi}{2}-\theta_z\right)}\right) \quad \text{II-15}$$

Les profondeurs optiques, τ_b , τ_g et $\tau_d$respectivement pour le rayonnement direct, global et diffus, sont obtenues par le calcul, le lecteur intéressé pourra trouver les détails dans les travaux de Ineichen (2008).

Les paramètres de profondeur optique et de hauteur de colonne d'eau sont des caractéristiques de l'état de l'atmosphère, celles-ci varient beaucoup d'un mois sur l'autre et même au cours d'une journée. Une première approximation est de considérer une moyenne mensuelle de ces paramètres. Ces deux paramètres devraient, en théorie, être mesurés de façon précise pour chaque lieu, cependant ce n'est pas le cas. Une base de données existe, nous pouvons y trouver des mesures in situ ou à partir de plates-formes spatiales (https://aeronet.gsfc.nasa.gov/new_web/networks.html).

Le Tableau II-3 et le Tableau II-4 rassemblent les moyennes mensuelles des mesures de la profondeur optique et de la hauteur de colonne d'eau selon le site Aeronet.

Tableau II-3: Moyenne mensuelle des mesures de la profondeur optique à 700nm pour les différents sites de mesure

AOD 700	Janv	Févr	Mars	Avri	Mai	Juin	Juil	Aout	Sept	Octo	Nove	Déce
Ajaccio	0,051	0,069	0,073	0,110	0,151	0,112	0,164	0,132	0,310	0,423	0,082	0,033
Tilos	0,104	0,101	0,203	0,133	0,30	0,116	0,134	0,152	0,131	0,082	0,101	0,071
Odeillo	0,003	0,003	0,047	0,027	0,041	0,064	0,037	0,049	0,023	0,008	0,016	0,011
Nancy	0,040	0,091	0,294	0,170	0,057	0,089	0,076	0,050	0,058	0,120	0,373	0,039

Tableau II-4: Moyennes mensuelles des mesures de la hauteur de colonne d'eau pour les différents sites de mesure

Hauteur d'eau (cm)	Janv	Févr	Mars	Avri	Mai	Juin	Juilt	Aout	Sept	Octo	Nove	Déce
Ajaccio	1,081	0,940	1,221	1,719	1,768	2,471	2,750	2,931	2,657	2,409	1,670	1,00
Tilos	0,121	0,214	0,232	0,179	0,191	0,201	0,223	0,274	0,189	0,120	0,160	0,089
Odeillo	0,288	0,288	0,146	0,071	0,339	0,546	0,496	0,558	0,396	0,341	0,239	0,353
Nancy	0,781	0,612	0,730	1,201	1,190	1,939	2,081	2,012	1,510	1,603	1,179	0,663

Ces mesures permettront de paramétrer correctement le modèle de ciel clair qui nécessite également les paramètres suivants :

- Latitude
- Longitude

- Altitude
- Type d'aérosol par zone (rurale, maritime, urbaine, troposphérique)
- Azimut
- Angle d'inclinaison

Tableau II-5: Paramètres des différents sites de mesure, nécessaires pour paramétrer le modèle SOLIS

	Latitude (°)	Longitude (°)	Altitude (m)	Type de zone	Azimut (°)	Angle d'inclinaison (°)
Ajaccio	41,9123	8,6531	4	Maritime	0	0
Tilos	36,4330	27,3667	78	Maritime	0	0 et 30°
Odeillo	42,4935	2,0292	1650	Rural	0	0
Nancy	48,6586	6,1509	271	Urbain	0	0

La paramétrisation permet d'avoir une modélisation la plus fidèle possible. Il est nécessaire de réaliser précisément cette paramétrisation, en effet, le modèle ciel clair est à la base de la stationnarisation de nos données.

Les Figures II-9 a II-11 présentent quelques courbes d'éclairement solaire par journée avec ciel clair mesuré et modélisé pour chaque jeu de données.

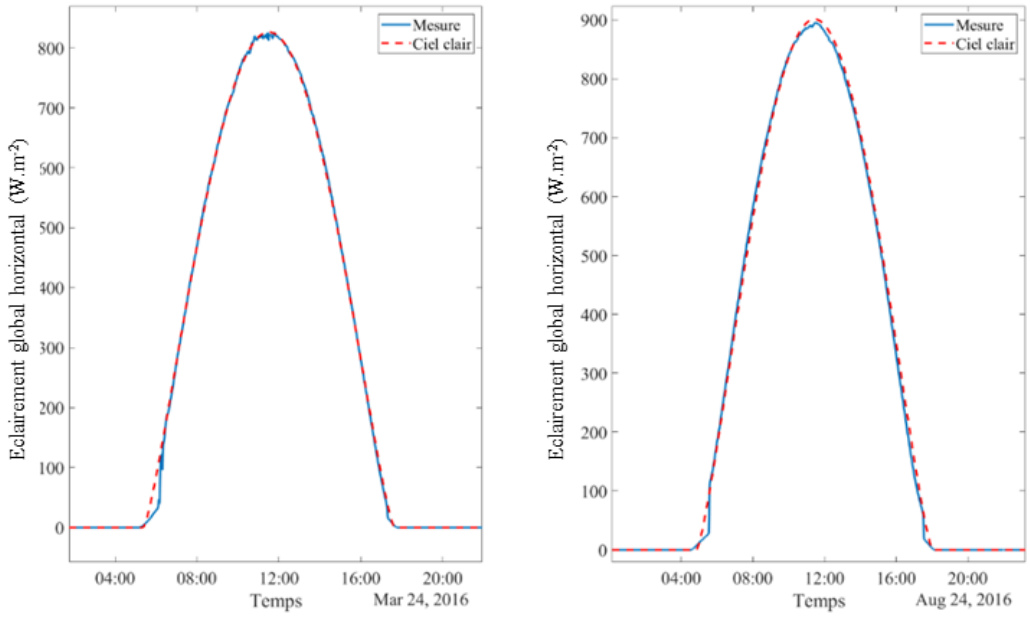


Figure II-8: Mesure d'éclairement global et modélisation ciel clair pour deux journées caractéristiques à Ajaccio

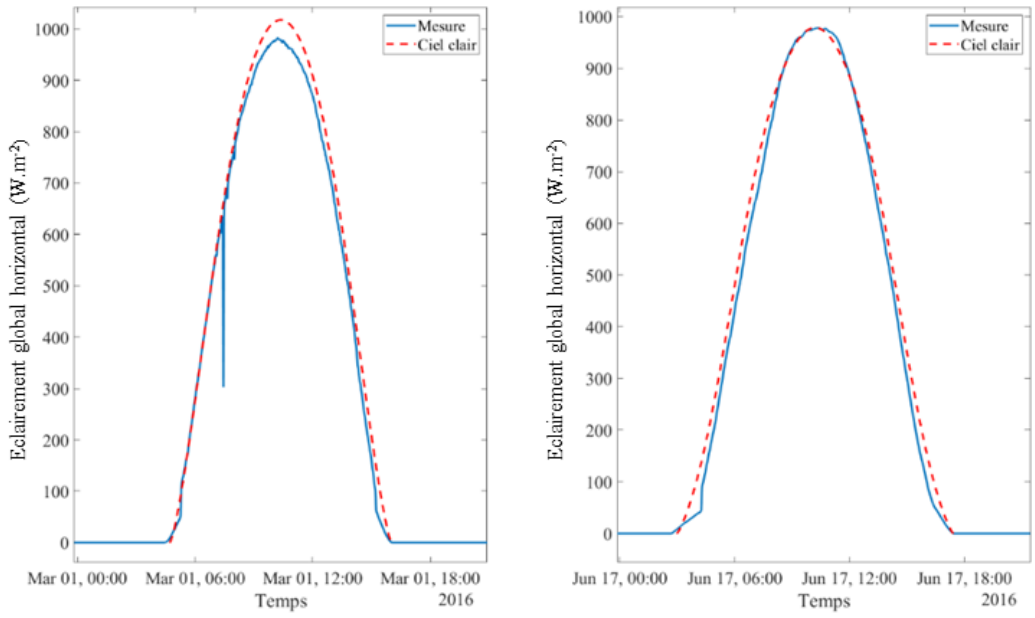


Figure II-9: Mesure d'éclairement global et modélisation ciel clair pour deux journées caractéristiques à Tilos

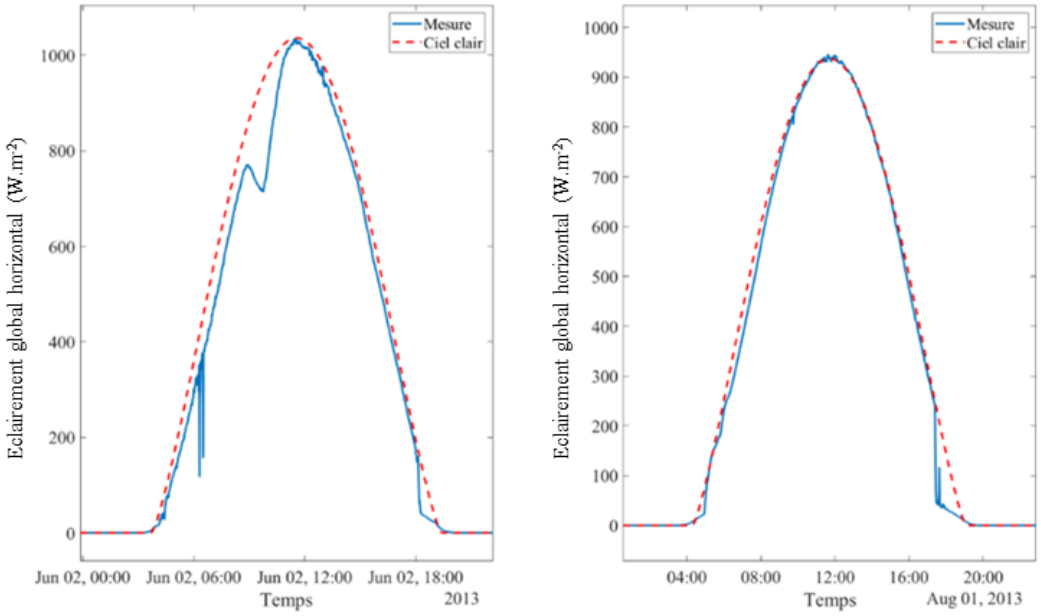


Figure II-10: Mesure d'éclairement global et modélisation ciel clair pour deux journées caractéristiques à Odeillo

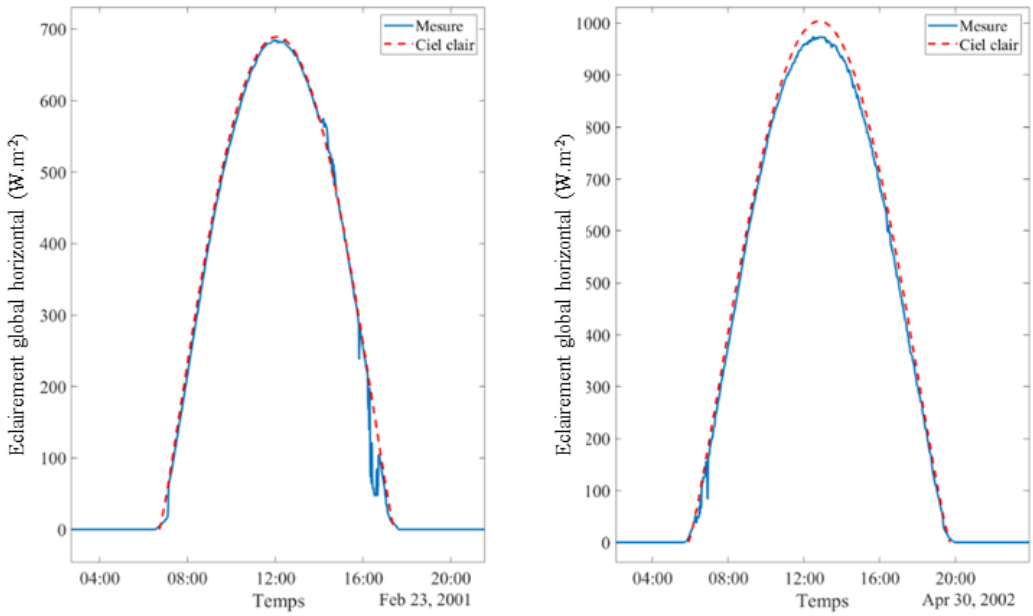


Figure II-11: Mesure d'éclairement global et modélisation ciel clair pour deux journées caractéristiques à Nancy

3.2.2. La stationnarité et l'indice de ciel clair

La stationnarité est une hypothèse importante, à la base de l'analyse des séries temporelles, de leur modélisation et donc de leur prévision (Bourbonnais et Terraza, 2008). Littéralement, un processus est dit « stationnaire » si sa structure ne varie pas au cours du temps, autrement dit si ses propriétés statistiques, comme l'espérance, sont indépendantes du temps. Mathématiquement, la stationnarité d'un processus est définie sous deux formes : la stationnarité forte et la stationnarité faible (Gladyshev, 1963; Pagano, 1976).

Soit une série temporelle, représentant un processus, notée $x_t : t \in [1, n]$:

- La stationnarité est dite « forte », si la fonction de répartition, notée P, qui caractérise la loi de probabilité de la série temporelle est constante pour tout intervalle « suffisamment grand », il vient :

$$P(x_1, \dots, x_n) = P(x_{1+h}, \dots, x_{n+h}) \qquad \text{II-16}$$

Cela signifie qu'un processus est stationnaire si ses propriétés ne sont pas affectées par un changement de « repère temporel » (translation dans le temps de h). Comme la loi de probabilité d'une distribution d'une série temporelle est particulièrement complexe à estimer, une définition moins stricte de la stationnarité a été introduite, il s'agit de la condition de stationnarité faible.

- La stationnarité faible, implique que les caractéristiques, aussi appelés « moments », dits d'ordre 1 et 2 sont indépendants du temps, il y a mathématiquement 3 conditions à respecter :

$$E(x_t) = \mu \qquad \forall\, t \in [1, n] \qquad \text{II-17}$$

$$Var(x_t) = \sigma^2 \neq \infty \qquad \forall\, t \in [1, n] \qquad \text{II-18}$$

$$Cov(x_t, x_{t+h}) = f(h) \qquad \forall\, t \in [1, n], \forall\, h \in [1, n] \qquad \text{II-19}$$

Cela signifie donc que l'espérance mathématique et la variance sont constantes au court du temps : il n'y a pas de tendance et la variabilité est stable au sein de la série. Le dernier point concerne la covariance (ici c'est l'auto covariance) et le fait qu'elle ne dépend que de l'ampleur du décalage h et non du moment où elle est calculée.

La stationnarité faible est aussi nommée stationnarité du second ordre, cette appellation résulte du fait que l'on se base exclusivement sur les deux premiers moments de la variable aléatoire de X_t. Dans la suite du document, les méthodes de stationnarisation utilisées sont supposées générer des séries stationnaires ou quasi-stationnaires pour lesquelles l'ensemble des conditions de stationnarité faible sont respectées (« presque » partout). Cette hypothèse nous permet d'utiliser la plupart des outils classiques de modélisation.

L'hypothèse stationnaire est facile à vérifier dans des simulations numériques avec des séries construites. Elle est plus complexe à réaliser dans le cas de processus liés à des signaux réels. C'est à ce niveau qu'intervient le modèle de rayonnement solaire par ciel clair qui va nous permettre de retirer la périodicité du phénomène que nous étudions.

L'éclairement solaire par ciel clair, calculé à partir de formules géométriques et de moyennes mensuelle de paramètres météorologiques, représente l'éclairement solaire en l'absence de nébulosité (ou plutôt avec une nébulosité moyenne). En divisant l'éclairement solaire mesuré au sol par l'éclairement solaire calculé par ciel clair, nous obtiendrons une variable appelée « indice de ciel clair » (rapport à la tendance) pour laquelle toute périodicité aura disparu, ce qui nous permet de justifier l'hypothèse de stationnarité des variables d'entrée de nos modèles.

L'indice de ciel clair est défini comme le rapport entre l'éclairement solaire mesuré au sol et le rayonnement modélisé par ciel clair. Pour chacune des 3 composantes (globale, directe et diffuse), un indice de ciel clair est défini par les équations (II-20) à (II-22) :

$$K_g(t) = \frac{i_g(t)}{i_g^{cs}(t)} \qquad \text{II-20}$$

$$K_b(t) = \frac{i_b(t)}{i_b^{cs}(t)} \qquad \text{II-21}$$

$$K_d(t) = \frac{i_d(t)}{i_d^{cs}(t)} \qquad \text{II-22}$$

Ces indices nous permettent de générer de nouvelles séries temporelles exploitables pour la plupart des modèles de prévision basés sur l'étude des séries temporelle et la modélisation de processus stochastiques ; ce sont elles qui sont prédites puis un post-traitement nous permet de revenir à la grandeur recherchée, l'éclairement ou l'irradiation.

Le rayonnement solaire comporte une tendance déterministe connue qui admet deux périodicités liées aux variations quotidiennes et annuelles, auxquelles s'ajoute une partie stochastique par nature inconnue et correspondant, notamment, à l'occurrence nuageuse et au bruit dans la série. Lorsque l'on cherche à réaliser une prévision c'est à cette composante que l'on s'intéresse. En effet, la partie déterministe est facilement calculable par le biais de la modélisation ciel clair. Le fait de stationnariser la série temporelle nous permet de ne pas prendre en compte cette partie déterministe. En général, une tendance saisonnière apparaît lorsqu'une série temporelle est influencée par des facteurs saisonniers, par

exemple, le mois de l'année, le jour de la semaine ou l'heure du jour (Hokoi *et al.* 1990). La saisonnalité est toujours d'une période fixe et connue (Franses et Paap, 1995).

De ces observations, nous déduisons la propriété suivante (Voyant *et al.*, 2018a) : La série temporelle observée d'éclairement solaire peut être considérée comme une série temporelle périodique avec deux périodes saisonnières fixes H (= 24 heures) et D (= 365 jours). La décomposition d'une série temporelle de rayonnement solaire met en évidence trois nouvelles séries temporelles : deux sont fortement saisonnières et la troisième est liée au bruit ou à la composante irrégulière. Soit une série temporelle de rayonnement notée $I(t)$, on a :

$$\{I(t), t \in Z\} = \{f\left(S_{24h}(t), S_{365d}(t), \varepsilon(t)\right), t \in Z\} \quad \text{II-23}$$

La fonction $f(\cdot)$ définit le type de décomposition : par modèle additif, multiplicatif ou hybride. Habituellement, le mode multiplicatif est préféré, et le terme $S_{24h}.S_{365d}$ à l'instant t correspond au rayonnement calculé par le modèle de ciel clair, c'est-à-dire :

$$I^{cs}(t) = \left(S_{24h}(t).S_{365d}(t)\right), \ t \in Z \quad \text{II-24}$$

Avec ce formalisme il vient :

$$\{I(t), t \in Z\} = \{S_{24h}(t).S_{365d}(t).\varepsilon(t), t \in Z\} = \{I^{cs}(t).\varepsilon(t), t \in Z\} \quad \text{II-25}$$

On comprend maintenant pourquoi l'indice de ciel clair (K(t)) correspond à une nouvelle série dans laquelle la périodicité est absente et qui est considérée comme stationnaire :

$$\{K(t), t \in Z\} = \{\varepsilon(t), t \in Z\} \quad \text{II-26}$$

La Figure II-12 est un exemple de série temporelle d'éclairement solaire global horizontal stationnarisée c'est-à-dire de l'indice de ciel clair. Dans le paragraphe 1, la Figure II-1, représentant la série de mesures, permet de voir distinctement les périodicités du rayonnement solaire, la **Erreur ! S ource du renvoi introuvable.** est issue du même jeu de données.

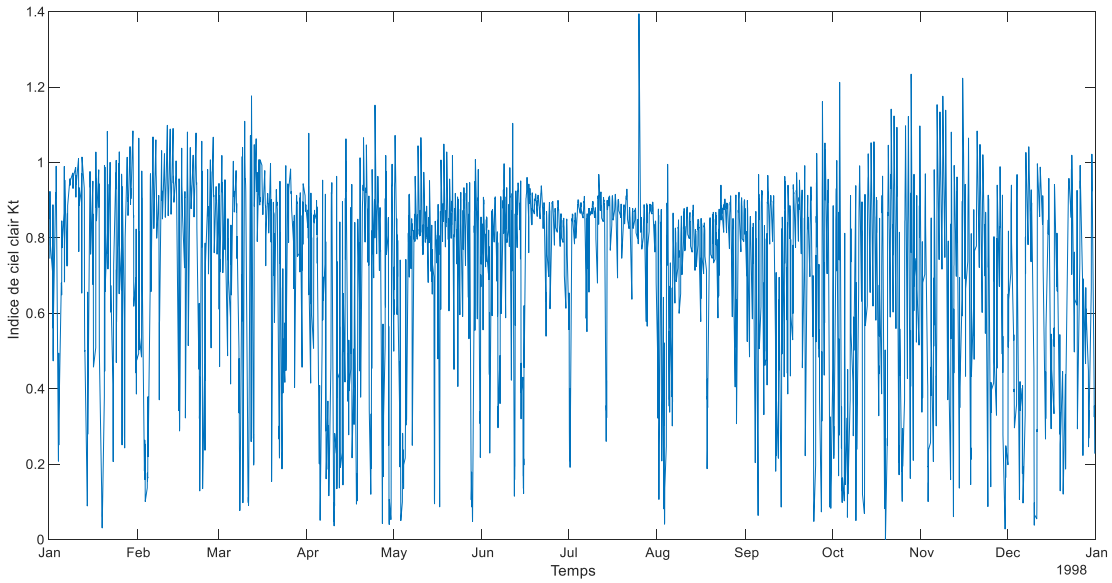


Figure II-12: Indice de ciel clair calculé pour le rayonnement global mesuré à Ajaccio

Le paragraphe suivant concerne les dernières étapes de prétraitement des données avant de commencer la modélisation, elle est composée de la filtration des données, de la détermination du nombre d'entrées et du partitionnement.

3.3. Filtration, sélection des entrées des modèles et partitionnement des données

Il s'agit ici d'expliquer, de manière détaillée, les dernières étapes du prétraitement des données, à l'issue de ces ultimes préparatifs, les données sont ensuite prêtes à alimenter les modèles pour réaliser la prévision.

3.3.1. La filtration

La filtration des données permet de s'affranchir de deux problèmes importants :

- D'une part les heures de nuit, durant lesquelles l'éclairement solaire est égal à 0, il est donc inutile d'alourdir les modélisations avec toutes ces données très facilement prévisibles
- D'autre part, les données mesurées au lever et au coucher du soleil ne sont pas toujours d'une grande fiabilité. Les rayons incidents sur le capteur arrivent avec un angle zénithal très faible, ce qui induit une erreur dans la mesure ; en effet, la réponse du capteur est faussée par tous les effets indésirables qui se produisent dans le dôme de mesure des pyranomètre, cette mesure est d'autant plus dégradée si on utilise une cellule étalon. On considère alors que la hauteur solaire seuil, au-delà de laquelle la mesure est considérée comme correcte est de 6° (angle solaire zénithal $\theta_z = 84°$) (Muneer et Fairooz, 2002). De plus, pour des hauteurs solaires très faibles, l'éclairement solaire peut être perturbé par l'environnement direct du lieu de la mesure : la présence de montagnes, d'arbres ou d'immeubles autour de la station de mesure peut générer des erreurs dans les données liées à l'ombre portée de ces obstacles.

Dans notre étude, nous avons décidé d'appliquer une filtration pour une limite de 10° de la hauteur solaire soit un angle zénithal solaire $\theta_z = 80°$.

L'utilisation de ce type de filtration ne doit pas faire oublier que même avec une faible hauteur de soleil, la production d'une centrale photovoltaïque de grande puissance peut être élevée.

3.3.2. Sélection du nombre d'entrées des modèles

Lorsque l'on utilise des modèles de machine learning nécessitant une phase d'apprentissage, il est nécessaire de choisir de manière optimale les données qui seront fournies au modèle pour qu'il soit correctement configuré. Il faut définir une méthode de sélection des variables d'entrée.

Pour réaliser cette sélection, nous utilisons l'information mutuelle basée sur la théorie de l'information de Shannon. Cette théorie basée sur les probabilités permet de quantifier le contenu en information d'un signal. C'est Claude Shannon qui pose les bases de cette théorie en 1948 à travers son article fondateur «A mathematical Theory of Communication» (Shannon, 1948). L'information mutuelle (IM) est une notion qui permet de quantifier l'information que partagent deux signaux.

La base mathématique de l'IM est basée sur l'entropie de Shannon. Par analogie elle peut être reliée à l'entropie thermodynamique quantifiant l'état de désordre d'un système.

Soit une variable x, pouvant prendre n valeurs discrètes, l'entropie S est définie par :

$$S(x) = -\sum_{i=1}^{n} P_i log_2(P_i) \qquad \text{II-27}$$

où $P_i = 1/N$ est la probabilité que x prenne la valeur x_i et N est le nombre de données, La présence du logarithme de base 2 permet d'exprimer l'entropie en bits. La formule (II-27) appliquée à deux variables *x* et *y*, pouvant prendre respectivement *n* et *m* valeurs, prend la forme de l'Equation (II-28) :

$$S(x,y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{i,j} log_2\left(P_{i,j}\right) \quad \text{II-28}$$

$P_{i,j}$ est la probabilité que *x* prenne la valeur x_i et *y* prenne la valeur y_j. Nous calculons ainsi l'IM entre *x* et *y* par l'Equation (II-29) (Bigdeli *et al.*, 2017):

$$IM(x,y) = S(x) + S(y) - S(x,y) \quad \text{II-29}$$

Il s'agit de la formulation pour le cas général. Dans notre cas, nous utilisons l'IM d'une façon différente, elle nous sert à déterminer le nombre de données historiques utiles pour réaliser la prévision. Cette sélection est nécessaire car trop peu de données ne permettraient pas d'appréhender les variations qui sont caractéristiques d'une série de mesures et trop de données seraient inutiles et compliqueraient la tâche des modèles notamment en termes de ressources et rendraient l'apprentissage hasardeux. On calcule alors l'information mutuelle de la série temporelle par rapport à elle-même, en effet, dans notre cas *x* et *y* sont de même nature puisque l'on réalise une prévision univariée.

L'astuce dans le cas de l'IM sur une seule variable est de calculer la dépendance statistique entre la variable x et elle-même mais décalée d'un instant i (Xuan *et al.*, 2019). On peut alors parler d'auto information mutuelle (AIM), la formulation devient :

$$AIM\big(x(t), x(t-i)\big) = S\big(x(t)\big) + S\big(x(t-i)\big) - S\big(x(t), x(t-i)\big) \quad \text{II-30}$$

On trace ensuite le résultat du calcul sous forme d'information en bits en fonction du nombre de lags (i) et on peut ainsi sélectionner le nombre de lags correspondant au nombre optimal de données d'historique qui correspond au premier minimum avant le rebond de l'AIM (Figure 2.13).

Si l'on remplace les grandeurs du cas général par l'indice de ciel clair (par exemple pour le rayonnement global que nous avons calculé dans l'équation (II-28), on obtient :

$$AIM\big(K_g(t), K_g(t-i)\big) = S\big(K_g(t)\big) + S\left(K_g(t-i)\right) - S\left(K_g(t), K_g(t-i)\right) \quad \text{II-31}$$

Et en remplaçant par les définitions de l'entropie, on obtient :

$$AIM\big(K_g(t), K_g(t-i)\big) = \sum\left(log\frac{P(K_g(t), K_g(t-i))}{P\big(K_g(t)\big).P(K_g(t-i))}\right).P\big(K_g(t), K_g(t-i)\big) \quad \text{II-32}$$

Cette formulation est bien entendu valable pour tous les indices de ciel clair, quelle que soit la composante.

La Figure II-13 est un exemple de représentation du calcul de l'auto-information mutuelle pour des données stationnarisées de rayonnement global horizontal mesurées à Ajaccio.

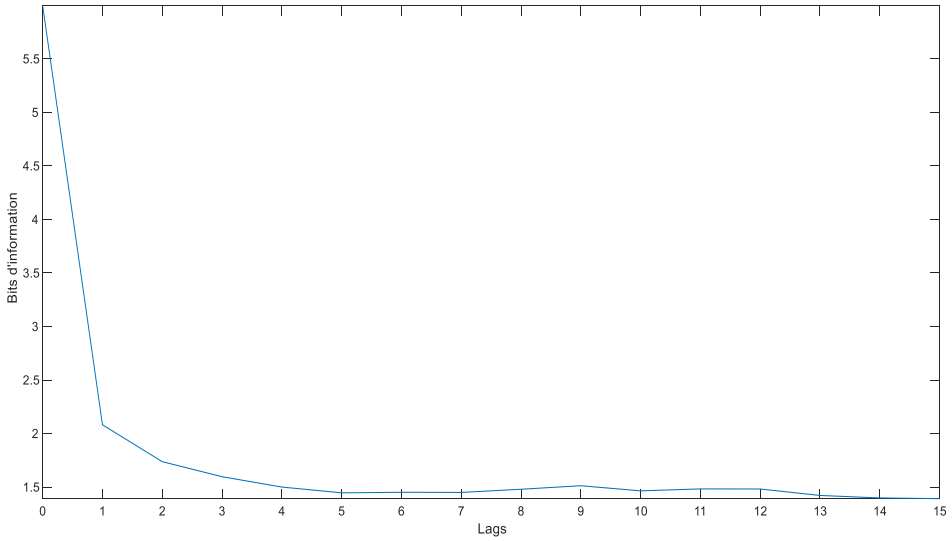


Figure II-13: Représentation graphique du calcul de l'AIM pour les données stationnarisées sur le site d'Ajaccio

La lecture graphique permet de déterminer le premier minimum qui correspond au nombre de lags optimal pour représenter cette série de données. Le fait de sélectionner le premier minimum, pour déterminer le nombre de données que nous allons utiliser, est effectué en accord avec le principe de parcimonie, il ne faut pas utiliser trop de données pour l'apprentissage au risque de phénomène de sur-apprentissage. Dans le cas de la Figure II-13 le nombre optimal de données d'entrées serait fixé à 5, cela signifie qu'il faudrait intégrer les 5 précédentes valeurs de K(t) pour effectuer la prévision de K(t+1). Ce nombre optimal de données d'entrée dépend du lieu de mesure, du pas de temps ou encore de la taille du jeu de données étudié ainsi que de l'horizon de prévision.

3.3.3. Partitionnement des données et validation croisée

Le partitionnement des données est une phase de préparation qui consiste à agencer les données pour alimenter les modèles qui ont besoin d'une phase d'apprentissage. Lorsque l'on cherche à paramétrer des modèles d'apprentissage automatique, il faut diviser les données en deux groupes :

- Le 1er groupe, qui contient la majeure partie des données, est consacrée à l'apprentissage, ces données serviront pour le paramétrage des modèles (détermination des poids pour un réseau de neurones, coefficients pour un modèle ARMA, architecture pour les arbres de régression…) ; ce set de données d'apprentissage est lui-même divisé en deux sous-groupes, un pour l'apprentissage proprement dit et un second pour la validation de cet apprentissage. Lorsque l'erreur commence à augmenter sur le set de validation, l'apprentissage est stoppé, on appelle cela le « early stopping ».
- Le 2nd groupe est appelé « set de test » et les données qu'il contient serviront à alimenter le modèle une fois paramétré, différents paramètres de niveau de fiabilité seront calculés pour juger la performance du modèle sur ce dernier set non utilisé lors de sa conception.

Les proportions de ce partitionnement peuvent être adaptées, traditionnellement au moins 60% des données sont utilisées pour la phase d'apprentissage et le reste pour le test. Dans notre étude, nous utilisons 80% des données pour le set d'apprentissage et 20% pour le test.

La Figure II-14 présente de manière schématique le partitionnement des données pour les modèles d'apprentissage automatique.

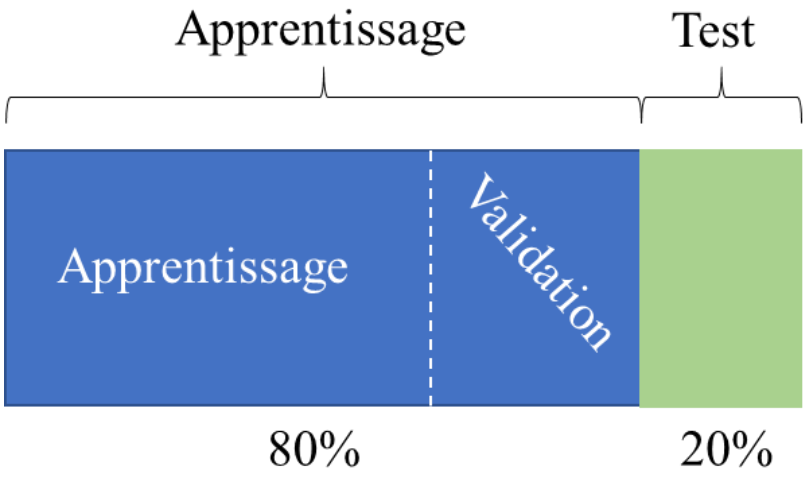


Figure II-14: Schéma du partitionnement des données pour l'apprentissage, la validation et le test des modèles

Lors de la phase d'apprentissage, le but est de modéliser les relations entre les données entrée-sortie. Le risque lors de cette phase est d'atteindre le moment où le modèle devient surentraîné, il risque alors de modéliser des évènements « accidentels » et il en résulterait une diminution de la qualité de la généralisation et donc des performances. Pour éviter ce phénomène de sur-apprentissage, il est obligatoire de poser une condition sur l'arrêt de l'apprentissage. De la même manière que pour le choix du nombre de données en entrée, le respect du principe de parcimonie conduit à stopper l'apprentissage le plus tôt possible. Suivant les modèles, les méthodes d'arrêt de l'apprentissage varient.

La Figure II-15 présente schématiquement l'évolution de l'erreur de prévision en fonction de la complexité du modèle, on entend ici par complexité non pas de son architecture propre mais plutôt son degré d'apprentissage.

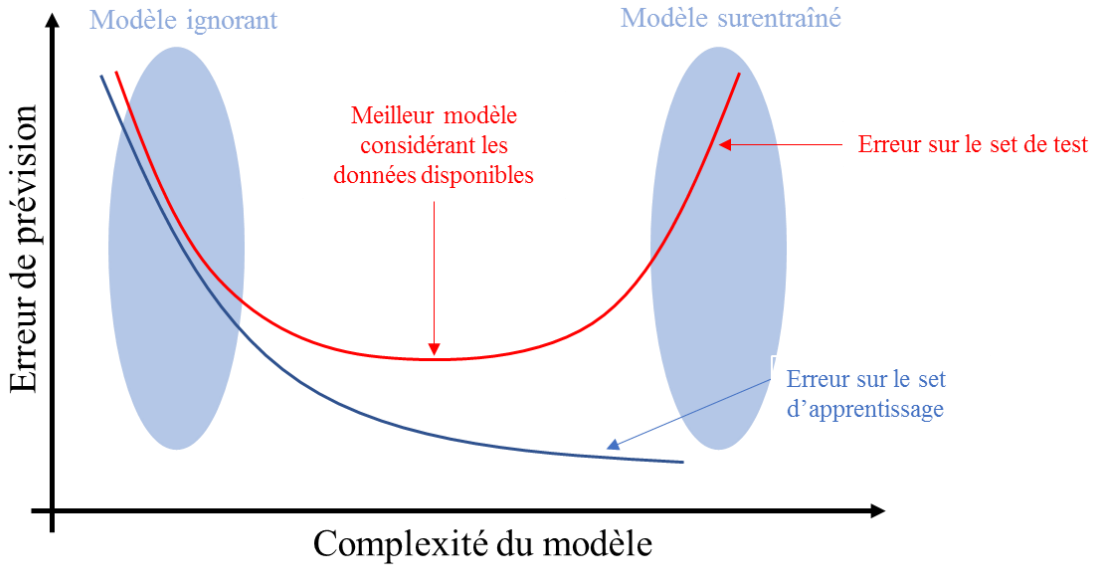


Figure II-15: Erreur de prévision en fonction de la complexité des modèles, effets du sur-apprentissage (Voyant *et al.*, 2017c)

La Figure II-15 montre bien l'importance de veiller au respect du principe de parcimonie durant la phase d'apprentissage. Durant toute la phase d'apprentissage, il faut se conformer, en permanence, au compromis biais-variance. Ce problème est complexe car il demande que l'on minimise simultanément deux sources d'erreur qui empêchent les algorithmes d'apprentissage d'effectuer une bonne généralisation en dehors des données du set d'apprentissage. Les deux erreurs à minimiser sont :

- Le biais, représente la déviation qui apparaît lorsque l'on réalise des hypothèses erronées dans l'algorithme d'apprentissage. Les valeurs élevées de biais font perdre aux algorithmes leur capacité d'établir des relations entre les valeurs d'entrées et de sortie.
- La variance, est l'erreur générée en appréhendant les petites fluctuations dans le set de données d'apprentissage. Il faut garder à l'esprit qu'une grande variance peut être la cause du sur-apprentissage qui débouche sur la modélisation du bruit aléatoire présent dans les données d'apprentissage, au lieu de la modélisation des variations attendues.

La résolution de ce problème de biais-variance est une alternative au problème de généralisation que l'on peut attendre d'un algorithme d'apprentissage pour un problème particulier. Cette erreur de généralisation peut se décomposer en trois termes, le biais, la variance et l'erreur irréductible qui est issue du bruit dans les données.

Les données doivent être classées sous forme matricielle, comme présentées sur la Figure II-16.

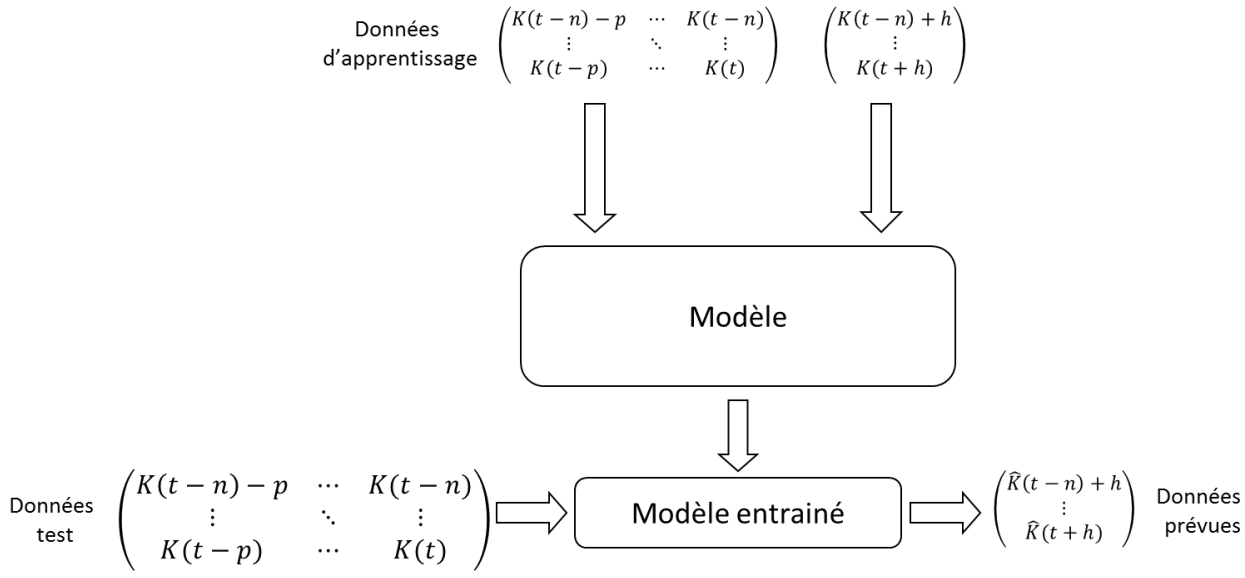


Figure II-16: Forme matricielle du partitionnement des données pour l'apprentissage (le symbole ^ indique qu'il s'agit d'une valeur prédite)

Sur cette Figure est présentée la forme générale de l'apprentissage et de la prévision avec :

- $K(t)$, l'indice de ciel clair à l'instant t ;
- h l'horizon de prévision ;
- n, la taille du jeu de données ;
- p, le nombre de données historiques obtenu par l'information mutuelle.

Dans cette phase d'apprentissage, on alimente les modèles avec les entrées et les sorties correspondantes afin que l'algorithme puisse ajuster les paramètres (problème classique d'optimisation). Cette opération est donc effectuée avec toutes les données qui servent à l'apprentissage.

Pour la phase de test, nous permettant de calculer l'erreur et d'estimer la précision, il suffit de recréer une matrice d'entrée, et de comparer les résultats de la matrice cible avec les mesures correspondantes.

Soit h l'horizon de prévision, en disposant des valeurs de la série temporelle de $x(t), t \in [1; n]$ on cherchera à prédire $x(t+h)$. Pour un horizon h, p étant le nombre de données passées et ε l'erreur entre la valeur prédite et la valeur mesurée aussi appelée résidu ou bruit, $\hat{x}(t+h)$ peut s'exprimer par :

$$\hat{x}(t+h) = f(x(t), x(t-1), x(t-2) \dots x\big(t-(p+1)\big) \qquad \text{II-33}$$

L'Equation (II-33) appliquée à la prévision de l'indice de ciel clair à l'horizon *h* devient :

$$\widehat{K}(t+h) = f(K(t), x(t-1), K(t-2) \dots K\big(t-(p+1)\big) \qquad \text{II-34}$$

On retrouve bien le terme prédit $\widehat{K}(t+h)$, en fonction d'un historique de données de même nature on parle ici du cas endogène. La fonction *f* représente le modèle que l'on choisit d'utiliser pour réaliser la prévision, ce modèle peut être linéaire (modèle ARMA), non linéaire mais continu (perceptron multi couche) ou non dérivable (arbres de régression).

Afin d'améliorer cette phase d'apprentissage et de valider les paramètres initiaux des modèles nous ajoutons une étape de validation croisée dite « méthode des k-fold ». Cette méthode permet de contrôler la validité des apprentissages par un échantillonnage « aléatoire » des données. On divise le jeu de données en *k* échantillons puis on se sert d'un de ces échantillons comme jeu de validation et les *k-1* autres échantillons comme jeu d'apprentissage. On répète l'opération *k* fois pour que tous les échantillons aient servi une fois comme échantillon de validation. On calcule pour finir la moyenne des *k* erreurs pour estimer l'erreur de prévision. Cette méthode permet d'effectuer une validation croisée objective des modèles moins sensibles aux particularités que le rayonnement global peut avoir sur certaines périodes : faible ou forte variabilité non réellement représentative du phénomène global étudié.

4. La variabilité des données

Ce paragraphe est consacré à la caractérisation mathématique des conditions météorologiques des données. Afin d'estimer, de manière quantitative, le degré de variation des données de rayonnement solaire, nous allons calculer un coefficient de variabilité pour chacun des sites météorologiques étudiés. Nous estimerons ainsi l'impact de cette variabilité sur les performances de nos prévisions.

4.1. Qu'est-ce que la variabilité et pourquoi la calculer ?

Que ce soit en économétrie ou en énergétique, dès lors que l'on utilise le formalisme des séries temporelles, la valeur future d'une grandeur physique (rayonnement) ou d'un produit financier, la provision peut être traitée de façon similaire. Dans le domaine de l'économétrie, de nombreuses études ont été réalisées sur l'impact des propriétés des données sur leur prévisibilité. Il est intéressant de pouvoir étendre ce type d'analyse à la prévision du rayonnement solaire.

La variabilité peut être définie comme la variation des données au cours du temps. Il est possible de faire l'analogie avec la volatilité en économétrie. En fait, il s'agit de quantifier de manière objective les variations que subissent les données au cours du temps. Dans le cas du rayonnement solaire, cette propriété est directement reliée aux conditions climatiques des lieux de mesure. Une fois la variabilité

déterminée, il devient possible de relier les conditions climatiques à la complexité de la prévision et répondre ainsi à la question : peut-on, à partir de la connaissance de la variabilité des données, s'orienter vers une méthode de prévision plutôt qu'une autre ?

Nous allons maintenant présenter la méthode de calcul de variabilité et le paramètre statistique que nous avons choisi pour notre étude.

4.2. Méthode de calcul de la variabilité

Les différents paramètres statistiques sont issus de l'économétrie et ont été testés sur les données de rayonnent solaire par Voyant *et al.* (2015). Il ressort que le paramètre statistique le plus adapté à nos besoins est le « mean absolute log return » défini par l'équation II-35 :

$$mean(abs\ logr) = E\left[\left|\ln\left(K(t)\right) - \ln\left(K(t-1)\right)\right|\right] \quad \text{II-35}$$

Avec *E* l'espérance mathématique et *K* l'indice de ciel clair pour la composante considérée. Les résultats obtenus pour chaque jeu de données sont présentés sur la Figure II-17.

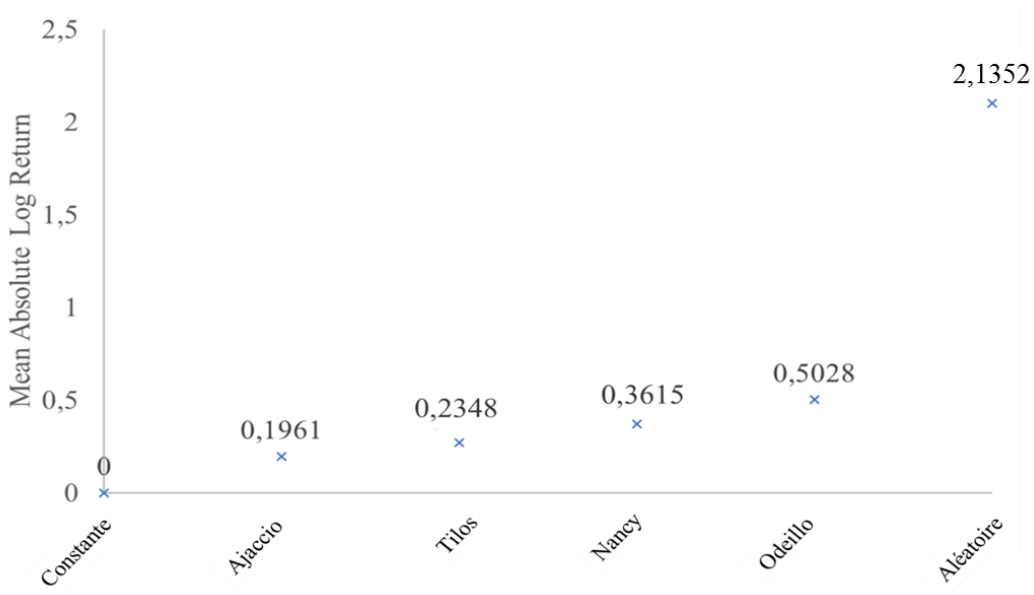


Figure II-17: Mean Absolute Log Return pour les différents lieux de mesure

Il a semblé judicieux de donner un ordre de grandeur possible de cette variabilité ; pour ce faire, nous avons calculé respectivement la variabilité au sens du MALR pour une série constante et une série de termes choisis totalement aléatoirement. Nous disposons ainsi d'un encadrement. A Ajaccio, les données d'éclairement solaire sont les moins volatiles, suivies par Tilos, puis Nancy et enfin Odeillo.

En plus de ce calcul de variabilité nous avons proposé une estimation de celle-ci sous forme de pourcentage, directement lié au calcul précédent en utilisant les bornes comme limites. La formule est la suivante :

$$var = 100.\frac{MALR}{MALR_{max} - MALR_{min}} \quad \text{II-36}$$

Ce qui nous a permis de classer les jeux de données dans le Tableau II-6 :

Tableau II-6: Résultats des calculs de la variabilité pour les différents jeux de données

Série de données	*Constante* $MALR_{min}$	*Ajaccio*	*Tilos*	*Nancy*	*Odeillo*	*Aléatoire* $MALR_{max}$
MALR (s.u.)	0	0,1961	0,2348	0,3615	0,5028	2,1352
var	0%	9,2%	11,0%	16,9%	23,5%	100%

La comparaison des performances de modèles de prévision et le classement des méthodes pourront désormais prendre en compte la variabilité du site. Pour évaluer la performance des modèles, nous avons à notre disposition plusieurs outils, mathématiques ou graphiques, qui seront détaillés dans la partie suivante.

5. Evaluation des modèles

Le but de l'évaluation des modèles est de mesurer leur capacité à effectuer une bonne prévision. Il est important de réaliser cette évaluation à l'aide de critères reconnus pour que cette évaluation soit objective. Cette évaluation est d'autant plus complexe que la définition d'une « bonne » prévision n'existe pas.

Dans le cas de la modélisation par les modèles d'apprentissage automatique, dont les relations entre les entrées et les sorties sont très complexes à interpréter, nous ne pouvons pas utiliser une technique de propagation des incertitudes. L'évaluation des modèles se fait alors, point par point, en comparant le résultat de la prévision avec la variable mesurée correspondante.

Il est très difficile d'évaluer les performances des modèles entre eux et surtout de les comparer aux résultats de la littérature. Effectivement, les jeux de données sont très différents, les périodes et les pas de temps de mesures également et pour finir il existe une multitude d'indices d'erreur qui parfois même, portant le même nom, ont des définitions distinctes.

Nous avons sélectionné les indices d'évaluation les plus utilisés dans la littérature et qui commencent à faire consensus (*COST action ES1002 Weather intelligence for Renewable Energies (WIRE)*, 2012). Rappelons que le calcul de cette erreur est réalisé sur des données qui n'ont pas servi durant l'apprentissage.

Les indices d'erreur servent à quantifier les différences entre les valeurs issues de l'observation (mesures) et les valeurs obtenues lors de la prévision. De manière concrète, on s'attache à comparer deux séries temporelles entre elles, celle constituée par les mesures $x(t)$ et celle constituée des prévisions $\hat{x}(t)$ correspondantes. Etant donné qu'il existe une erreur commise dans la mesure (appareillage, acquisition…) nous n'évaluons pas exactement la différence entre prévision et réalité stricte, la différence est donc évaluée entre la prévision et l'observation de la réalité. L'erreur systématique ou biais est mesurée par certains indice tandis que d'autres appréhendent aussi l'erreur aléatoire (bruit).

La Mean Bias Error (MBE) quantifie le biais moyen des prévisions :

$$MBE = \frac{1}{n} \sum_{t=1}^{n} (\hat{x}(t) - x(t)) \qquad \text{II-37}$$

Avec :

- $x(t)$ la variable mesurée ;
- $\hat{x}(t)$ la variable prévue ;

- t l'indice temporel ;
- n le nombre total de données.

La MBE n'est pas un outil pour mesurer la précision d'un modèle mais uniquement pour estimer la sous-estimation ou à la surestimation.

La Mean Absolute Error (MAE) est la valeur absolue (norme 1) de l'écart moyen entre les mesures et les prévisions, elle combine l'erreur systématique et l'erreur aléatoire. Elle permet de juger la précision d'un modèle, son unité est la même que la grandeur mesurée :

$$MAE = \frac{1}{n} \sum_{t=1}^{n} |\hat{x}(t) - x(t)| \qquad \text{II-38}$$

Sa forme normalisée est obtenue en divisant la MAE par la valeur moyenne des mesures, on obtient une erreur en pourcentage :

$$nMAE = \frac{\frac{1}{n} \sum_{t=1}^{n} |\hat{x}(t) - x(t)|}{\bar{x}} . 100 \qquad \text{II-39}$$

Ou $\bar{x}$ est la moyenne des valeurs mesurées.

La Root Mean Square Error (RMSE) est la racine carrée de la Mean Square Error (MSE), calculée à partir du carré de l'écart entre les mesures et les prévisions (norme 2). Elle a donc tendance à pénaliser les grands écarts. La MSE est utilisée lors de la phase d'apprentissage des modèles durant laquelle on cherche à la minimiser. Pour une évaluation de la précision, elle est peu intéressante ; son unité n'étant pas la même que celle de la variable étudiée. C'est pour cette raison que l'on utilise plutôt la RMSE qui a la même dimension que les variables étudiées :

$$MSE = \frac{1}{n} \sum_{t=1}^{n} (\hat{x}(t) - x(t))^2 \qquad \text{II-40}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left(\hat{x}(t) - x(t)\right)^2} \qquad \text{II-41}$$

Dans le domaine statistique, la RMSE et sa version normalisée sont largement utilisées car elles permettent une bonne lisibilité ainsi qu'une bonne évaluation de la précision des modèles. Les auteurs ne donnent pas toujours la définition de la RMSE, certains auteurs divisent la RMSE par l'écart maximal entre les données, ou même le maximum des données. Dans notre cas, nous avons choisi la version « classique » de la normalisation pour laquelle on divise la RMSE par la moyenne des donnés :

$$nRMSE = \frac{\sqrt{\frac{1}{n} \sum_{t=1}^{n} \left(\hat{x}(t) - x(t)\right)^2}}{\bar{x}} . 100 = \frac{RMSE}{\bar{x}} . 100 \qquad \text{II-42}$$

Le Skill Score ou encore Forecast Score (FS) est un indice qui est de plus en plus utilisé, il permet de comparer les performances du modèle par rapport à un modèle de référence bien connu. Il est intéressant en phase de développement et permet d'identifier très rapidement si un modèle est correctement paramétré ou non. En phase de résultats, ce type de paramètre permet de déterminer si un modèle très complexe a de bien meilleures performance qu'un simple modèle naïf.

$$FS = 1 - \frac{MSE_{modèle}}{MSE_{référence}} \qquad \text{II-43}$$

Nous avons choisi de nous limiter à ces indices car ils sont assez utilisés dans le domaine de la prévision et permettent d'avoir une idée assez juste des performances des modèles.

6. Synthèse

Dans ce chapitre, nous avons présenté le formalisme des séries temporelles, particulièrement important car à la base des simulations qui seront réalisées par la suite. Nous avons ensuite détaillé l'ensemble des données disponibles ainsi que les caractéristiques météorologiques des sites de mesures. Nous avons pu constater qu'il était compliqué de disposer de mesures fiables de rayonnement solaire car les appareillages de mesures doivent faire l'objet d'une attention particulière et d'un suivi sérieux.

Nous avons ensuite énuméré les phases de préparation des données (ou prétraitement) avec :

- Leur contrôle qualité ;
- La gestion des données manquante ;
- La gestion des données aberrantes.

Nous avons montré la nécessité de stationnariser les données et de définir un indice de clarté basé sur un modèle de rayonnement par ciel clair. La méthode nous permettant d'obtenir la dimension des matrices qui serviront à l'apprentissage des modèles a été décrite à partir de la théorie de l'information et plus particulièrement l'auto-information mutuelle. Nous avons également détaillé la manière de partitionner les données en trois échantillons : apprentissage, validation et test. Nous avons réalisé durant cette étape une validation croisée par la méthode des k-fold, qui permet d'améliorer la phase de développement des modèles.

III. Méthodologie et modèles de prévision

1. Introduction

Ce chapitre présente les différentes méthodes qui sont utilisées dans le cadre de la prévision du rayonnement solaire. Nous détaillerons les modèles utilisés pour réaliser cette étude. En préambule, il est important de définir une notion intervenant à de nombreuses reprises dans la suite du manuscrit : les modèles et la modélisation.

Un modèle est une série d'équations ou de représentations graphiques qui décrivent des relations entre des variables de la manière la plus précise possible. Des modèles mathématiques sont utilisés dans une multitude de domaines, en biologie, en ingénierie électrique, en sociologie, ou encore en économie.

En physique, l'utilisation des modèles (qui conduit à la modélisation) est éprouvée, elle trouve tout son sens dans la représentation de systèmes physiques qui mettent en jeu de grandes quantités de données. L'expansion de l'utilisation des modèles mathématiques s'est réellement accrue parallèlement au développement de l'informatique. Les premiers modèles permettaient de simuler en laboratoire des phénomènes dont l'expérimentation sur le terrain semblait impossible, en raison de leur échelle temporelle ou spatiale (mouvement des glaciers, phénomènes météorologiques exceptionnels).

Le travail du « modélisateur » est de confronter les résultats de son modèle avec la réalité pour le confirmer ou l'infirmer. Grace à la modélisation il est possible de réaliser une multitude de simulations et de scénarios. Par la suite, les scientifiques se sont naturellement tournés vers une discipline qui jusque-là était inconnue : la prévision. Cette nouvelle thématique ouvre un très grand nombre d'applications dans un large éventail de domaines.

Il existe plusieurs façons de réaliser la prévision de la ressource solaire. Le choix de l'une ou l'autre méthode dépend de l'horizon de prévision et des besoins de l'utilisateur. En effet, tous les modèles n'ont pas la même précision en fonction du type de prévision souhaité, du lieu, de l'horizon temporel ou encore du type voire du nombre de données disponibles. Les 3 grandes familles de modèles prédictifs sont réparties en fonction des horizons de prévision :

- Le très court terme, dont l'horizon est compris entre quelques minutes et une heure : on utilise alors des modèles basés sur l'imagerie du ciel (en anglais «sky imaging») (Kurtz *et al.*, 2017), ou encore des modèles basés sur le formalisme des séries temporelles et la persistance de la nébulosité. D'autres modèles basés sur les images satellites peuvent aussi être utilisés dans ces horizons de prévision (Lara-Fanego *et al.*, 2012).

- Le court terme (aussi appelé «nowcasting»), dont l'horizon de prévision s'étend de 1 heure à 6 heures: on utilise alors le formalisme des séries temporelles avec des méthodes d'apprentissage automatique ou bien des images satellites (Lorenz *et al.*, 2012; Paulescu *et al.*, 2013).

- Le moyen terme, de 1 jour à 1 semaine : typiquement le domaine où les modèles numériques, basés sur la résolution des équations de l'atmosphère, appelés NWP pour «Numerical Weather Prediction», sont les plus pertinents (Heinemann *et al.*, 2006b; Perez *et al.*, 2010a). Ces modèles de prévision sont parfois associés à des modules de post-traitement et sont complétés avec des informations issues des satellites (Lara-Fanego *et al.*, 2012). Ils demandent des ressources informatiques relativement importantes.

La Figure III-1 est constituée de deux parties, en haut les différentes techniques de prévision en fonction de la résolution temporelle et de l'horizon recherché, et en bas quelques détails sur les techniques de prévision en incluant les objectifs.

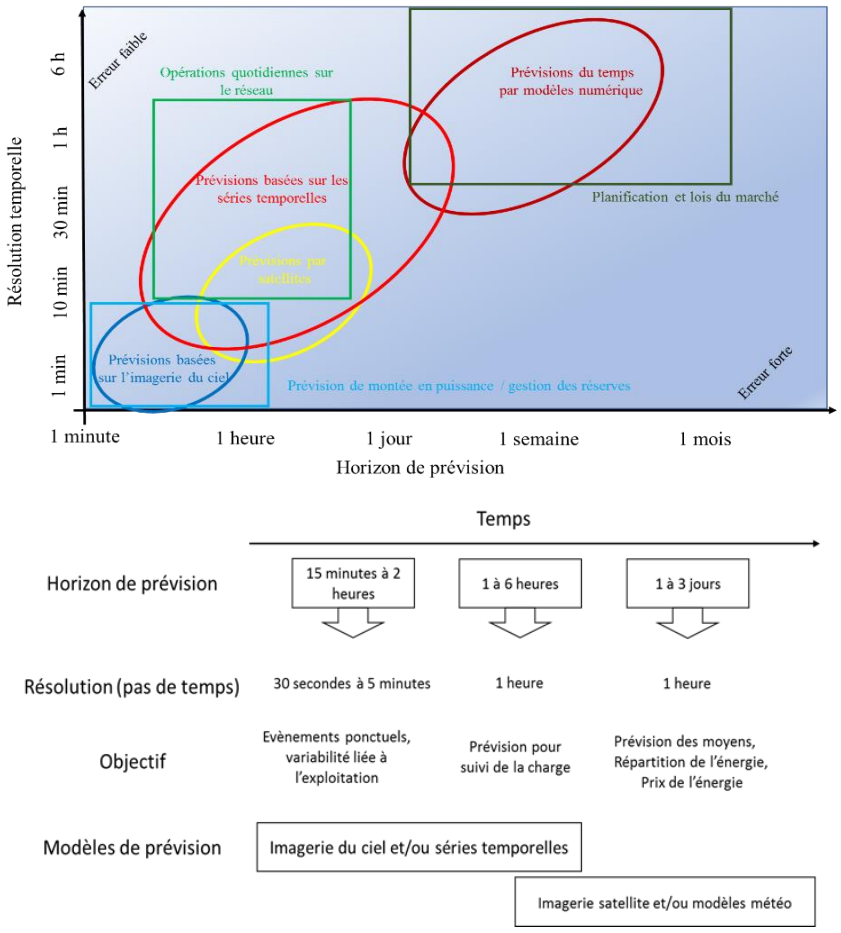


Figure III-1: En haut la résolution temporelle et horizon pour les différentes méthodes de prévision. En bas les relations entre l'horizon de prévision, les différents modèles et les objectifs (Diagne *et al.*, 2013)

Les modèles NWP permettent de générer une probabilité d'occurrence nuageuse qui sert d'entrée à des modèles dynamiques de l'atmosphère afin de déterminer l'éclairement solaire au niveau du sol. Les modèles d'extrapolation ou statistiques analysent des séries temporelles de rayonnement global, à partir de la télédétection par satellite (Lorenz *et al.*, 2004) ou de mesures au sol (Reikard, 2009) afin d'estimer le mouvement des nuages et projeter leur impact dans le futur (Diagne *et al.*, 2012; Perez *et al.*, 2010a, 2010b).

Des approches des méthodes de prévision, avec leurs limites et leur précision, peuvent être trouvées dans la littérature (COST, 2012; Diagne *et al.*, 2012; Elliston et MacGill, 2010; Espinar *et al.*, 2010; Heinemann *et al.*, 2006a; Paulescu *et al.*, 2013).

Des études comparatives de multiples approches ont été effectuées pour évaluer la précision des prévisions d'irradiation solaire (Lorenz *et al.*, 2009; Mihalakakou *et al.*, 2000; Perez *et al.*, 2010a, 2010b, 2007; Remund *et al.*, 2008). De plus, les métriques utilisées pour estimer l'erreur sont souvent différentes, rendant ainsi plus complexe encore la comparaison des modèles d'entre eux ; certains paramètres tels que le coefficient de corrélation, l'erreur quadratique moyenne sont souvent utilisés, mais pas toujours adaptés pour comparer les performances des modèles; la période utilisée pour évaluer la précision varie considérablement d'un article à l'autre: certains d'entre eux estiment la précision du modèle sur une période d'une ou plusieurs années, d'autres sur une période de quelques semaines introduisant alors un biais saisonnier potentiel.

Dans ces conditions, il n'est pas facile de faire des comparaisons des résultats présentés, les différentes études bibliographiques doivent donc être soigneusement analysées.

La partie suivante est consacrée à la présentation des différentes méthodes de prévision de l'éclairement solaire, de plus, nous détaillerons l'état de l'art réalisé dans le but d'orienter nos travaux.

2. La prévision de l'éclairement solaire

Nous allons dans cette section présenter les différentes manières de réaliser la prévision de l'éclairement solaire. Nous entrerons ensuite dans le détail des techniques qui nous intéressent.

2.1. Techniques basées sur l'imagerie du ciel et les données satellitaires

Les techniques basées sur l'imagerie du ciel et les données satellitaires utilisent soit des images provenant de dispositifs d'imagerie du ciel (caméras pointées vers le ciel), soit des images provenant des satellites. A partir de ces « photographies », ces techniques peuvent prévoir le mouvement des nuages dans un futur proche, et à partir de ces informations couplées à l'utilisation d'algorithmes spécifiques, prévoir le rayonnement solaire.

2.1.1. Imagerie du ciel ou « Sky Imaging »

Cette catégorie d'approches de prévision solaire utilise des images du ciel, réalisées à partir d'acquisitions faites au sol (objectifs grand angle), pour prévoir le mouvement des nuages, et les effets sur le rayonnement solaire pour un site géographique particulier. L'imagerie totale du ciel peut être utilisée pour effectuer des prévisions en temps réel (immédiat) jusqu'à 15-30 minutes, en appliquant des techniques de traitement d'images et de suivi des nuages à partir de photographies du ciel.

Sur la Figure III-2 à gauche, on voit une lentille d'un dispositif « sky imaging » vue de haut et à droite, l'exemple d'une photographie du ciel obtenue avec les vecteurs de mouvement des nuages.

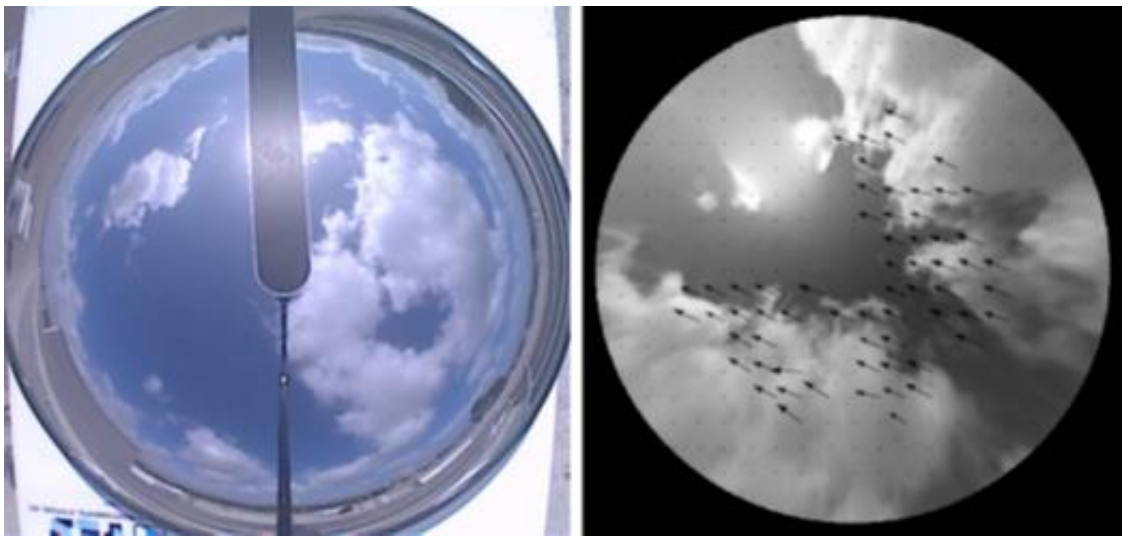


Figure III-2: A gauche, lentille de dispositif de "sky imaging" vue d'en haut. A droite, exemple de photographie du ciel avec les vecteurs de mouvements des nuages.

Le Total Sky Imager (TSI) a une application accrue ces dernières années et est représenté sur la Figure III-3.

Ce type de dispositif et les méthodes associées sont utilisés pour la prévision du rayonnement à très court terme (Rudd, 2011), leurs caractéristiques rendent leur utilisation répandue dans les dispositifs de production d'énergie solaire et de conversion (solaire thermique par exemple).

Figure III-3: Dispositif TSI, composé d'un miroir hémisphérique convexe avec une caméra au-dessus. Le miroir contient une bande d'ombre de suivi du soleil pour protéger la caméra des reflets du soleil.

2.1.2. Modèles physiques et statistiques basés sur les données satellites

Les mesures de rayonnement solaire par satellite constituent également une alternative à un réseau de mesures et d'acquisitions au sol, à l'échelle nationale ou mondiale. Selon le traitement de l'interaction du rayonnement solaire et de l'atmosphère, les modèles satellitaires peuvent être classés en deux catégories : les modèles physiques et les modèles statistiques (Noia *et al.*, 1993a, 1993b). Les avantages des modèles physiques sont leur généralisation due à l'utilisation de modèles de transfert radiatif (MTR) et le fait que les mesures in situ de données au sol ne sont pas nécessaires. Cependant, les modèles de transferts radiatifs nécessitent une mesure précise et complète de la morphologie atmosphérique ainsi qu'un étalonnage soigneux des appareils de mesures satellitaires.

Les modèles satellitaires statistiques s'appuient sur des régressions statistiques simples entre les mesures réalisées par satellite et au sol. En conséquence, ces modèles statistiques sont beaucoup plus simples en raison de leur indépendance par rapport à la mesure précise de la composition de l'atmosphère, mais ils souffrent de leur perte d'universalité et de la nécessité de disposer de données mesurées au sol.

En France les algorithmes Heliosat pour l'estimation de la ressource solaire ont été développés dans les années 80 (Cano *et al.*, 1986), la méthode Héliosat 2, qui est encore utilisée, trouve son origine dans les travaux de Rigollier *et al.* (2004) et elle sert aujourd'hui à alimenter les bases de données HelioClim. Ces bases de données diffèrent entre elles par les images satellitaires utilisées en entrée et par la résolution temporelle des observations (quart-horaire, horaire, journalière). Le nombre d'accès à ces bases de données (2 millions en 2011) démontre l'intérêt de HelioClim pour la science et l'industrie. D'un point de vue plus global, le modèle d'estimation du rayonnement solaire à partir d'images satellites le plus utilisé est celui de Perez, développé en 2002 (Perez *et al.*, 2002).

2.1.3. Les modèles numériques de prévision météorologiques NWP (Numerical Weather Prediction)

La méthode numérique de prévision météorologique (NWP) utilise un ou plusieurs des modèles météorologiques existants pour prévoir l'éclairement solaire. Cette méthode est généralement considérée comme la plus précise pour des horizons de 6 heures à plusieurs jours (4 jours pour ARPEGE de Météo France, 10 jours pour le modèles européen ECMWF, 16 jours pour le modèle américain GFS).

Un modèle NWP est un programme informatique qui simule le mouvement atmosphérique dans l'espace et dans le temps des nuages et particules. Une variété de phénomènes météorologiques peut être analysée par ces types de modèles de prévision numérique. Dans ce type de modèle, l'atmosphère est représentée par une grille 3D. Plus l'espacement de la grille est fin, plus la simulation est élaborée et complexe, mais le temps d'obtention des résultats est plus long, ce qui peut être un frein.

Il convient donc de faire un choix judicieux entre taille du maillage, complexité des modèles et durée d'exécution du programme. La simulation, faite avec ce type de modèle génère l'état futur de l'atmosphère dans chaque voxel (pixel en 3D) à partir de son état initial (Radnoti *et al.*, 1995). D'abord un domaine est défini, il est ensuite spatialement discrétisé suivant la résolution désirée. Enfin, le NWP prévoit l'ensemble des variables météorologiques usuelles en résolvant les équations du mouvement et les lois de la thermodynamique.

La Figure III-4 présente les étapes de mise en œuvre des modèles NWP.

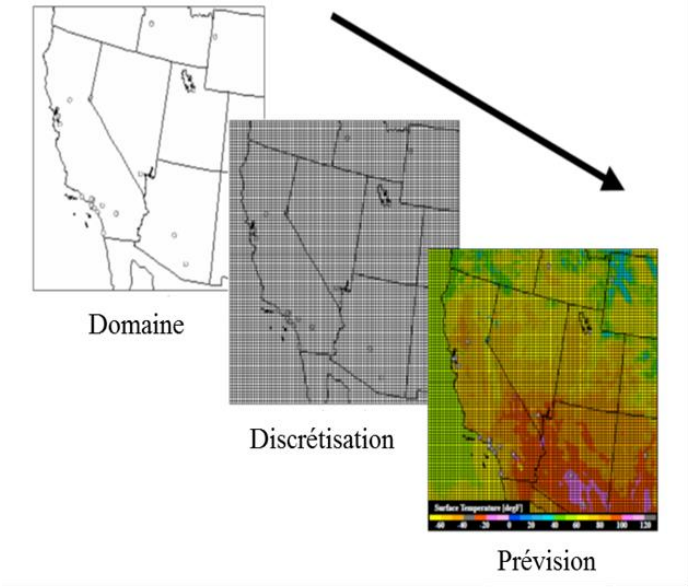


Figure III-4: Etapes de mise en œuvre de modèles NWP, définition du domaine, discrétisation et prévision, en 2D.

Ces étapes sont répétées pour toutes les couches de l'atmosphère afin de réaliser des prévisions les plus fidèles possibles à la réalité,

La Figure III-5 présente un exemple de discrétisation 3D de la Terre.

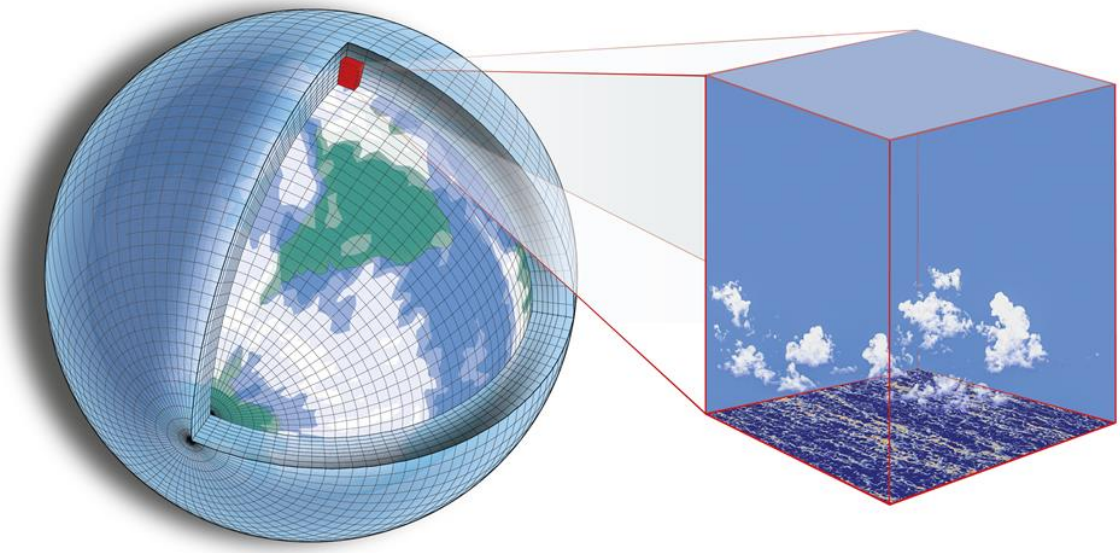


Figure III-5: Discrétisation de l'atmosphère sur le globe terrestre. (Source : https://climate-dynamics.org/).

En réalité, un modèle NWP n'est pas un outil unique mais plutôt un système de prévision numérique de la météo basé sur l'exploitation d'une chaîne de modèles complémentaires. Les principaux modèles météo sont les suivant :

- Le modèle GFS (américain) pour « Global Forecast System » est produit par le National Centers for Environmental Prediction (NCEP). Cet organisme est un regroupement de plusieurs centres nationaux de prévisions météorologiques aux États-Unis. Il fait également partie du National Weather Service (NWS).
  Parmi les 9 centres nationaux, l'Environmental Modeling Center (EMC) développe particulièrement le modèle GFS. Le modèle GFS est initialisé quatre fois par jour : 00 h, 06 h, 12 h et 18 h. Les calculs de prévisions brutes vont jusqu'à 384 h (16 jours). Sa résolution horizontale est de 27 km jusqu'à 192 h et 70 km de 192 h à 384 h. Il est à noter que GFS est un modèle libre et gratuit.
- Le modèle ECMWF (« European Centre for Medium-Range Weather Forecasts ») est un modèle utilisé pour la prévision allant jusqu'à 10 jours. Contrairement au modèle GFS, une grande partie des paramètres du modèle CEPMMT (Centre européen pour les prévisions météorologiques à moyen terme) ne sont pas accessibles gratuitement. Le modèle CEPMMT est initialisé deux fois par jour à 00 h et 12 h.
- Le modèle WRF (américain) pour « Weather Research and Forecasting » est un modèle météo utilisé par le National Weather Service des États-Unis et pour la recherche en simulation de l'atmosphère. C'est un modèle dit de méso-échelle avec une résolution horizontale entre 2 et 15 km. Il est aussi libre et gratuit. Le Weather Research and Forecasting est initialisé quatre fois par jour à 00 h, 06 h, 12 h et 18 h.
- Le modèle ARPEGE (Monde) pour « Action de Recherche Petite Echelle Grande Echelle » est un modèle qui couvre l'ensemble de la planète avec une maille fluctuante selon les zones géographiques (7,5 km en moyenne pour l'Europe). L'échéance de la prévision est de 4 jours. Le modèle ARPEGE est initialisé quatre fois par jour à 00 h, 06 h, 12 h et 18 h.
- Le modèle AROME (français) pour « Application of Research to Operations at MEsoscale » est un modèle avec une maille très fine (maille de 1,3 km) pour la prévision en France. L'échéance de la prévision est limitée à 36 heures. Ce modèle développé par Météo France appartient à la dernière génération de modèles. Grâce à sa maille très fine, il permet de mieux appréhender les

phénomènes convectifs tels que les orages, et ce, grâce à l'intégration de nouvelles données d'observation ou encore la prise en compte de la topographie, des villes, des cours d'eau, de la végétation, etc. Le modèle AROME est initialisé quatre fois par jour : 00 h, 06 h, 12 h et 18 h.

Lorsque des prévisions ont été générées, celles-ci peuvent être améliorées en les comparant à des données mesurées pendant une période d'évaluation au cours de laquelle des modèles améliorés sont développés et ajustés. Cette approche MOS (Model Output Statistic) fonctionne pleinement si les corrections de prévision sont mises à jour dans le temps (récursif) et élaborées séparément selon différentes conditions ou régimes. En effet, les erreurs de prévision dépendent souvent de l'heure et de l'année, des conditions du ciel, etc. Les modèles sont fréquemment modifiés et mis à jour, ce qui nécessite des corrections de prévision pour s'adapter, en conséquence on parlera de modèles de prévision et de correction adaptatifs.

Cette méthode a été appliquée par Lorenz *et al.* (2011) à la prévision de l'éclairement solaire en Allemagne en couplant la correction MOS avec l'indice ciel clair et à l'angle zénithal solaire afin de corriger le biais lié à la prévision. Une procédure similaire a été appliquée par Mathiesen et Kleissl (2011) pour les États-Unis et différentes structures de biais ont été trouvées pour différents modèles NWP.

L'élimination du biais a également été étudié par Pelland *et al.* (2011) pour les prévisions de rayonnement global en Amérique du Nord issues des modèles NWP, en utilisant un filtre linéaire de Kalman et une fenêtre d'entraînement glissante de 30 à 60 jours.

D'autres approches de prévision à partir des sorties de modèles NWP et de données historiques ont été réalisées. Perez *et al.* (2007), par exemple, ont proposé une méthodologie de prévision de l'éclairement solaire basée sur l'utilisation des prévisions de la couverture nuageuse du NDFD (National Digital Forecast Database) américain.

Diverses autres approches basées sur le MOS ont été mises en œuvre pour prévoir le rayonnement global horizontal ou la production photovoltaïque d'un à plusieurs jours pour différents types de méthodes :

- Linéaire : pour les méthodes autorégressives (Bacher *et al.*, 2009; C. Chen *et al.*, 2011);
- Intelligence artificielle : comme par exemple les RNA (réseaux de neurones artificiels) (Cao et Lin, 2008; A Chaouachi *et al.*, 2010; Mellit et Pavan, 2010; Yona *et al.*, 2007), des modèles boite grise avec RNA (Wang *et al.*, 2011) et des machines à vecteurs de support (Shi *et al.*, 2012).

2.2. Méthodes basées sur le formalisme des séries temporelles

L'apprentissage automatique (en anglais « machine learning ») est un domaine qui fait partie de l'informatique et est classé dans la famille des méthodes d'intelligence artificielle. L'utilisation de l'apprentissage automatique est aujourd'hui largement répandue dans de nombreux domaines, Ce type d'approche permet, de manière avantageuse, de résoudre des problèmes qui sont impossible à représenter par des algorithmes explicites. Les modèles d'apprentissage automatique sont en capacité de trouver des relations entre les entrées et les sorties même si la représentation est impossible, cela les rend particulièrement adaptés aux taches de prévisions. Pour la prévision du rayonnement solaire, les modèles se répartissent en différents sous-groupes (Aggarwal et Saini, 2014) :

- Modèles structuraux, basés sur des paramètres géographiques et météorologiques (utilisant aussi des séries temporelles).

- Modèles basés sur les séries temporelles, admettant en entrée les données de mesures du rayonnement solaire (modélisation univariée).
- Modèles hybrides prenant en compte les données de mesures du rayonnement ainsi que d'autres paramètres météorologiques (modélisation multivariée).

Afin d'orienter les travaux de recherche de cette thèse, nous avons tout d'abord réaliser un état de l'art de la prévision de l'éclairement solaire.

- **Etat de l'art de la prévision de l'éclairement global horizontal**

Les horizons de prévision et les granularités temporelles (pas de temps des données) nécessaires à nos recherches concernent les modèles basés sur le formalisme des séries temporelles et les méthodes de machine learning. L'analyse de cette étude bibliographique (non exhaustive) montre que certains modèles sont plus utilisés que d'autres ; la fréquence d'apparition des différents modèles dans l'état de l'art que nous avons réalisé, est assez révélatrice des travaux menés dans ce domaine.

Le Tableau III-1 contient les informations révélées par l'étude bibliographique qui a été réalisée.

Tableau III-1: Etude bibliographique non exhaustive des travaux de prévision par méthodes d'apprentissage automatique jusqu'en 2016

Référence	*Localisation*	*Horizon*	*Indice d'erreur*	*Commentaires/Conclusions*
(Burrows, 1997)	Canada	18 heures	-	Arbres de régression>Régression linéaire
(Mihalakakou *et al.*, 2000)	Grèce	1 heure	-	RNA équivalent à AR
(Mori, 2001)	Japon	1 jour	nMAE = 1,75%	{Arbres de régression -RNA}>RNA
(Podestá *et al.*, 2004)	Argentine	1 jour	-	Régression généralisée utile
(Tso et Yau, 2007a)	Chine	-	-	Arbres de régression > RNA > Régression linéaire
(Cao et Lin, s. d.)	Chine	1 heure	R^2 = 0,72	{RNA-wavelet}>RNA
(Reikard, 2009)	Etats Unis	1 heure	nRMSE = 26%	{ }=ARMA>RNA
(Aymen Chaouachi *et al.*, 2010)	Japon	1 heure	MAPE = 4%	{RNA}>RNA
(Gastón *et al.*, 2010)	Espagne	1 heure	-	{SVM-KNN}>climatologie
(Paoli *et al.*, 2010)	France	1 jours	nRMSE = 21%	RNA>AR>kNN>Bayesian>Markov
(Marquez et Coimbra, 2011)	Etats Unis	24 heures	nRMSE = 17,7%	RNA>persistance
(Moreno *et al.*, 2011)	Espagne	1 jour	-	RNA= régression généralisée
(Ben Taieb *et al.*, 2012)	Benchmark	Plusieurs	MAPE=18.95%	MIMO-ACFLIN (lazy learning)

(Chakraborty *et al.*, 2012)	Etats Unis	1 heure	-	Bayésienne>{SVM-RNA}
(Demirtas *et al.*, 2012)	Turquie	10 min	nRMSE = 18%	kNN>RNA
(Ferrari *et al.*, 2012)	Italie	1 heure	-	SVM>RNA>KNN>persistance
(Hossain *et al.*, 2012)	Australie	6 heures	-	{RNA-least median square}>least median square>RNA>SVM
(Mori et Takahashi, 2012)	Japon	1 heure	-	Arbres de régression pour la sélection des variables
(Olaiya et Adeyemo, 2012)	Nigeria	NA	nRMSE = 24%	RNA= Arbres de régression
(Bouzerdoum *et al.*, 2013)	Italie	10 min	nRMSE = 9,4%	{SARIMA-SVM}>SARIMA>SVM
(Chu *et al.*, 2013)	Etats Unis	10 min	Skill score = 20%	{GA-RNA}>RNA
(Prokop *et al.*, 2013)	République tchèque	10 min	-	{RNA-SVM}>SVM>RNA
(Aggarwal et Saini, 2014)	Etats Unis	1 jour	-	{RNA-régression linéaire}>RNA>régression linéaire
(Aggarwal et Saini, 2014)	Etats Unis	1 jour	-	{LSR-RNA}>LSR régularisée=LSR ordinaire
(Alobaidi *et al.*, 2014)	Emirats Arabes Unis	10 min	rRMSE = 9,1%	{RNA}>RNA
(Bilionis *et al.*, 2014)	Etats Unis	30 min	-	{ACP-Processus Gaussien}>NWP
(Fernández *et al.*, 2014)	Espagne	1 jour	-	SVM>persistance
(Huang *et al.*, 2014)	Australie	1 min	MAPE = 38%	Forêts aléatoires>régression linéaire
(Krömer *et al.*, 2014)	Canada	1 heure	-	SVM>NWP
(Long *et al.*, 2014)	Macao	1 jour	MAPE = 11.8%	RNA >SVM>KNN>linear regression
(Salcedo-Sanz *et al.*, 2014)	Espagne	1 jour	-	Extreme machine learning is useful
(J. Wu *et al.*, 2014)	Singapour	NA	-	{GA-kmean- RNA}> RNA >ARMA
(Y.-K. Wu *et al.*, 2014)	Malaisie	1 heure	nRMSE = 5%	{GA-SVM- RNA -ARIMA}>SVM>ANN>ARIMA
(Yang *et al.*, 2014)	Taiwan	1 jour	nMAE = 3%	{ANN-SVM}>SVM>ANN
(Zamo *et al.*, 2014)	Benchmark	10 min	nRMSE = 10%	Forêts aléatoires>SVM>régression généralisée>boosting>bagging>persistance
(Almeida *et al.*, 2015)	Espagne	24 heures	nMAE=3,73% - 9,45%	Quantile regression forests et NWP
(Wolff et Kramer, 2015)	Allemagne	1 heure	nRMSE = 6,2%	SVR>K-NN

(Chu *et al.*, 2015b)	Etats Unis	10 min	Skill score = 6%	{GA- RNA}>persistance
(De Felice *et al.*, 2015)	Italie	1 jour	MdAPE = 6%	SVM>linear model
(Dong *et al.*, 2015)	Etats Unis	1 heure	nRMSE = 22%	{RNA-SVM}>ARMA
(Lauret *et al.*, 2015)	Iles Françaises	1 heure	nRMSE = 19,6%	RNA =Processus Gaussiens=SVM> persistance
(Lazzaroni *et al.*, 2015)	Italie	1 heure	-	SVR> RNA >AR>KNN> persistance
(McGovern *et al.*, 2015)	Benchmark	1 heure	nRMSE = 13%	Arbres de régression >NWP
(Pedro et Coimbra, 2015)	Etats Unis	30 min	Skillscore = 23,4%	KNN>persistance
(Samanta *et al.*, 2015)	Etats Unis	1 heure	-	{SVR}>SVR>{SVR-ACP}>ARIMA>régression linéaire
(Cheng, 2016)	Taiwan	20 min	nRMSE = 25%	{Filtre de Kalman-Régression}>Filtre de Kalman> Régression

RNA=réseau de neurones artificiels, AR=auto régression, ACP=analyse en composante principale, SVR=support vector régression, SVM=support vector machine, kNN=K-nearest neighbour (k plus proches voisins), GA=algorithme génétique, LSR=least square régression, NWP=numerical weather prediction. Le signe > pour « meilleur que », {} pour hybridation.

La Figure III-6 présente une synthèse du nombre d'apparitions des principaux modèles dans les différents articles rassemblés au cours de l'étude bibliographique.

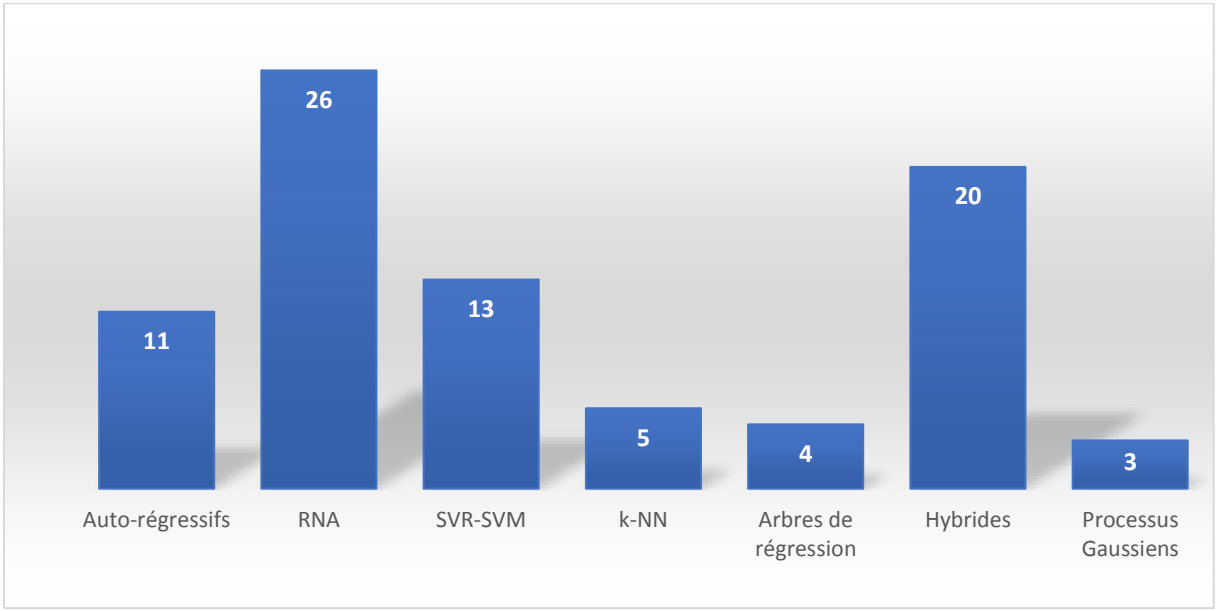


Figure III-6: Nombre d'apparitions des méthodes dans l'état de l'art réalisé

Les RNA ont été largement utilisés dans le domaine de la prévision de l'éclairement solaire, ils ont été étudiés dans de nombreuses régions du monde et des chercheurs ont démontré la capacité de ces techniques à réaliser une prévision basée sur l'utilisation des séries temporelles de données météorologiques avec précision. En effet, les RNA, et plus particulièrement, les perceptrons multicouches (PMC) sont assez régulièrement utilisés pour leur propriété « d'approximation universelle » et leur capacité de réaliser des régressions non linéaires sur des données.

Les méthodes basées sur la régression linéaire (ARMA, ARIMA, SARIMA) sont aussi très utilisées, même si elles sont généralement légèrement moins efficaces que les méthodes basées sur les RNA, elles peuvent être probantes dans certains cas.

La plupart des articles dans la littérature concernent des comparaisons entre différents modèles plus ou moins complexes, les modèles sont comparés entre eux, et souvent hybridés pour observer les améliorations possibles dans leurs performances dues à ces hybridations. On entend par hybridation le fait d'assembler différents modèles pour les utiliser ensemble et observer les effets sur les performances lors des prévisions. Cette méthode est souvent génératrice d'amélioration mais aussi de complexité des modèles.

Certains modèles sont plus largement développés dans la prévision que d'autres. Cette tendance montre que certaines méthodes encore peu utilisées peuvent être intéressantes à étudier, dans ce cas, toutes les méthodes dérivées des arbres de régression (simples, élagués, renforcés, ensachés, forêts aléatoires) peuvent être utilisées.

Notons que, le deep learning, branche du machine learning basée sur un ensemble d'algorithmes modélisant des abstractions de haut niveau dans des données en utilisant des architectures de modèle particulièrement complexes, composées de multiples transformations non linéaires, n'est pas pris en compte.

- **Etat de l'art de la prévision des composantes directe normale et diffuse horizontale du rayonnement**

Etant donnés les résolutions temporelles et les horizons de prévision que nous avons choisis, nous nous sommes concentrés sur les méthodes de prévision à court terme et basées notamment sur le formalisme des séries temporelles.

Dans Law *et al.* (2014), trois références seulement ont été citées pour la prévision de l'éclairement direct normal : en utilisant le formalisme des séries temporelles et du PMC pour les horizons courts et les modèles NWP, de vecteurs de mouvement de nuages et satellitaires pour les autres horizons. Law *et al.* (2014) ont conclu que des recherches supplémentaires étaient nécessaires. Marquez et Coimbra (2011) ont constaté que l'éclairement direct est généralement beaucoup plus difficile à prévoir que l'éclairement global. Pour les mêmes jours et pour les prévisions avec un horizon de 1 jour, ils ont trouvé une plage de nRMSE comprise entre 15 et 22% pour l'éclairement global, et entre 28 et 35% pour les prévisions d'éclairement direct (Zhu *et al.*, 2017). Lara-Fanego *et al.* (2012) ont évalué la fiabilité des prévisions de l'éclairement global horizontal et de l'éclairement direct normal pour un horizon de 3 jours par un modèle atmosphérique à méso échelle ; pour l'éclairement global la nRMSE allait de moins de 10% par temps clair à 50% par temps nuageux alors que pour l'éclairement direct, l'erreur variait de 20% à 100% entre le ciel clair et le ciel nuageux.

Ghofrani *et al.* (2016) ont développé un modèle informatique utilisant des techniques de clustering, une méthode de classification des séries temporelles et une méthode basée sur les réseaux de neurones artificiels pour l'éclairement global horizontal et direct normal sur plusieurs sites aux Etats-Unis et pour des horizons temporels allant de 1 heure à 48 heures. Les résultats pour l'horizon 1 heure étaient en termes de nRMSE entre 16,9% et 40,5% pour l'éclairement direct normal et entre 5,5% et 11,6% pour l'éclairement global horizontal.

Un modèle de prévision basé sur les ondelettes, utilisant l'indice de ciel clair en entrée, a été proposé pour estimer l'éclairement direct normal pour un horizon de 10 minutes. Ses performances en termes de nMAE étaient entre 0,84 et 7,66% (erreur absolue moyenne normalisée) et en termes de nRMSE entre 1,89 et 10,99% pour quatre stations différentes (Zhu *et al.*, 2017).

Li *et al.* (2016) ont analysé les prévisions d'éclairement direct normal à 5, 10 et 15 minutes à l'aide d'un modèle déterministe et d'un perceptron multicouche et ils ont quantifié l'impact du coefficient de transmission et de la vitesse des nuages sur les prévisions à court terme.

Chu *et al.* (2015a) ont proposé des modèles de prévisions intra horaires pour l'éclairement direct normal en associant le traitement d'image du ciel et l'optimisation par réseaux de neurones artificiels ; Les modèles de prévision hybrides permettent d'obtenir des prévisions statistiquement robustes avec une amélioration de nRMSE de 20% par rapport à la persistance pour les prévisions à 5 et 10 minutes.

Un modèle hybride, qui intègre des techniques d'imagerie du ciel, et qui prend en charge des sous-modèles de machine à vecteurs support (SVM) et des réseaux de neurones artificiels a été développé par Chu *et al.* (2015a) pour quatre horizons de prévision infra horaires : 5, 10, 15 et 20 minutes ; les RMSE obtenues variaient de 93,1 W/m² pour l'horizon de 5 minutes à 131,8 W/m² pour l'horizon de 20 minutes.

Deux approches différentes (Chauvin *et al.*, 2014) basées sur les systèmes à inférences floues avec un réseau adaptatif (ANFIS) ont été développées afin de prévoir l'éclairement direct normal par ciel clair pour des horizons allant de 30 minutes à 5 heures. La nRMSE du meilleur modèle variait entre 1,5% et 2,8%, tout en gardant à l'esprit que cette prévision ne s'appliquait que par ciel clair.

Pedro *et al.* (2018) ont comparé les performances des méthodes d'apprentissage automatique (k-plus proches voisins (kNN) et gradient boosting (GB)) pour la prévision intra horaire d'éclairement global horizontal et direct normal ; ils ont prévu la valeur de l'éclairement solaire et les intervalles de prévision correspondants et ils ont montré que les modèles d'apprentissage automatique amélioraient considérablement les prévisions par rapport au modèle de référence (réduction de la nRMSE de 8% à 24% et de 10% à 30% respectivement pour l'éclairement global horizontal et direct normal).

De cette brève étude bibliographique (car les articles qui en traitent sont peu nombreux) sur les prévisions de l'éclairement direct normal, nous pouvons conclure que les prévisions de cette composante sont obtenues avec une précision inférieure à celle de l'éclairement global horizontal et que seul un petit nombre d'articles ont été écrits sur ces prévisions à court terme, comme le confirment Law *et al.* (2014).

Cette étude bibliographique nous a fourni des informations sur les différents modèles couramment utilisés. L'état de l'art alors réalisé a permis d'orienter notre étude vers d'autres modèles, moins utilisés ou utilisés dans d'autres domaines mais néanmoins intéressants à développer.

Nous avons développé et utilisé douze modèles différents, classés en deux grandes familles :

- Les modèles sans phase d'apprentissage
- Les modèles avec phase d'apprentissage.

2.2.1. Les modèles sans phase d'apprentissage

Nous avons utilisé 3 modèles dits sans apprentissage, exposés par ordre croissant de complexité.

- **Persistance simple et persistance intelligente :**

Le premier modèle utilisé est le modèle de persistance (Diagne *et al.*, 2013). Il s'agit d'un modèle « naïf ». Il consiste à répéter la valeur d'une grandeur de l'instant t à l'instant t+h. Il est formalisé de manière mathématique par :

$$\hat{\dot{I}}(t+h) = \dot{I}(t) \qquad \text{III-1}$$

On parle de persistance en raison de la persistance des conditions météorologiques. Ce modèle peut faire l'objet d'une amélioration par couplage avec un modèle de connaissance prenant en compte la variation de la position du soleil et donc de l'angle d'incidence, par l'intermédiaire d'un modèle d'éclairement hors atmosphère ou par ciel clair (ce dernier est plus performant mais plus difficile à élaborer). Cette amélioration, facile de mise en œuvre, améliore considérablement les résultats par rapport à la persistance simple.

Le modèle de persistance intelligente (Voyant *et al.*, 2015), utilisant un modèle de rayonnement solaire par ciel clair pour tout horizon h est défini par :

$$\hat{\dot{I}}(t+h) = \dot{I}(t).\frac{\dot{I}^{cs}(t+h)}{\dot{I}^{cs}(t)} \qquad \text{III-2}$$

Ce modèle permet de réaliser une persistance sur l'indice de ciel clair. Ce type de prédicteur est quelques fois le seul à pouvoir être utilisé en mode opérationnel car il ne nécessite pas de données historiques et peut donc prédire même en cas de défauts sur les mesures.

- **Persistance intelligente couplée au filtre de Kalman :**

Le filtre de Kalman est utilisé dans une multitude de domaines technologiques. C'est un pilier de l'automatique et du traitement du signal, initialement voué au traitement des signaux bruités, notamment dans le domaine des radars et des communications.

Ses qualités et sa flexibilité ont conduit à étendre son utilisation dans beaucoup d'autres domaines, comme la finance, la navigation, la météorologie ou l'océanographie pour l'assimilation de données par les modèles numériques. Un autre champ d'applications possible est la prévision des séries temporelles et, en particulier, du rayonnement solaire.

La description mathématique complète des filtres de Kalman est très complexe ; les lecteurs intéressés peuvent se référer au livre « Stochastic Processes and Filtering Theory » (Jazwinski, 1970).

Dans le cadre de notre étude, nous avons utilisé les filtres de Kalman comme un outil d'amélioration du modèle de persistance intelligente. Il s'agit d'un estimateur récursif (Cheng, 2016), c'est une compilation de l'estimation à l'instant présent seulement avec l'estimation à l'instant précèdent et la mesure de l'instant présent (Chaabene et Ben Ammar, 2008).

Le filtre de Kalman peut être écrit sous la forme d'une seule équation cependant il est plus souvent défini par deux parties qui sont complémentaires :

- La partie de prévision
- La partie de mise à jour.

La phase de prévision est basée sur l'état de la grandeur à l'instant précédent pour produire une prévision de l'instant présent. Cet état prédit est appelé « a priori », en effet, bien qu'il concerne l'instant présent, au moment où il est généré, il n'est pas possible de disposer d'observation (mesure) de cet instant.

Dans la phase suivante dite de « mise à jour » il est fait une compilation de la prévision « a priori » avec la mesure de l'instant présent pour affiner la prévision. Le résultat de la compilation est alors appelé prévision « a posteriori ». Pour le cas général, l'algorithme de prévision d'un éclairement noté $\dot{I}$ pour cette adaptation du modèle s'écrit :

$$\dot{I}(t+1) = A(t).\dot{I}(t) + \omega(t) \qquad \text{III-3}$$

avec ω un bruit blanc auquel correspond une covariance $Q(= N(0, Q))$ et :

$$A(t) = \frac{i^{cs}(t+h)}{i^{cs}(t)} \quad \text{III-4}$$

A l'instant t, une observation (mesure) de l'état réel du rayonnement I(t), z(t) est de la forme suivante :

$$z(t) = M(t).\dot{I}(t) + v(t) \quad \text{III-5}$$

Dans laquelle :

- $v(t)$ est le bruit de l'observation qui est supposé être un bruit blanc Gaussien moyen avec une covariance $R(= N(0, R))$.
- M(t) est la matrice qui relie l'état réel à la mesure, dans notre cas nous sommes partis de l'hypothèse que la mesure est l'état réel donc M est une matrice identité.

L'état initial du modèle et le bruit à chaque pas de temps sont supposés mutuellement indépendants. Dans le cadre de notre problématique, la prévision définissant le vecteur d'état est définie par l'équation suivante :

$$\hat{I}(t|t) = \hat{I}(t-1|t-1).A(t-1) = \hat{I}(t-1|t-1).\frac{i^{cs}(t+h)}{i^{cs}(t)} \quad \text{III-6}$$

$\hat{I}_{t|t}$ est la valeur prédite de l'éclairement à l'instant t. Pour résumer, il s'agit en fait d'un état prédit « a posteriori » au temps t, qui provient d'observations jusqu'à l'instant t. Il est nécessaire ensuite de juger de la précision du modèle.

On calcule la matrice de covariance de l'erreur « a posteriori » qui est appelée P (une estimation de l'état prédit). Lors de la compilation de cette matrice P, nous avons défini une « fenêtre glissante » pour le calcul de la matrice de covariance permettant de prendre en compte la saisonnalité due à la nature des données constituant le signal.

La définition de la forme de P est donc :

$$P(t|t-1) = A(t-1).P(t-1|t-1).A(t-1)^T + Q \quad \text{III-7}$$

On déduit ensuite le gain du filtre de Kalman noté G défini par :

$$G(t) = P(t|t-1).M(t).(M(t).P(t|t-1) + R)^{-1} \quad \text{III-8}$$

On introduit alors le facteur de correction décrit et définit par A.C. Harvey (1990):

$$\hat{I}(t|t) = \hat{I}(t|t-1) + G(t).\left(z(t) - M(t).\hat{I}(t|t)\right) \quad \text{III-9}$$

A partir de ces formulations, on définit le modèle de prévision pour un horizon h=1 :

$$\hat{I}(t+1|t) = A(t).\hat{I}(t|t) \quad \text{III-10}$$

et la formule généralisée pour un horizon h quelconque :

$$\hat{I}(t+h|t) = \prod_{i=1}^{h} A(t+h-i)\,.\hat{I}(t|t) \quad \text{III-11}$$

De façon moins formelle, le filtre de Kalman est un modèle qui :

- Compare la prévision à la mesure ;
- Tente à chaque instant de faire recoller la prévision à la mesure.

L'amélioration de la persistance intelligente par les filtres de Kalman est intéressante, elle nous permet de réaliser des prévisions de qualité sans pour autant avoir accès à une grande quantité de données.

Une fois que les mesures commenceront à alimenter la base de données il sera possible de réaliser le réajustement du modèle en poursuivant la phase de prévision.

Ce modèle améliore la prévision réalisée par la persistance intelligente. Elle donne la possibilité d'effectuer de la prévision dès la première mesure disponible. Les lecteurs intéressés par ces travaux pourront se référer au papier de Voyant *et al.* (2017a).

Cette famille de modèles est intéressante puisqu'il n'est pas nécessaire de disposer d'historiques de données importants. Cependant, ils présentent des performances relativement restreintes avec une appréhension difficile en raison des variations brusques dans le rayonnement. Ils sont très souvent utilisés en tant que modèle de référence pour mieux estimer l'apport des modèles prédictifs plus complexes.

2.2.2. Les modèles avec phase d'apprentissage

En termes d'apprentissage automatique, on distingue trois types d'apprentissage :

- L'apprentissage supervisé, le modèle est alimenté avec des exemples d'entrées et leurs résultats souhaités, le but est d'apprendre une règle générale qui fait correspondre les entrées aux sorties.
- L'apprentissage non supervisé, le modèle est capable de trouver une structure cachée dans ses entrées sans connaître les résultats correspondants.
- L'apprentissage d'ensemble, pour lequel il est nécessaire de former plusieurs apprenants dits « faibles » en tant que membres d'un ensemble plus grand. On combine ensuite leurs prévisions en un seul résultat pour parvenir à une meilleure performance.

Plusieurs modèles plus ou moins complexes ont été utilisés pour réaliser la prévision. Cette section concerne les modèles pour lesquelles une phase d'apprentissage est nécessaire avant de pouvoir les utiliser.

2.2.2.1. Modèle autorégressif à moyenne mobile (ARMA)

Dans les années 1970, Box et Jenkins (1976) ont décrit la méthodologie de l'analyse des séries temporelles et ils ont ainsi développé la famille de modèles constitués par les modèles auto régressifs à moyenne mobile.

Il s'agit d'une approche statistique de l'analyse des séries temporelles. Une des combinaisons de modèle auto régressif (AR) avec un modèle de moyenne mobile (MA) permet le développement du modèle ARMA.

Ce modèle, utilisé pour la prévision de séries temporelles stationnaires, est utilisé dans de nombreux domaines et plus particulièrement dans le domaine de la prévision du rayonnement solaire, il s'agit d'un modèle linéaire.

On définit les modèles AR et MA, respectivement d'ordre p et q, pour une série temporelle x(t) et $t \in Z$ par :

AR(p):

$$x(t) = \sum_{i=1}^{p} \varphi(i).x(t-i) + \varepsilon(t) \qquad \text{III-12}$$

MA(q) :

$$x(t) = \sum_{i=1}^{q} \theta(i).\varepsilon(t-i) \qquad \text{III-13}$$

La combinaison des deux modèles devient alors ARMA (p, q) :

$$x(t) = \varepsilon(t) + \sum_{i=1}^{p} \varphi(i).x(t-i) + \sum_{i=1}^{q} \theta(i).\varepsilon(t-i) \qquad \text{III-14}$$

Où $\varphi(i)$ et $\theta(i)$ sont les paramètres du modèle, p et q les ordres et ε le résidu qui est un bruit blanc si le modèle est bien paramétré et la série temporelle est stationnaire.

Le modèle ARMA est donc un modèle basé sur une combinaison de mesures et d'erreurs antérieures pour caractériser une donnée actuelle.

La phase d'optimisation de ce modèle détermine les ordres du modèle. Les paramètres du modèle (φ et θ) sont déterminés pendant la phase d'apprentissage. Ce modèle est très utilisé dans l'analyse des séries temporelles, il est assez peu gourmand en ressources et est rapide à mettre en œuvre.

Il existe une multitude d'évolutions de ce type de modèle telles que ARMAX, SARMA, SARIMA… Le lecteur intéressé peut se référer à l'ouvrage de Box et Jenkins sur l'analyse des séries temporelles (Box et Jenkins, 1976).

2.2.2.2. Réseau de neurones artificiels, le perceptron multi couches (PMC)

Les réseaux de neurones artificiels (RNA) sont directement inspirés du fonctionnement des neurones dans un cerveau réel. Dans un RNA, les neurones artificiels sont interconnectés de manière semblable aux neurones réels dans un cerveau.

Chaque neurone reçoit une entrée, la traite et donne une sortie qui alimente le neurone suivant etc. Ce domaine de recherche a connu un succès grandissant tout au long de la seconde moitié du XX^{ème} siècle. Les premiers travaux sont dus à Mc Culloch et Pitts et datent de 1943. La démarche de base des travaux est de tenter de comprendre et d'expliquer le fonctionnement des systèmes nerveux, et ce à partir de composants élémentaires. Ils commencent, alors, par représenter un neurone par un modèle simple : le neurone formel, et prouvent qu'il est possible de construire des systèmes à base de neurones simples pour le calcul de fonctions logiques.

C'est en 1949 que Donald Hebb donne une première explication au travers de la loi de Hebb, il se concentre sur les connexions entre les cellules, ces dernières modifient l'intensité de ces connexions (poids) qui les relient entre elles. Peu de temps après, le développement de modèles software et hardware a permis de concrétiser ces travaux ; pour la première fois, on a pu simuler le fonctionnement de base du système nerveux.

En 1957, Frank Rosenblatt pose les bases du premier perceptron, il s'agit alors du premier système artificiel capable d'apprendre par expérience. On doit la première utilisation concrète des réseaux de neurones artificiels à Widrow et Hoff en 1960, ils réalisent le formalisme de « l'Adaline » pour ADAptative LINear Element. Ce réseau fait du traitement du signal en supprimant les échos sur une ligne téléphonique.

La publication, en 1965, de « Machine learning » par Nilsson pose les bases mathématiques de l'apprentissage automatique et notamment pour la reconnaissance des formes.

Les débuts des réseaux de neurones sont très prometteurs. Cependant, ce type de modèle perd en popularité face à des techniques de calculs plus classiques.

En 1969, Minsky et Paper sont les principaux détracteurs des réseaux de neurones, ils mettent en évidence les lacunes des réseaux trop simples (une couche) et la complexité de mise en œuvre des réseaux multicouches. Cette période marquera le début du déclin des réseaux de neurones qui seront

délaissés pendant une dizaine d'années, les autres techniques d'intelligence artificielle montreront également leurs faiblesses aux termes de cette dizaine d'années.

Dans les années 80, Hopfield (1982-1984) remettra les réseaux de neurones au goût du jour en créant les premiers réseaux totalement interconnectés, Kohonen (1982) développe, dans le même temps, des cartes auto-organisatrices.

En 1986, la contribution majeure aux réseaux de neurones est due à McClelland et Rumelhart (1981) qui développent l'algorithme d'apprentissage avec rétropropagation de l'erreur.

- **Du neurone biologique au neurone formel :**

Pour comprendre le fonctionnement des réseaux de neurones artificiels il faut partir du modèle originel, en effet, comme nous l'avons dit dans le paragraphe précèdent, l'origine de cette modélisation est l'étude des mécanismes du cerveau et notamment de l'apprentissage. C'est dans cette optique que les premiers neurones formels ont été représentés en suivant la construction du neurone biologique.

La Figure III-7 présente le schéma de la structure du neurone biologique.

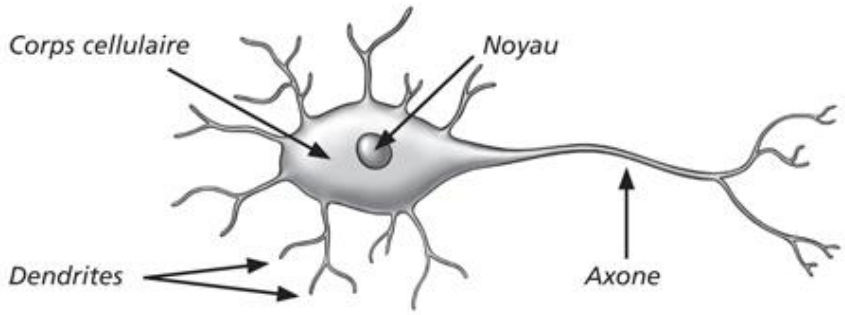


Figure III-7: Représentation schématique du neurone biologique

Les neurones sont un type de cellule bien spécifique, ils ne servent qu'au fonctionnement du système nerveux.

Ils sont composés d'un noyau entouré par le corps cellulaire et prolongé par les dendrites et axones. La transmission entre les cellules est faite par des signaux de nature électrochimique.

Les neurones reçoivent ces signaux par le biais de neurotransmetteurs chimiques au niveau des zones synaptiques, zones de connexion des dendrites aux neurones voisins.

Chaque synapse produit un signal unitaire celui-ci est appelé signal postsynaptique. Au niveau de la connexion du neurone vers l'axone a lieu une sommation des potentiels postsynaptiques qui parviennent au neurone. Si cette somme dépasse le seuil d'excitabilité du neurone, alors le signal est propagé via l'axone vers les synapses suivantes qui libèrent, à leur tour, les neurotransmetteurs pour les synapses et dendrites suivants et ainsi de suite. En partant de cela, il a été possible de représenter, de façon formelle, les neurones artificiels utilisés au sein de RNA.

En effet, le neurone formel est directement inspiré de la représentation du neurone biologique. Avec ses entrées, les poids attribués à celles-ci, la sommation de ces entrées pondérées avec l'introduction d'un biais et la fonction d'activation avant d'alimenter le neurone suivant par la sortie (axone).

La nature de la fonction d'activation du neurone détermine la nature du réseau lui-même, linéaire ou non, il est donc important de choisir la fonction d'activation selon le type de phénomènes plus ou moins complexes à modéliser. La nature même des données à traiter nous permet de faire ce choix afin d'adapter au mieux le paramétrage du modèle.

La Figure III-8 représente le neurone formel avec les différents paramètres qui lui sont propres.

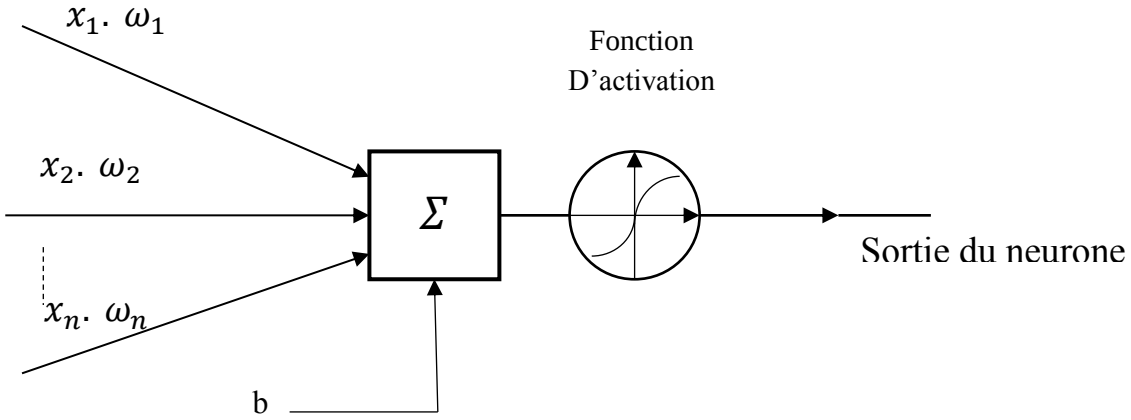


Figure III-8: Représentation du neurone formel avec les entrée x, leurs poids w et le biais b

Les valeurs des poids et des biais sont déterminées lors de la phase d'apprentissage du réseau réalisée à l'aide de plusieurs algorithmes.

- **Du neurone formel au perceptron multicouche (PMC) :**

Le perceptron multicouche (PMC) est une déclinaison des réseaux de neurones remarquable par sa construction dite « feedforward » c'est-à-dire non bouclée. Il est composé d'au minimum de deux couches de neurones, une cachée et une couche de sortie. Il est possible d'utiliser ce type de modèle pour traiter des problèmes de classification ou encore de régression. Il est utilisable pour réaliser de la prévision de séries temporelles. Son architecture ainsi que la nature de la fonction de transfert en fait un approximateur universel (Hornik *et al.*, 1989), il peut appréhender des relations non linéaires entre les données.

La Figure III-9 représente un PMC avec une couche cachée et une couche de sortie, n variables d'entrée et un neurone de sortie, $\hat{x}$ est la valeur prédite du vecteur d'entrée de la variable x.

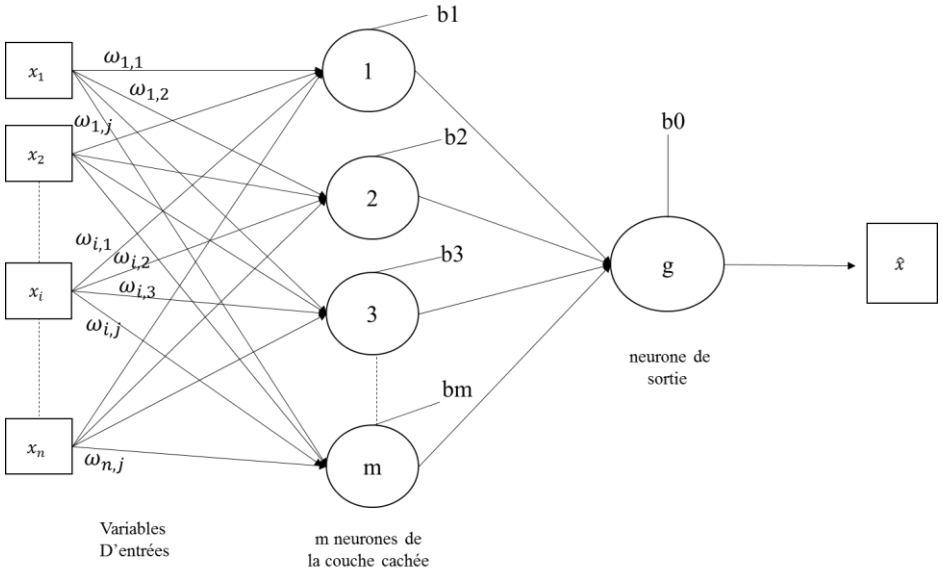


Figure III-9: Représentation d'un réseau de neurones à une couche cachée et une couche de sortie, avec les n entrée, les m neurones cachés, des exemples de poids, les biais de la couche cachée et de la couche de sortie, le neurone de sortie, son biais b0 et sa fonction g

Pour le $j^{ème}$ neurone de la couche cachée, un poids $w_{k,j}$ dont les valeurs sont déterminées pendant la phase d'apprentissage, est lié à chaque entrée x_k .

Une fonction d'activation f est appliquée à la somme pondérée ($S_j = \sum_{k=1}^{n} w_{k,j} x_k$) pour calculer une sortie du neurone si cette somme dépasse un biais donné (b_j) ($o_j = f(\sum_{k=1}^{n} w_{k,j} x_k + b_i)$. Cette sortie est alors distribuée aux neurones suivants : une fonction sigmoïde pour les couches cachées et une fonction linéaire pour la couche de sortie sont définies comme fonctions d'activation (Cybenkot, 1989). Pour la régression de la série temporelle $x(t)$, l'expression mathématique d'un PMC avec une couche cachée de m neurones, un neurone de sortie et n variables d'entrée est une fonction décrite par :

$$\hat{x}(t+h) = \sum_{j=1}^{m} \omega_{*j}\, o_j + b_o \qquad \text{III-15}$$

Avec x le vecteur d'entrées contenant les n valeurs de la variable ($K_g, K_b\ ou\ K_d$), $\hat{x}(t+h)$ la valeur de sortie qui correspond à la prévision par le modèle à l'horizon t+h, b_j et b_o les biais correspondants au neurone caché j et à la sortie, et $\omega_{k,j}$ les poids entre la k-ème entrée et le j-ième neurone caché, f désigne la fonction de transfert (tangente hyperbolique dans notre cas), ou d'activation des neurones de la couche cachée, ω_{*j} le poids entre la sortie et le j-ième neurone caché.

Cette fonction de transfert à l'avantage, par nature, d'être infiniment dérivable. Cette propriété de la fonction de transfert est importante durant la phase d'apprentissage, elle rend le PMC non linéaire. La formule mathématique de cette fonction est la suivante :

$$tanh(x) = \frac{2}{(1+e^{-2x})} - 1 \qquad \text{III-16}$$

Ainsi sa représentation graphique est présentée en Figure III-10 :

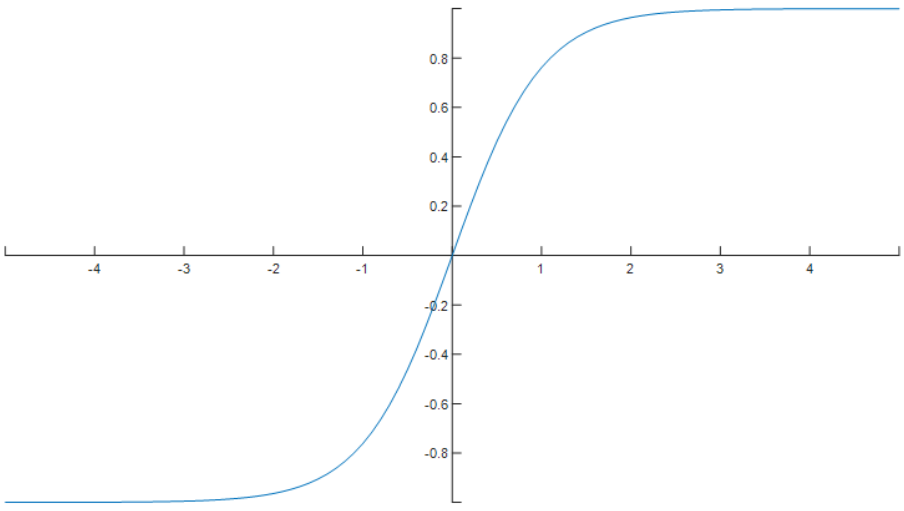


Figure III-10: Représentation de la fonction tangente hyperbolique sigmoïde

La fonction de transfert du neurone de sortie est la fonction $f(x) = x$. C'est une fonction de transfert linéaire qui revient à appliquer un coefficient multiplicateur intégré aux paramètres du neurone de sortie.

2.2.2.3. Les processus Gaussiens (PG)

Les processus Gaussiens sont des modèles dont le développement est assez récent (Rasmussen et Williams, 2006). Ce type de modèles est, en réalité, une généralisation d'une distribution gaussienne (ou normale) multivariée. Il s'agit de modèles non linéaires. Le lecteur intéressé se référera à Lauret *et al.* (2012).

Pour une prévision univariée, la formulation mathématique des modèles est la suivante :

$$\hat{x}(t+h) = \sum_{i=1}^{n} \alpha_i . k_f(\boldsymbol{x_i}, \boldsymbol{x}_{test}) \qquad \text{III-17}$$

Avec n le nombre de données d'apprentissage, $\boldsymbol{x_i}$ est le i[ème] vecteur d'entrées pour l'apprentissage et $\boldsymbol{x}_{test}$ le vecteur d'entrées test. Avec la fonction de covariance définie par :

$$k_f(x_p, x_q) = \sigma_f^2 . \exp\left(\frac{-(x_p - x_q)^2}{2l^2}\right) \qquad \text{III-18}$$

Dans laquelle σ_f^2 et l sont les hyperparamètres de la fonction de covariance, ils définissent la complexité du modèle et sont aussi déterminés lors de la phase d'apprentissage. Dans ce cas le coefficient α_i est déterminé lors de la phase d'apprentissage (issu de l'application de la fonction de covariance sur les données d'apprentissage) en reliant les données d'entrée avec le vecteur des n valeurs cibles de l'apprentissage.

2.2.2.4. Machines à vecteurs supports pour la régression (MVS)

Les machines à vecteurs de support sont un ensemble de techniques d'apprentissage supervisé destinées à traiter des problèmes de discrimination ou de régression. Ce sont des classifieurs linéaires généralisés.

Le développement de ces modèles est directement issu des travaux de Valdimir Vapnik dans les années 80 (Vapnik, 1986). Les « support vector regression » (SVR) sont issus de l'adaptation des machines à vecteur de support utilisés pour la régression sur les données. Cette méthode a été appliquée avec succès à la prévision de séries temporelles. Le formalisme des SVR est similaire à celui des processus Gaussiens.

La formule mathématique des SVR appliquée à la prévision de séries temporelles est (Lauret *et al.*, 2015):

$$\hat{x}(t+h) = \sum_{i=1}^{n} \alpha_i . k_{rbf}(\boldsymbol{x_i}, \boldsymbol{x}_{test}) + b \qquad \text{III-19}$$

Dans laquelle, $\boldsymbol{x_i}$ est le i[ème] vecteur d'entrées pour l'apprentissage et $\boldsymbol{x}_{test}$ le vecteur de test. La fonction de base radiale est donnée par :

$$k_{rbf}(x_p, x_q) = exp\left[\frac{-(x_p - x_q)^2}{2\sigma^2}\right] \qquad \text{III-20}$$

Le paramètre b (paramètre de biais) est déduit de l'équation précédente, l'hyperparamètre σ également, il définit notamment la complexité du modèle.

Concernant les SVR, les coefficients α_i sont liés à la différence de deux multiplicateurs de Lagrange qui sont déduits par la résolution d'un problème de programmation quadratique. Contrairement aux réseaux de neurones artificiels, sujets aux problèmes de minimums locaux, pour les SVR le problème

est strictement convexe et ne possède donc pas une seule et unique solution. En outre, il faut souligner (contrairement aux processus Gaussiens) que tous les modèles d'entraînement ne participent pas à la relation précédente. En effet, un choix pratique d'une fonction de coût (fonction ε de Vapnik) dans le problème quadratique, permet d'obtenir une solution particulière construite à partir d'une régression. A noter alors que certains des coefficients α_i seront non-nuls.

2.2.2.5. Les arbres de régression

Les arbres de régression sont une évolution des arbres de décision, utilisables dans de nombreux domaines de traitement des données.

Un arbre de décision représente une suite de choix en choix qui aboutit à une décision finale. De manière graphique, on représente le processus par un arbre binaire, chacun des choix étant une fourche entre plusieurs branches et les décisions étant les feuilles. Ce type de modèle à l'avantage de proposer une représentation graphique particulièrement lisible et compréhensible par l'utilisateur. Ce type de méthode, établi dans les années 1960, fait partie des méthodes de partitionnement récursif.

Le formalisme tel que nous le connaissons aujourd'hui a été réalisé par Breiman *et al.* (1984) sous l'acronyme CART : « Classification and Regression Tree ». Cet acronyme compile les deux formes de cette modélisation, qualitative (classification) et quantitative (régression).

Ce type de modèle peut être utilisé pour de la prévision, dans différents domaines (De'ath, 2007; Troncoso *et al.*, 2015; Tso et Yau, 2007b). Il existe plusieurs déclinaisons des arbres de régression. Dans les sections suivantes, nous développerons les types d'arbres de régression utilisés, les arbres de régression simples, élagués (pruned), renforcés (boosted), ensachés (bagged) et forêts aléatoires (random forest).

- **Les arbres de régression simples :**

Les arbres de régression classiques ont été la première déclinaison des arbres de décision utilisés pour traiter des problèmes de régression sur les données. Ce sont Hastie et Tibshirani (1986) qui ont proposé le premier formalisme de ce modèle régressif, la formule mathématique est :

$$\hat{x}(t+h) = \sum_{i=1}^{t-1} k_i \times H\big(x(t-i)\big) \qquad \text{III-21}$$

Où k_i est un facteur, H est une fonction qui retourne 1 si la donnée est utilisée ou 0 si elle ne l'est pas.

Une fois que l'arbre est construit, un modèle de régression est appliqué à chaque nœud. Lors de la phase d'apprentissage, un processus itératif, on cherche à minimiser l'erreur entre la valeur à l'instant t est la même valeur prédite. Dans notre cas, lors de la phase d'apprentissage nous utilisons l'erreur quadratique en fixant une tolérance sur la valeur de cette erreur. Cette tolérance est fixée proche de 0 (de l'ordre de $1x10^{-3}$). Dans la famille des arbres de régression pour la prévision, cette méthode est la moins optimisée et peut être gourmande en ressources et ne donner que des performances assez limitées.

- **Les arbres de régression élagués « pruned » :**

Il s'agit de la première méthode, souvent très efficace, d'amélioration des arbres de régression.

En effet, dans un arbre de régression classique (comme ceux du paragraphe précédent), on laisse grandir les branches de l'arbre sans prendre en compte les effets de la croissance des branches. Dans le cas de l'élagage le principe est sensiblement le même, excepté le fait, que l'on réduit le nombre de nœuds de l'arbre. Les arbres élagués sont construits en augmentant la tolérance d'erreur quadratique par nœud.

La division des nœuds s'arrête lorsque l'erreur quadratique par nœud tombe en dessous d'une tolérance donnée.

Nous avons vu que, dans le cas des arbres de régression classique, la tolérance est fixée proche de 0, dans le cas des arbres élagués ; il est choisi une valeur plus élevée en utilisant une méthode heuristique consistant à minimiser l'erreur de prévision. De manière plus formelle, comme dans l'équation précédente, la fonction I retourne plus souvent 0 que 1.

La Figure III-11 est un exemple de représentation graphique d'un arbre de régression, la différence avec un arbre de décision vient du fait que les divisions à chaque nœud sont réalisées de manière quantitative, et non qualitative. Sur un arbre de classification les conditions de séparation au niveau d'un nœud peuvent être des réponses à des questions logiques (oui/non, si/alors…).

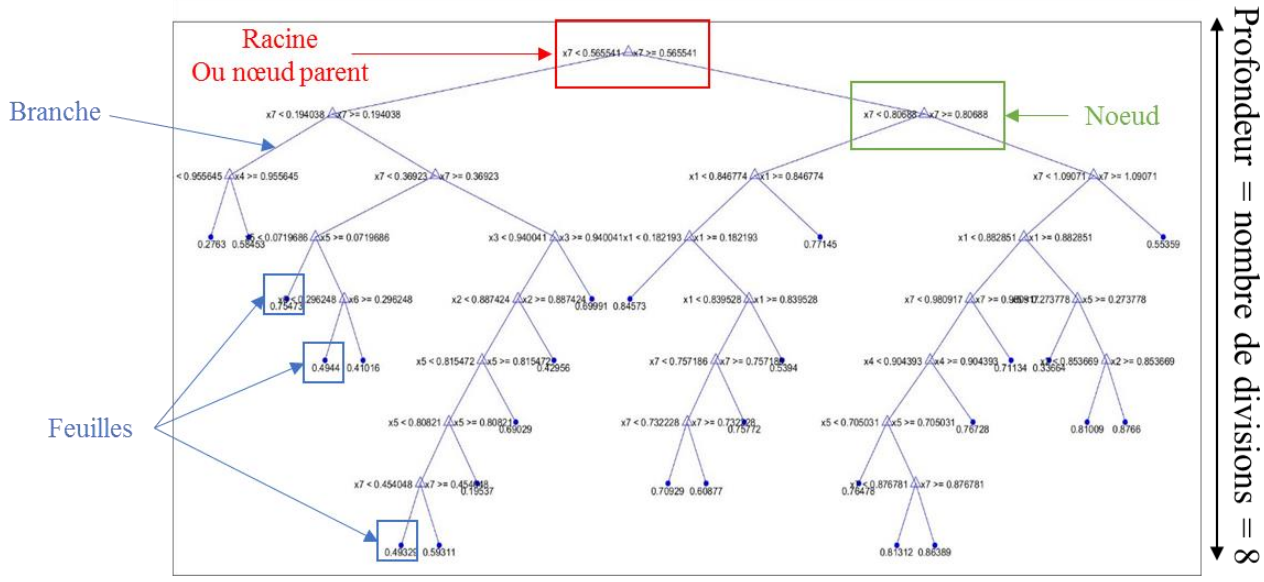


Figure III-11: Exemple d'arbre de régression construit sur des données d'indice de ciel clair, avec le vocabulaire associé aux arbres de régression

- **Les arbres de régression renforcés « boosting » :**

L'arrivée de la méthode de boosting dans le monde de l'apprentissage automatique découle de l'introduction de l'algorithme Adaboost (Freund et Schapire, 1999), un algorithme de classification binaire.

L'idée de base est de combiner les réponses de plusieurs prédicteurs basics (utilisable sur n'importe quel modèle de machine learning) pour obtenir une réponse plus précise obtenant ainsi un prédicteur plus complexe.

Le classifieur de base, le plus souvent utilisé comme prédicteur basique par l'algorithme, est un arbre simple de classification à une seule division.

L'algorithme crée des arbres successifs à partir de différentes pondérations des données. A chaque étape, les données sont classifiées par l'état actuel et ces classifications sont utilisées comme poids pour générer les arbres suivants.

Les données mal classifiées sont pondérées de manière plus grande que les données correctement classifiées,

Les données difficiles à classifier reçoivent des poids toujours plus élevés, cela augmente les chances de les classifier correctement.

Le résultat final de la classification est obtenu par la majorité pondérée de la classification à travers la génération des arbres du modèle.

Cette méthode d'amélioration, d'abord développée pour les arbres de classification, a été adaptée aux arbres de régression. C'est la première méthode qui utilise la technique des ensembles. En effet la famille des modèles à base d'arbres de régression a vu une naître un nouveau sous-groupe de modèles, les modèles dits d'« ensemble learning ». Les arbres dont la réponse améliore la prévision sont pondérés et la prévision finale est obtenue par une combinaison linéaire des arbres (De'ath, 2007).

La formulation mathématique des arbres de régression améliorés par la méthode de boosting est la suivante :

$$\hat{x}(t+h) = \sum_m \beta_m b(\hat{x}(t+h), \gamma_m) \qquad \text{III-22}$$

- $b(\hat{x}(t+h), \gamma_m)$ fonction de base représentant chaque arbre ~~que~~ généré lors de la modélisation ;
- γ_m la variable de division, définie par différentes valeurs à chaque nœud et les résultats de la prévision ;
- β_m est le poids affecté par l'algorithme global à chaque résultat obtenu par les arbres. Pour ce type de modèle il convient de déterminer le nombre d'arbres généré lors de l'apprentissage, l'erreur est compilée pour déterminer le nombre optimal d'arbres à générer.

La Figure III-12 illustre cette optimisation :

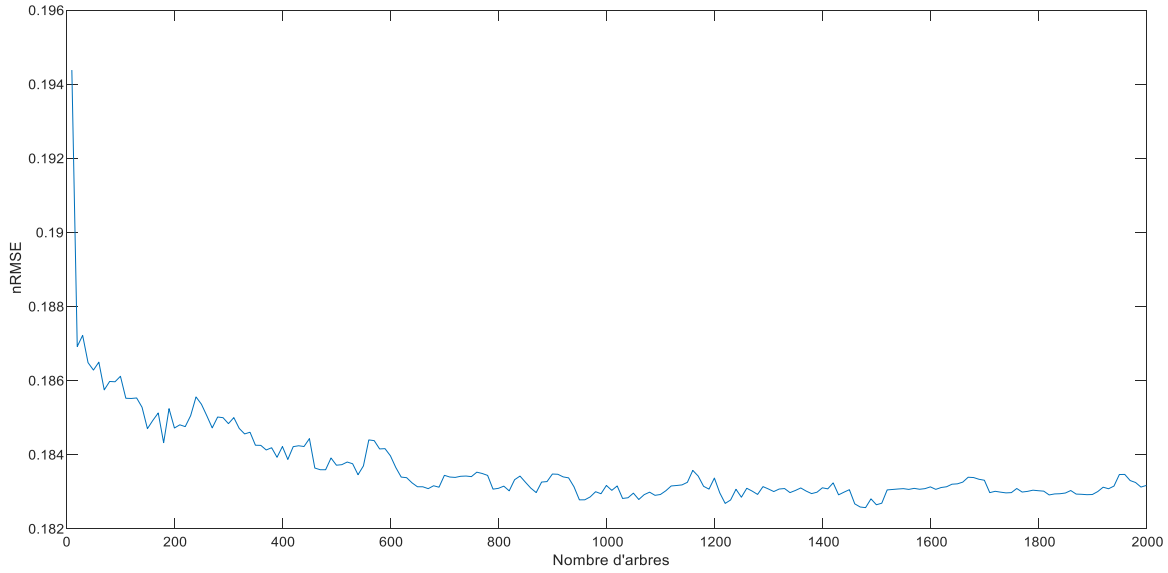


Figure III-12 : nRMSE en fonction du nombre d'arbres générés, cas des arbres de régression renforcés

Par exemple, dans ce cas où il s'agit d'apprentissage sur des données d'indice de ciel clair de rayonnement global ; nous constatons que le nombre d'arbres nous donnant les meilleures performances en termes de nRMSE est de 1480.

- **Les arbres de régression ensachés « bagging » :**

Le terme de bagging est l'acronyme de « bootstrap aggregating ».

Il s'agit d'un autre niveau d'amélioration des modèles de prévision et notamment des arbres de régression.

La description de cette méthode a été réalisée par Breiman (1996). Cette méthode consiste à générer des arbres de régression sur des échantillons du jeu de données, mais contrairement à la méthode de boosting, les arbres successifs ne dépendent pas des réponses des arbres précédents.

En effet, chaque arbre est construit en utilisant un échantillon d'amorce, dit « Bootstrap », du jeu de données. Un Bootstrap est défini comme un nouvel échantillon créé à partir du jeu de données originel. Chaque échantillon est construit par tirage avec remise, c'est ce que l'on appelle le ré échantillonnage.

L'hypothèse gaussienne sur la distribution des échantillons n'est pas nécessaire. Lorsqu'on a généré les arbres et que l'on dispose des réponses correspondantes, un simple vote majoritaire est utilisé pour réaliser la prévision. Cette méthode d'ensemble donne des améliorations significatives dans le traitement de données concernant des phénomènes complexes et dont les relations entrées/sorties sont difficiles à appréhender.

La formule mathématique de cette modélisation est comme suit :

$$\hat{x}(t+h) = av_k \varphi_k\big(\hat{x}(t+h)\big) \qquad \text{III-23}$$

- φ_k correspond à chaque arbre qui est généré à partir d'un échantillon du jeu de données ;
- av_k qui est la moyenne des réponses des k arbres pour réaliser la prévision finale.

Le nombre d'arbres générés k est déterminé par une méthode heuristique, le test est effectué sur un nombre croissant d'arbres. On calcule l'erreur sur la prévision pour déterminer k.

La Figure III-13 représente les résultats sur la sélection du nombre d'arbres.

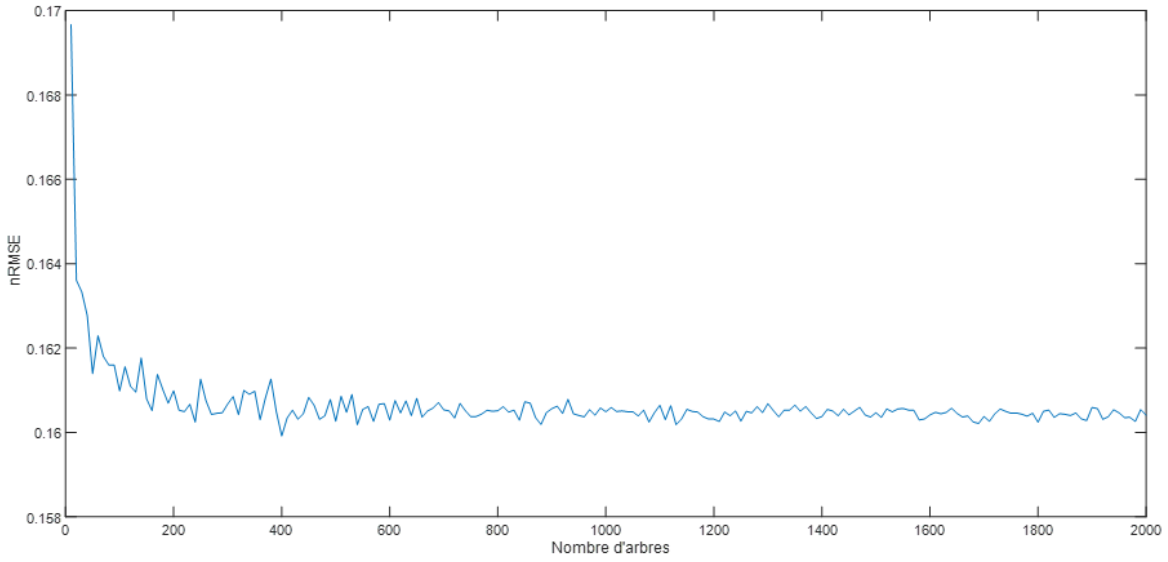


Figure III-13: nRMSE en fonction du nombre d'arbres générés sur des données d'indice de ciel clair de rayonnement global, cas des arbres de régression ensachés

Grâce à ce travail, nous déterminé que le nombre d'arbre optimal est, dans ce cas de figure, de 400. Naturellement cette étape devra être réalisée dès que l'on changera de données d'entrées.

- **Les forêts aléatoires « Random forest » :**

C'est Breiman (2001) qui a proposé une définition des forêts aléatoires, pour la classification et la régression.

Cette méthode d'optimisation des arbres de régression est directement inspirée de la méthode de bagging (Lahouar et Ben Hadj Slama, 2015), il est ajouté un degré de complexité supplémentaire dans l'échantillonnage des données d'apprentissage et la façon de diviser les nœuds.

Dans le cas des forêts aléatoires, un échantillon de données du jeu d'apprentissage complet est pris aléatoirement pour générer chaque arbre de régression, ensuite l'échantillon de données est replacé dans le jeu complet avant de générer un autre arbre. Cette étape étant réalisée à chaque fois, cela implique qu'il est possible de retrouver plusieurs fois les mêmes valeurs dans des échantillons différents.

Lorsque les nœuds sont divisés au sein de chaque arbre, l'algorithme des forêts aléatoires cherche à sélectionner la meilleure façon de diviser les nœuds parmi les différentes possibilités, le nombre de façons de diviser chaque nœud est fixé. La sélection à chaque nœud est réalisée aléatoirement, ce qui implique que, pour p caractéristiques de division à chaque nœud, la probabilité de choix d'une méthode plutôt qu'une autre est de 1/p.

Pour terminer, on réalise la prévision en prenant la moyenne des réponses de tous les arbres De la même manière que pour les méthodes précédentes nous avons déterminé le nombre optimal d'arbres générés au cours de l'apprentissage. La Figure III-14 illustre cette étape.

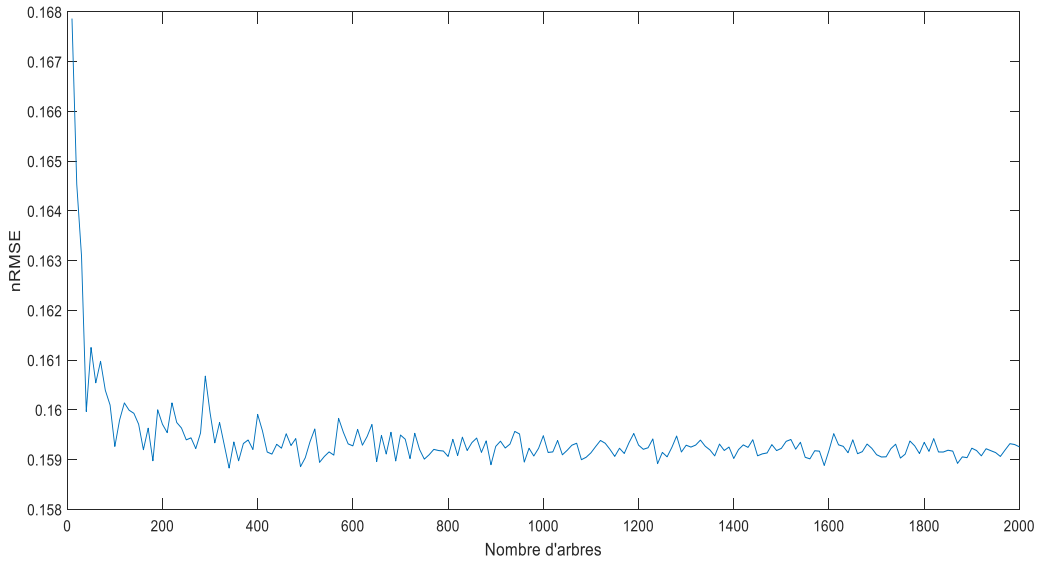


Figure III-14: : nRMSE en fonction du nombre d'arbres générés sur des données d'indice de ciel clair de rayonnement global, cas des forêts aléatoires

De cette manière, il a été possible de déterminer que le nombre optimal d'arbres est de 340 pour la réalisation de l'apprentissage. A partir de ces méthodes, il est possible de générer des quantiles de prévision pour réaliser des intervalles de prévision (Voyant *et al.*, 2018b). Nous avons participé à des travaux sur cette amélioration possible, la démarche sera explicitée dans la section suivante.

La Figure III-15 est une représentation graphique d'un modèle d'arbres de régression ensachés ou de forêts aléatoires. La différence fondamentale entre ces deux modèles ne se trouve pas dans leur architecture (qui est la même) mais dans la façon de réaliser les divisions à chaque nœud. Dans le cas des forêts aléatoires la division du nœud est réalisée sur l'échantillon aléatoire correspondant, alors que

dans la méthode bagging l'échantillon est remis au sein du jeu originel donc la division se fait sur l'exemple du jeu originel.

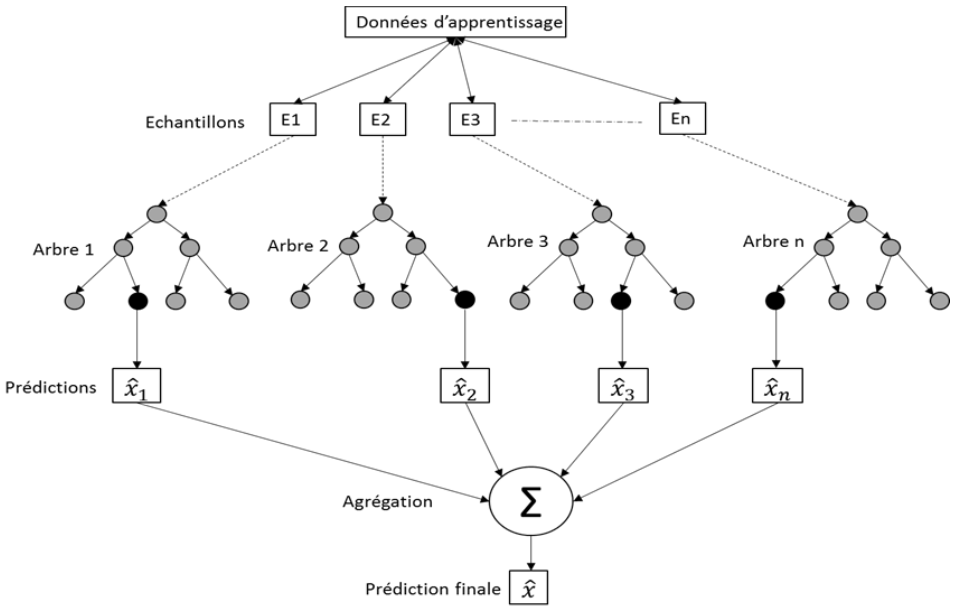


Figure III-15: Exemple général d'architecture des arbres de régression ensachés ou des forêts aléatoires

Le fait de complexifier un peu plus le modèle confère aux forêts aléatoires une grande robustesse, même pour des données très bruitées.

La sélection d'un sous ensemble aléatoire de caractéristiques nodales pour chaque division permet de décorréler les arbres de régression entre eux. Le grand nombre d'arbres générés a pour inconvénient de supprimer le caractère « intelligible » des arbres de régression. En effet, pour les méthodes d'ensemble il n'est plus possible de représenter les arbres.

La famille des modèles, constituée par les arbres de régression, offre une multitude de possibilités et d'outils pour la prévision. Ce type de modèles est de plus en plus utilisé que ce soient dans le prétraitement des données, dans la sélection et dans la prévision. Les différentes déclinaisons de ces modèles les rendent robustes et utilisables dans l'appréhension de problèmes complexes. Ces modèles ont mis du temps à être utilisés mais sont désormais des incontournables des problèmes de régression ou de prévision de phénomènes naturels et notamment en météorologie.

2.3. Synthèse sur les modèles

Tous ces modèles choisis à fins de développement constituent un panel de techniques assez intéressant pour entamer la résolution des problèmes de prévision.

Dans la famille des modèles sans apprentissage et d'une manière générale, il est possible d'affirmer que sont plus simples à mettre en œuvre (persistance et smart persistance), affinés par l'amélioration basée sur les filtres de Kalman couplés à la smart persistance.

Dans la famille des modèles avec apprentissage (qui sont déjà bien plus complexes) il est possible de « classer » les modèles par la complexité de mise en œuvre :

- Les méthodes dites « classiques » sont généralement les moins complexes (ARMA, Réseau de Neurones),
- Les arbres de régression simples et élagués

- Les méthodes basées sur les fonctions Kernel (processus Gaussien et Machine à vecteur de support).

- Les méthodes plus compliquées sont celles dites «d'ensemble». Bien qu'elles soient basées sur le couplage de méthodes simples, elles nécessitent l'utilisation simultanée d'un grand nombre de «petits modèles». Les méthodes de renforcement, d'ensachage et de forêts aléatoires sont basées sur de simples arbres de régression. Cependant les nuances dans leur fonctionnement, que ce soient l'algorithme de renforcement ou l'échantillonnage des données d'apprentissage de plusieurs méthodes, donnent à ces modèles une plus grande complexité de mise en œuvre et moins de lisibilité.

3. Prévision probabiliste

Cette partie concerne les intervalles de prévision, que l'on peut aussi appeler les prévisions probabilistes.

Le but de cette méthodologie est d'apporter des éléments prédictifs supplémentaires à la simple prévision ponctuelle à horizon. Au lieu de ne lui donner qu'une prévision déterministe (valeur discrète), il est possible de donner une probabilité de la prévision. De ce fait, on obtient un intervalle au sein duquel la prévision est la plus susceptible de se trouver. Autrement dit, une prévision probabiliste représente une estimation des probabilités qu'une variable donnée prenne certaines valeurs dans le futur et ainsi permet de déduire un indice de confiance.

En termes de gestion de la production PV, par exemple, l'intérêt est de déterminer des limites hautes et basses de la production possible après un certain laps de temps. Cela donne plus de latitude à l'utilisateur (gestionnaire de réseau) pour gérer ses installations.

Nous avons suivi un protocole pour réaliser cette prévision probabiliste, nous permettant de choisir, de manière objective, le meilleur intervalle. Il est constitué de plusieurs étapes qui sont détaillées dans les sections suivantes.

3.1. Génération des intervalles de prévision

Plusieurs méthodes sont disponibles pour produire une prévision bornée (Voyant *et al.*, 2017b), dans le cadre de notre étude, nous avons choisi d'utiliser la méthode basée sur l'échantillonnage (bootstrap) du jeu de données d'apprentissage (B. Chen *et al.*, 2011; Efron, 1979).

Toutes les méthodes basées sur le bootstrap (Alonso *et al.*, 2002; Bühlmann, 2002) sont construites sans faire d'hypothèses sur les distributions sous-jacentes à partir desquelles nos observations auraient pu être échantillonnées.

Avec ce type de méthodes, les données, elles-mêmes, sont utilisées pour estimer les distributions d'échantillonnage des prévisions, à partir des k sous-ensembles et des k modèles de prévision associés. Ces distributions d'échantillonnage estimées sont alors utilisées pour calculer les intervalles de confiance basés sur une estimation en centiles (Wilcox, 2012).

Dans les statistiques descriptives, un centile est chacune des 99 valeurs qui divisent les données triées en 100 parties égales, de sorte que chaque partie représente 1/100 de l'échantillon de population.

Dans notre étude, cet échantillonnage fait référence à la construction de plusieurs modèles de prévision basés sur les différents sous-ensembles de données d'apprentissage créés. Chaque modèle construit renvoie à une prévision ponctuelle. Les trois principales étapes de génération de ces intervalles sont :

- Génération des k prévisions ;
- Génération de la fonction de distribution des k prévisions, nous faisons ici l'hypothèse que nous disposons d'assez de données pour considérer que la distribution est équivalente aux histogrammes des prévisions (D'Haultfoeuille et Givord, 2014) ;
- Intégration de la fonction de distribution pour obtenir la fonction de répartition (Wilks, 1995) des prévisions et ainsi déduire les intervalles de prévision.

Etant donné que l'on construit plusieurs modèles sur des échantillons différents, on réalise un ensemble de prévisions (9 au total) représentant un ensemble de prévisions.

La Figure III-16 représente schématiquement cette étape.

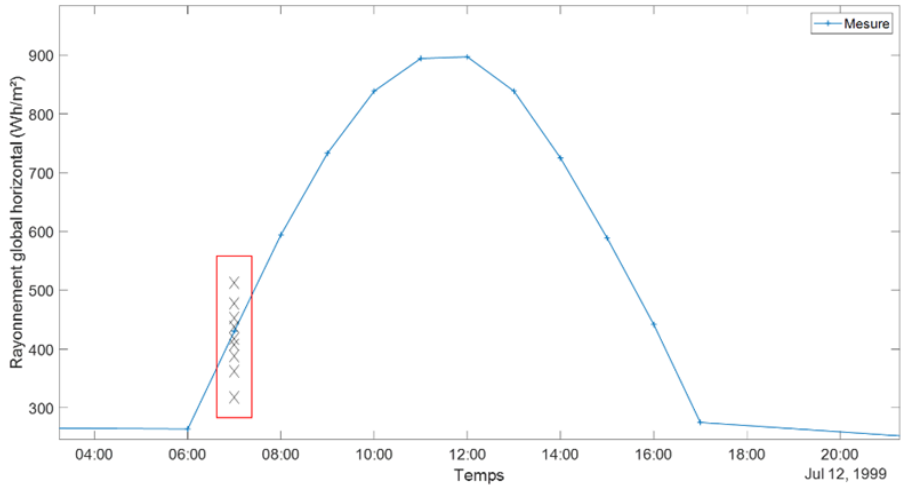


Figure III-16: Exemple d'ensemble de prévisions pour un point de mesure par plusieurs modèles construits sur des échantillons différents. L'encadrement en rouge rassemble les différentes valeurs obtenues

A partir de ces différentes prévisions, une distribution des valeurs prédites est générée chaque fois que la prévision est faite. Cette fonction représente la fréquence des prévisions en fonction du rayonnement global, autrement dit cette fonction est générée et représente toutes les prévisions obtenues.

La Figure III-17 est un exemple de cette fonction de distribution des prévisions. On réalise alors une intégration de cette fonction par rapport au rayonnement pour générer la fonction de répartition après normalisation par la valeur maximale de l'intégration et obtenir ainsi les intervalles de prévisions.

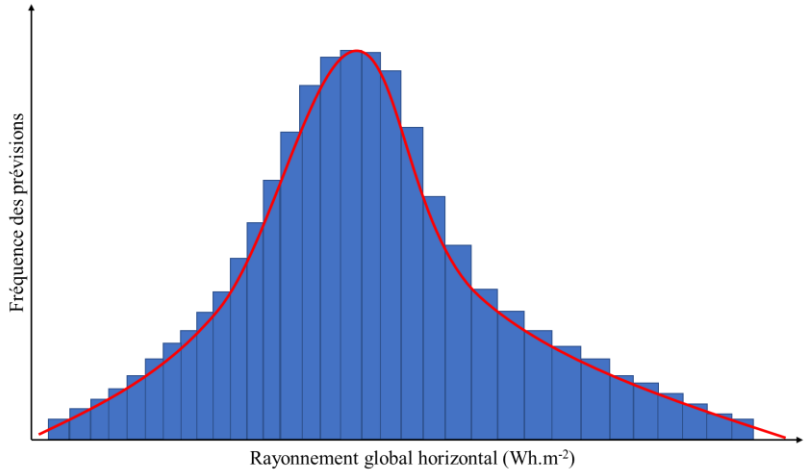


Figure III-17: Exemple de fonction de distribution obtenue pour un point de mesures et k modèles de prévision.

La Figure III-18 représente un exemple de fonction de répartition avec les quantiles associés à l'encadrement.

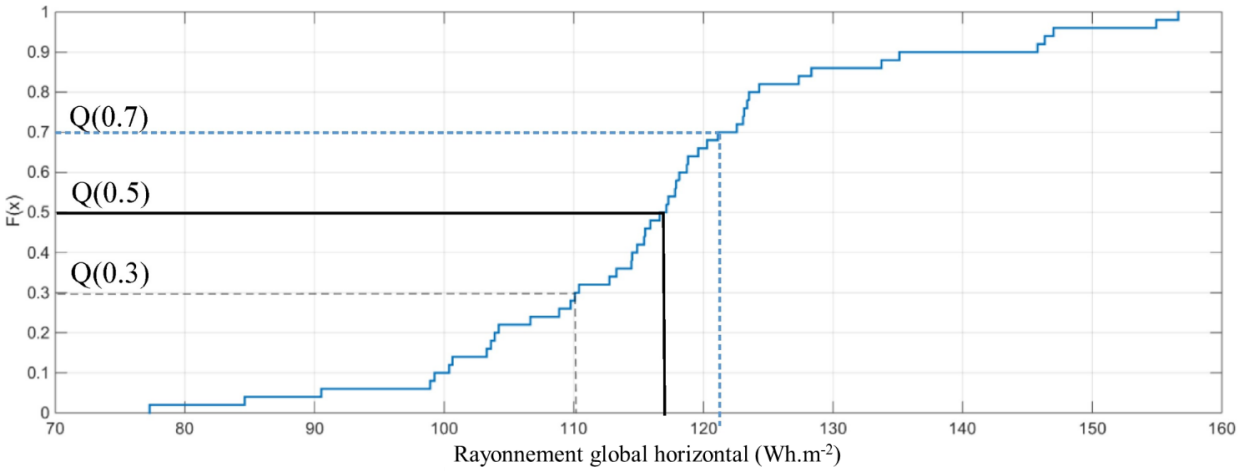


Figure III-18: Exemple de fonction de répartition obtenue par intégration de la fonction de distribution utilisée pour la génération des intervalles de prévision.

Avec cet outil, tous les centiles sont générés afin de calculer les intervalles de prévision. Un centile est une mesure statistique indiquant la valeur en dessous de laquelle un pourcentage donné de prévision tombe.

Par exemple :

- Le 30ème centile Q (0,3) est la valeur en dessous de laquelle 30% des modèles proposent une prévision ayant une valeur inférieure à 110 Wh/m² de la prévision.
- Le 25ème centile Q (0,25) est appelé premier quartile,
- Le 50ème Q (0,5) est la médiane (environ 115 Wh/m² dans notre exemple)
- Le 75ème centile le troisième quartile Q (0,75).

La valeur médiane Q (0,5) de la fonction de répartition est considérée comme une prévision ponctuelle. Les autres quantiles sont utilisés pour définir l'intervalle de prévision ad hoc.

Il est possible de définir alors un triplet d'équations qui représente l'intervalle que l'on a généré.

$$\hat{I}(t+h) = Q(0{,}5)|_{FR(t+h)} \qquad \text{III-24}$$

$$\overline{\hat{I}_n\,(t+h)} = Q(0{,}5 + n.0{,}01)|_{FR(t+h)} \qquad \text{III-25}$$

$$\underline{\hat{I}_n\,(t+h)} = Q(0, -n.0{,}01)|_{FR(t+h)} \qquad \text{III-26}$$

Dans ces équations, $FR(t+h)$ désigne la fonction de répartition correspondant à la prévision à t+h, le trait vertical étant un séparateur dans la notation pour dire que l'on se place dans la fonction de répartition désirée, $\overline{\hat{I}_n\,(t+h)}$ est la valeur haute de l'encadrement de la prévision et $\underline{\hat{I}_n\,(t+h)}$ la valeur basse de ce même encadrement. La fonction de répartition FR peut être définie de la façon suivante :

$$FR(t+h) = FR\{\hat{I}_k(t+h)\} \qquad \text{III-27}$$

Où *k* est le nombre d'échantillons bootstrap que l'on a choisi lorsque l'on a divisé le jeu de données d'apprentissage, et *n* est le paramètre qui définit l'amplitude de l'encadrement.

Naturellement, ces informations ne suffisent pas à fournir un encadrement objectif de la prévision, en fonction de la valeur de *n*. On pourra obtenir soit une bonne sensibilité mais un intervalle trop grand pour être intéressant soit un encadrement objectif très étroit, mais aucune des futures mesures ne sera comprises entre les deux bornes. Il faut donc trouver le bon compromis pour relier le tout.

La Figure III-19 représente les enveloppes de la prévision pour plusieurs intervalles, choisis arbitrairement.

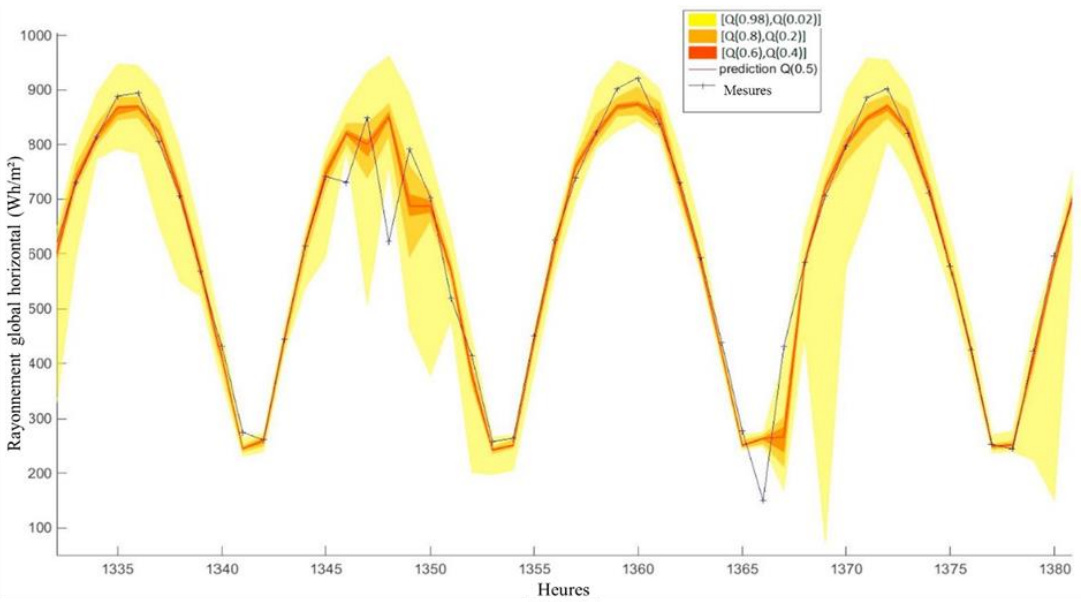


Figure III-19: Exemple d'enveloppes générées pour différents intervalles de prévision plus ou moins grands basés sur les centiles. Prévision de rayonnement global horizontal pour un horizon de 1 heure à Ajaccio.

Cette figure montre l'importance de choisir une taille pertinente d'intervalle de prévision. En effet, si l'on se place du côté de l'utilisateur (gestionnaire de réseau, producteur d'énergie…) il est possible de voir qu'un intervalle trop étendu ne donne pas d'informations exploitables (voir même aberrantes comme vers l'heure 1369) et donc n'apporte que peu d'information sur ce qu'il est susceptible de se passer. A l'inverse un intervalle trop petit est inexploitable d'un point de vue opérationnel. C'est donc pour cela qu'il est nécessaire de continuer cette optimisation de la prévision par l'estimation de la pertinence de l'intervalle. C'est l'objet de la section suivante.

3.2. Pertinence de l'intervalle de prévision

Les incertitudes induites par la prévision du rayonnement solaire peuvent être décomposées en trois parties (Voyant *et al.*, 2017b): la première est liée à la mesure elle-même, la seconde aux caractéristiques intrinsèques de la série temporelle et la dernière à la méthode utilisée pour la prévision.

Ici, la méthode est basée sur une méthodologie d'estimation des incertitudes dues à la méthode de prévision,

Plusieurs auteurs (David *et al.*, 2016; Grantham *et al.*, 2016; Trapero *et al.*, 2015) ont proposé une prévision probabiliste et ils ont exposé l'incertitude associée. Deux types d'approches, assez similaires, non concurrents mais complémentaires, sont proposés pour tracer une enveloppe de confiance autour des prévisions.

Si les méthodes présentées par David *et al.* (2016) et Grantham *et al.* (2016) sont assez similaires à notre approche, les outils que nous utilisons pour choisir objectivement la largeur de l'intervalle sont probablement plus intuitifs que le diagramme de fiabilité (en anglais « reliability diagram »), le classement par histogramme (en anglais « rank histogram »), le score de probabilité classée continue (en anglais « continuous ranked probability score ») et le skill score associé. Les deux outils que nous avons choisi pour notre étude sont, la probabilité de couverture d'intervalle de prévision (en anglais « prediction interval coverage probability : PICP ») et la longueur d'intervalle moyenne (en anglais « mean interval length : MIL »).

Dans le but d'améliorer un peu plus cette approche, nous avons adapté un outil mathématique qui nous permet de quantifier l'efficacité de l'intervalle déterminé en combinant le MIL et le PICP, il s'agit du test Gamma.

3.2.1. Longueur d'intervalle moyenne (MIL) et probabilité de couverture de l'intervalle (PICP)

Nous avons choisi deux outils et nous nous proposons de la décrire dans cette partie.

La longueur d'intervalle moyenne (MIL) définie comme la différence entre les limites haute et basse de l'intervalle de prévision.

La formulation mathématique est la suivante :

$$MIL_n = \overline{\hat{I}_n\,(t+h)} - \underline{\hat{I}_n\,(t+h)} \qquad \text{III-28}$$

La probabilité de couverture de l'intervalle de prévision (PICP), définie par la probabilité que la mesure à t+h se trouve entre les limites hautes et basses de l'intervalle de prévision. Le lecteur intéressé pourra se référer à (Rana *et al.*, 2015).

La définition mathématique est donnée par l'équation III-29 :

$$PICP_n = \frac{100}{N}.count(j)\ avec\ j : \overline{\hat{I}_n\,(t+h)} \leq I(t+h) \leq \underline{\hat{I}_n\,(t+h)} \qquad \text{III-29}$$

Dans cette équation N correspond au nombre de données disponibles, la fonction $count(j)$ retourne le nombre de fois pour laquelle la condition j est remplie. Ce qui donne une probabilité de présence de la mesure dans l'intervalle. Pour obtenir un PICP proche de 100%, c'est-à-dire pour être sûr que la prévision est immanquablement contenue dans le MIL, il faut choisir un MIL grand. Cependant l'information donnée est alors peu pertinente pour l'utilisateur final.

Le but que l'on cherche à atteindre lorsqu'on réalise un intervalle de prévision est de déterminer comment obtenir le meilleur compromis possible entre un PICP élevé et une MIL faible. Lorsqu'on réalise un intervalle de prévision, on cherche à déterminer comment obtenir le meilleur compromis possible entre un PICP élevé et un MIL faible.

Le test Gamma

Une méthodologie appelée test gamma (Voyant *et al.*, 2014) a été développée afin de comparer deux cartes d'irradiation en deux dimensions. Dans le cadre de notre étude, cette méthode est adaptée à la comparaison d'intervalles. Avec les deux paramètres précédents le MIL et le PICP, un facteur gamma est calculé à partir de l'équation d'ellipse suivante :

$$\Gamma_n = \sqrt{\left(\frac{MIL_n}{Tol_{MIL}}\right)^2 + \left(\frac{1-PICP_n}{Tol_{1-PICP}}\right)^2} \qquad \text{III-30}$$

Dans ce cas, n étant le paramètre d'encadrement dans les équations 3.24 et 3.25, Tol_{MIL} et Tol_{1-PICP} sont deux tolérances qui dépendent du problème considéré.

Plus l'indice est élevé, moins l'intervalle de prédiction est efficace. Avec cet indice, on construit un test d'hypothèse statistique. Au début de la procédure, il y a deux hypothèses, l'hypothèse nulle (H_0) et l'hypothèse alternative (H_1) définies par :

- H_0 : "l'intervalle de prédiction est pertinent" si Γ_n<1,
- H_1 : "l'intervalle de prédiction n'est pas pertinent" si Γ_n> 1.

En calculant Γ_n pour les n compris entre 1 et 50, nous proposons une règle simple (ou test) permettant de valider l'intervalle de prédiction et, en particulier, de déterminer la meilleure valeur de n permettant d'avoir un bon compromis entre un MIL faible et un fort PICP.

Ce test permet de délimiter le plan de coordonnées cartésien défini par les deux variables MIL et (100-PICP). Cette limite est une ellipse :

- à l'intérieur de l'ellipse, l'hypothèse H_0 est retenue, il s'agit de la zone pour laquelle l'intervalle de prédiction est pertinent,
- en dehors de l'ellipse, H_1 est conservée, c'est la zone pour laquelle l'intervalle de prédiction n'est pas pertinent.

Dans le cas de la prévision du rayonnement global les valeurs de tolérances ont été fixées $Tol_{MIL} = 0{,}5 < I(t) >$ et $Tol_{1-PICP} = 50\%$, ce qui signifie qu'un bon intervalle propose un MIL inférieur à 50% de la valeur moyenne du rayonnement global et permet également d'obtenir un PICP supérieur à 50%.

Graphiquement il est possible de représenter ce test pour faciliter la compréhension. La Figure III-20 est un exemple de tracé de PICP en fonction du MIL pour déterminer l'intervalle le plus pertinent.

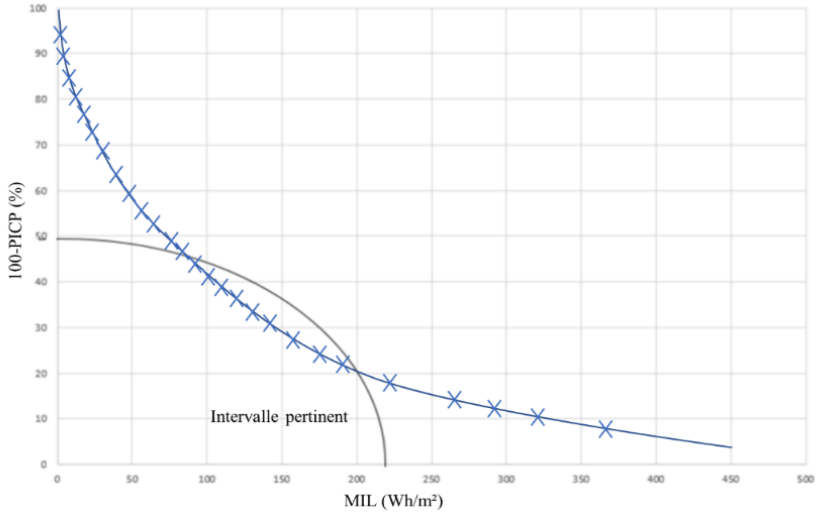


Figure III-20: Schéma de tracé de 100-PICP en fonction du MIL.

Chaque croix correspond à une valeur particulière de n, côté gauche du graphique $n = 1$ et à droite du graphique $n = 50$. Lorsque la courbe rentre dans l'aire de l'ellipse, les intervalles sont pertinents et l'intervalle (donc le n) optimal est celui pour lequel la distance entre l'origine du graphique et le point est minimale. Lorsque l'on sort de l'ellipse, le résultat est amélioré sur la condition 100-PICP (qui doit tendre vers 0) mais on constate que le MIL augmente, ce qui nous fait encore perdre en pertinence. Cette

méthode de détermination de la pertinence de l'intervalle de prévision présente l'avantage d'adapter les conditions sur les deux caractéristiques au problème considéré, cela donne de la flexibilité à ce test et permet de l'appliquer à de nombreux cas.

Une fois que le n est déterminé, il est alors possible de borner la prévision en remplaçant dans les équations III-24 et III-25, il en résulte alors, pour chaque point de prévision, un encadrement optimal et ainsi une bonne vision la grandeur prévue.

Une autre amélioration de ce bornage est effectuée en parallèle par l'utilisation du modèle de ciel clair Solis (cf. Chapitre 2). En effet la mesure du rayonnement global trouve sa borne supérieure dans le rayonnement global horizontal estimé par ciel clair, et sa borne inferieure dans le rayonnement diffus horizontal estimé lui aussi par le modèle ciel clair (le rayonnement diffus horizontal est la seule composante solaire maximale par ciel nuageux et minimale par ciel clair). Ces deux valeurs issues du modèle de connaissances nous permettent d'affiner encore un peu plus les intervalles. En effet si un intervalle optimal donné voit ses bornes en dehors de ces deux limites physiques il est alors possible de réduire cet intervalle ce qui a pour effet de réduire le MIL pour un même PICP. Sur la Figure III-21 nous pouvons voir les effets de cette correction sur un encadrement de prévision.

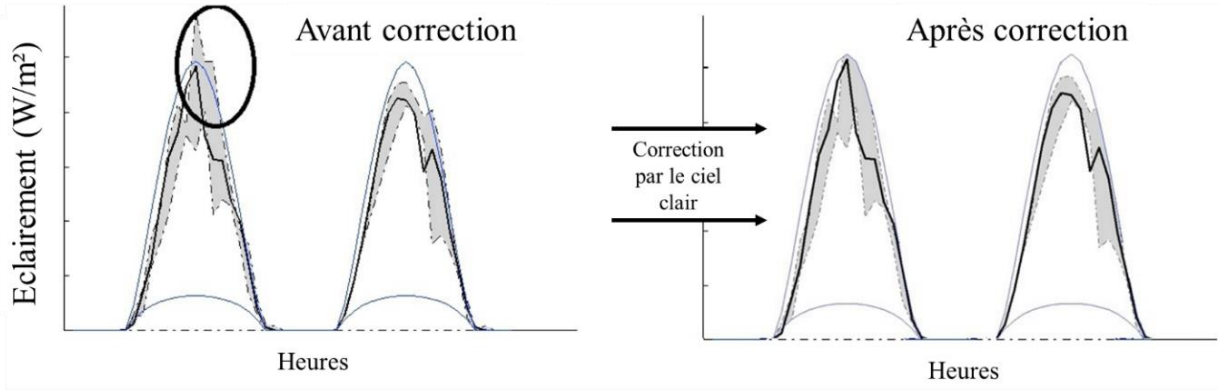


Figure III-21: Correction de l'encadrement de la prévision par l'utilisation du modèle de ciel clair

Cette utilisation est en théorie très intéressante, cependant on retiendra que le modèle de ciel clair contient lui-même une petite incertitude et que, de ce fait, la condition d'égalité du PICP n'est pas toujours respectée. Néanmoins la différence est assez faible (moins de 5%) ce qui n'enlève rien aux avantages apportés par ce bornage supplémentaire.

4. Synthèse

Dans ce chapitre nous avons détaillé les méthodes que nous avons utilisées pour réaliser les prévisions.

Il existe plusieurs catégories de modèles avec différentes utilités en fonction de l'horizon de prévision choisi et l'incertitude que l'on peut admettre. L'état de l'art actuel nous a conduit à nous orienter vers les techniques dites d'apprentissage automatique. Ce type de techniques regroupe une grande variété de modèles pouvant se diviser en deux grandes catégories : les modèles sans apprentissage et les modèles avec apprentissage.

La première famille de modèles, dite sans apprentissage, plus simples à mettre en œuvre, nous serviront de référence lors des expérimentations, il s'agit du modèle de persistance (P) et de persistance intelligente (PI). Nous avons choisi de tester, également, un autre modèle ne nécessitant pas de phase

d'apprentissage, un modèle de prévision basé sur l'utilisation des filtres de Kalman en add-on à la persistance intelligente (PIK).

La seconde famille de modèles, dite avec apprentissage, regroupe de nombreux modèles. Certains d'entre eux sont classiques comme le modèle autorégressif à moyenne mobile (ARMA), le perceptron multicouche (PMC) de la famille des réseaux de neurones artificiels, les processus Gaussiens (PG) et les machines à vecteurs supports pour la régression (MVS).

La famille de modèles moins répandus pour la prévision regroupe les arbres de régression et leurs différentes déclinaisons, simples (ARs), puis élagués (ARel), renforcés (ARr), ensachés (ARen) et enfin les forêts aléatoires (FA).

Ce panel de modèles étendu (12 prédicteurs) permet de comparer leur performance sous différents angles tels que complexité, robustesse, erreur de prévision, etc…

Prédire est important mais connaître la « confiance » que l'on peut accorder à cette prévision est nécessaire en particulier pour le gestionnaire du système qui aura à les utiliser. Nous avons donc développé une méthode de prévisions probabilistes afin de donner un encadrement de la prévision.

IV. Simulations et résultats

1. Introduction

Dans les chapitres précédents, nous avons décrit les principaux outils que nous avons utilisés pour réaliser la prévision des irradiations et des éclairements solaires. Nous avons autant que possible essayé de suivre l'ordre chronologique des différentes étapes de nos travaux pour apporter de la cohérence et de la lisibilité au document. Ainsi, après avoir discuté de la nécessité de la prévision, nous avons passé en revue toute la méthodologie prédictive, qui va de la collecte et du prétraitement des données jusqu'à la mise en évidence des méthodes utilisables dans la cadre de la prévision. Le troisième chapitre était quant à lui tourné vers l'élaboration d'une méthodologie opérationnelle en détaillant les modèles que nous avons mis en œuvre et en proposant une approche prévisionnelle probabiliste.

De manière logique, le présent chapitre a pour but d'exposer les résultats obtenus après avoir appliqué ces modèles sur les données collectées. Il est divisé en quatre sections relatives à des conditions expérimentales différentes…. Les conclusions les plus objectives possibles seront ensuite dressées afin de mettre en lumière les points forts et les points faibles des différents modèles suivant les situations dans lesquelles ils sont utilisés.

La première section de ce chapitre concerne le développement des modèles, décrits dans le chapitre 3, pour la prévision du rayonnement global horizontal pour un horizon de 1 à 6 heures par pas de temps horaire (6h/1h).

Ces modèles seront validés et comparés sur les quatre sites météorologiques (Ajaccio, Tilos, Nancy et Odeillo) puis nous tenterons de relier les résultats obtenus, en particulier, le classement des modèles, à la variabilité des données solaires sur chacun des sites.

Comme nous l'avons vu dans le Chapitre 3 (Partie 3), la prévision probabiliste est de plus en plus demandée par les gestionnaires de réseaux électriques. En effet, outre la valeur prédite de la ressource, ces utilisateurs ont besoin de disposer d'un encadrement de la prévision, en d'autres termes, d'un indice de confiance, semblable à celui donné pour les prévisions météorologiques. Ainsi, dans la seconde section, nous proposerons un type de prévision probabiliste qui nous permettra d'obtenir cet encadrement ; le mode opératoire diffèrera un peu de ce que l'on peut trouver dans la littérature et ainsi posera les bases d'un nouveau mode de prévision à horizon.

La troisième section de ce chapitre traitera de la prévision des deux composantes de l'éclairement solaire, à savoir l'éclairement direct normal et diffus horizontal, dont les effets sont différents à la fois dans les applications thermiques et photovoltaïques. Nous montrerons qu'il est plus difficile de prédire ces composantes du fait de leur plus forte variabilité et de leur plus rapide et profonde variation.

Enfin la quatrième section présentera les travaux et expérimentations menés dans le cadre du projet TILOS. En effet, nous avons dû tester nos modèles en suivant les contraintes du cahier des charges imposé par le gestionnaire de la centrale hybride in situ. Cette étude porte sur des mesures inclinées à 30° (dans le même plan que les panneaux photovoltaïques) et pour des horizons de prévision et des résolutions temporelles particulières : 10min/1min (au début du projet), puis 2h/10min et 2h/15min (à la fin du projet). Ces modèles sont, à ce jour, opérationnels dans le système de gestion de l'énergie (EMS en anglais pour Energy Management System) du système développé à Tilos.

2. Prévision du rayonnement global horizontal (6h/1h)

Comme nous l'avons souligné, la première partie des simulations concerne le développement des différents modèles de prévision. Par ailleurs, cette étape nous a permis d'éprouver notre protocole de préparation et d'utilisation des données pour la réalisation de la prévision.

Dans cette section nous allons détailler les prévisions réalisées à l'aide de tous les modèles que nous avons développés pour les différents sites de mesures. La plus grande partie de ce travail a été de réaliser le développement des modèles pour la prévision de l'irradiation horaire globale pour un horizon de 1 à 6 heures par pas de temps de 1 heure.

2.1. Construction des modèles de prévision

Les modèles de prévision développés et testés se répartissent en trois groupes :

- Les modèles sans apprentissage : la persistance (P), la persistance intelligente (PI) puis améliorée par l'utilisation des filtres de Kalman (PIK) ;
- Les modèles avec apprentissage : autorégression à moyenne mobile, ARMA (de manière plus rigoureuse, pour ce modèle on parlera plutôt de phase d'optimisation que d'apprentissage), les réseaux de neurones artificiels avec le perceptron multi couche (PMC), les processus Gaussiens (PG) et les machines à vecteurs de support appliquées à la régression (MVS) ;
- Les modèles avec apprentissage, basés sur les arbres de régression : arbres de régression simples (ARs), arbres de régression élagués (ARel), arbres de régression renforcés (ARr), arbres de régression ensachés (ARen) et les forets aléatoires (FA).

Le protocole suivi pour la construction des modèles de prévisions est présenté sur la *Figure IV-1*. Les premières étapes jusqu'au prétraitement ont été détaillées dans le chapitre 2 de ce document ; à l'issue de cette phase, les données sont prêtes à être utilisées par les modèles.

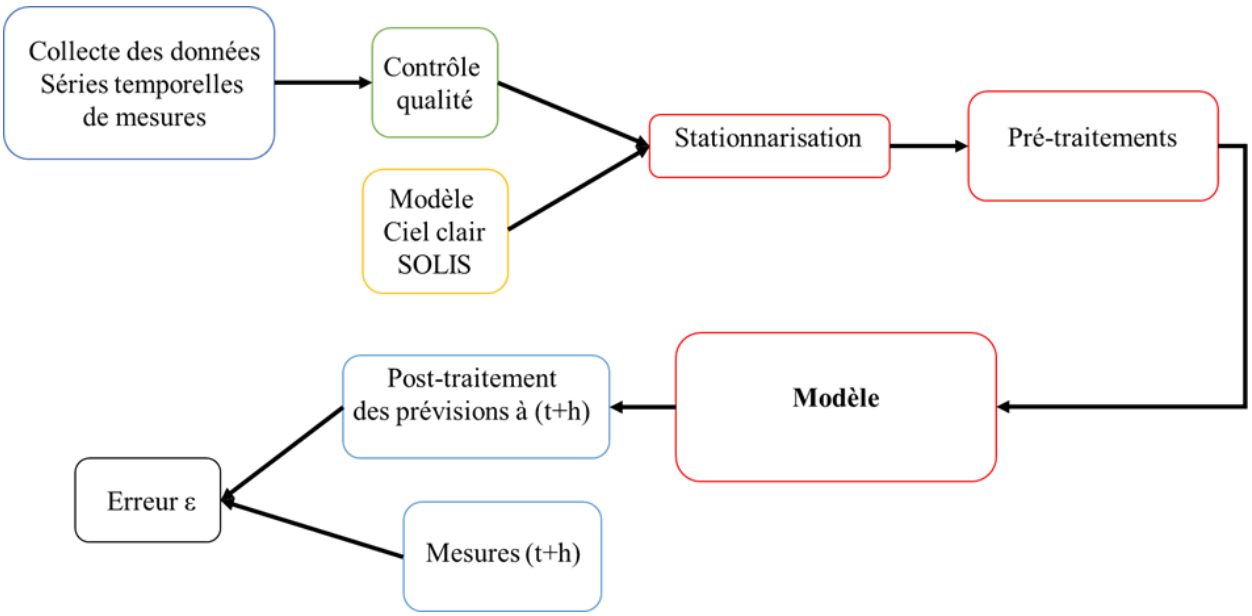


Figure IV-1: Représentation de la méthodologie de construction des modèles de prévision

Les différents modèles ont été présentés en détail dans le chapitre 3.

Au cours de cette étape de construction, nous avons pu constater que les différents modèles ont des complexités différentes d'une part dans leur architecture et d'autre part dans leur mise en œuvre avec des nombres de paramètres intervenant variant de manière significative. Il nous est apparu intéressant de classer les modèles dans le *Tableau IV-1* qui rassemble les différentes caractéristiques inhérentes à leur construction et de tenter d'établir un classement suivant la complexité de mise œuvre. La complexité du modèle prend en compte la préparation des données à lui fournir, le fait qu'elles nécessitent d'être au préalable stationnarisées ou non, le nombre de données antérieures nécessaires (en entrée du modèle déterminée par Information Mutuelle) et le nombre de paramètres à déterminer pour la configuration du modèle lui-même.

Tableau IV-1: Classement des modèles en fonction de leur complexité

Modèles	Stationnarisation des données	Information mutuelle	Paramètres de configuration	Complexité
P	-	-	-	-
PI	oui[1]	-	Ciel clair	+
PIK	oui[1]	-	Ciel clair Fenêtre de mise à jour	+ +
ARs	oui	oui	Ciel clair Nombre de données historiques Coefficients de régression	++
ARMA	oui	oui	Ciel clair Nombre de données historiques Coefficients de régression	+ +
ARel	oui	oui	Ciel clair Nombre de données historiques Tolérance quadratique	+ +
PMC	oui	oui	Ciel clair Nombre de données historiques Algorithme d'apprentissage Poids et biais	+ + +
PG	oui	oui	Ciel clair Nombre de données historiques Fonction de covariance Hyper paramètres (2)	+ + +
MVS	oui	oui	Ciel clair Nombre de données historiques Hyper paramètres (2)	+ + +
ARr	oui	oui	Ciel clair Nombre de données historiques Algorithme de renforcement Nombre d'arbres	+ + +
ARen	oui	oui	Ciel clair Nombre de données historiques Nombre d'arbres Nombre d'échantillons	+ + + +
FA	oui	oui	Ciel clair Nombre de données historiques Nombre d'arbres Méthode d'échantillonnage Nombre d'échantillons	+ + + +

[1] Il ne s'agit pas d'une stationnarisation proprement dite cependant on utilise quand même le modèle de ciel clair pour corriger la persistance

Le modèle le plus simple à mettre en œuvre est le modèle de persistance car il consiste uniquement à considérer que la donnée prédite est égale à la donnée précédente tout en gardant à l'esprit qu'il est très peu performant et ne sert en général que de référence. Vient ensuite le modèle de persistance intelligente simple qui malgré sa simplicité a en général de meilleures performances. Les modèles un peu plus complexes sont la persistance intelligente récursive avec filtre Kalman (dont la complexité réside dans le fait qu'il faut déterminer la fenêtre de mise à jour), les arbres de régression simples, les processus auto régressifs à moyenne mobile et arbres de régression élagués, ces trois modèles ont sensiblement le même nombre de paramètres. Les modèles encore plus complexes et très présents dans la littérature sont le Perceptron Multi Couches, les processus Gaussiens, les Machines à Vecteurs supports et les arbres de régression élagués et renforcés. Enfin, les modèles les plus complexes sont basés sur les techniques d'ensemble et l'échantillonnage des données, comme les arbres de régression ensachés et les forets aléatoires ; ils nécessitent un plus grand nombre de paramètres pour être mis en œuvre et l'échantillonnage des données peut s'avérer complexe, rendant ces modèles les plus gourmands en ressource de calculs.

Cette phase de construction a été particulièrement importante et nous a permis de nous familiariser avec le fonctionnement et l'utilisation des modèles basés sur l'apprentissage automatique. Par la suite nous allons confronter ces différents modèles sur les séries de données que nous avons à notre disposition.

La prochaine section est consacrée à l'utilisation des modèles sur les 4 jeux de données que l'on a à disposition, pour ainsi tenter de dégager des règles de sélection de modèles liées à l'étude des caractéristiques intrinsèques des séries temporelles étudiées.

2.2.Comparaison des modèles

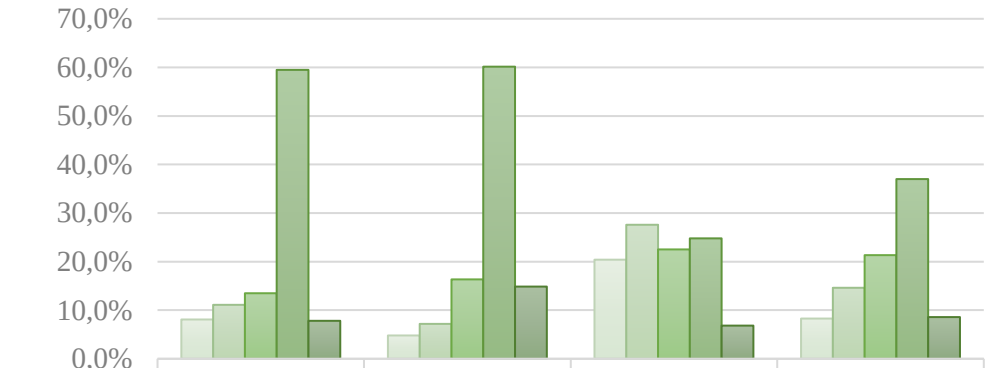
Le but de ce paragraphe est de comparer les différents modèles que nous avons choisis en réalisant les prévisions pour des jeux de données en provenance des quatre sites de mesures. Comme nous l'avons vu dans le Chapitre 2, ces sites ont des caractéristiques météorologiques très différentes. Pour prendre en compte cette différence, nous avons voulu quantifier la variabilité de la série temporelle pour chaque site ; une telle étude a été réalisée par (Voyant *et al.*, 2015) qui ont comparé 20 paramètres pour qualifier la variabilité de différents jeux de données sur plusieurs stations réparties sur le globe (St Pierre à La Réunion, Raizet en Guadeloupe, Ajaccio en Corse, Marseille, Nice et Montpellier en France continentale et Melbourne en Australie) et ont conclu que le paramètre le plus pertinent était le « mean absolute log return » (*MALR*) , qui a été détaillé dans le Chapitre 2. Pour rappel, le *Tableau IV-2* consigne les résultats de ces calculs (visibles sur la Figure II-17 du Chapitre 2). Pour mieux estimer si la variabilité du site est forte ou faible, nous avons calculé le *MALR* pour une série temporelle de valeurs constantes et de valeurs générées aléatoirement (suivant une loi de probabilité uniforme continue ; bruit blanc). La valeur obtenue pour la série aléatoire peut paraître élevée par rapport à celle d'Odeillo, mais il faut relativiser car même dans le cas d'une station météorologique pour laquelle la variabilité du rayonnement solaire serait très importante, il existe néanmoins des tendances indéniables entre les données mesurées, tendance totalement inhibée dans le cas de la série totalement aléatoire. Sur ce tableau, nous voyons que le *MALR* est compris entre 0 (série constante ; $MALR_{min}$) et 2,1352 (bruit blanc ; $MALR_{max}$), il est alors possible de proposer une estimation de la variabilité (*var*) en % à partir de $MALR_{min}$ et $MALR_{max}$ en utilisant $var = 100.\frac{MALR}{MALR_{max} - MALR_{min}}$.

Tableau IV-2: Résultats des calculs de variabilité sur les différents jeux de données

Série de données	*Constante* $MALR_{min}$	*Ajaccio*	*Tilos*	*Nancy*	*Odeillo*	*Aléatoire* $MALR_{max}$
MALR (s.u.)	0	0,1961	0,2348	0,3615	0,5028	2,1352
var	0%	9,2%	11,0%	16,9%	23,5%	100%

Ajaccio et Tilos, deux villes situées en zone Méditerranéenne à fort potentiel solaire présentent, comme nous pouvions nous y attendre une faible variabilité alors que Nancy, soumis à un climat continental a une variabilité que l'on peut qualifier de moyenne. Enfin, Odeillo, du fait de sa situation en zone montagneuse est sujet à de fréquents passages de brume et d'orages même en période estivale induisant ainsi une forte variabilité de l'ensoleillement.

De plus, pour compléter notre raisonnement, et selon la suggestion de P. Lauret, 2017 (Communication personnelle, 05/05/2017), nous avons déterminé la répartition des différentes classes d'indice de ciel clair sur les 4 jeux de données. Il s'agit de calculer les effectifs de données pour plusieurs intervalles d'indices de ciel clair ; nous avons choisi de distinguer cinq classes, de 0 à 0,25, de 0,25 à 0,5, de 0,5 à 0,75, de 0,75 à 1 et supérieur à 1 (données pour lesquelles l'irradiation solaire mesurée dépasse celle calculée par ciel clair). La *Figure IV-2* illustre cette répartition par classes.



	Ajaccio	Tilos	Nancy	Odeillo
[0;0,25[	8,1%	4,8%	20,4%	8,3%
[0,25;0,5[	11,1%	7,2%	27,6%	14,6%
[0,5;0,75[	13,5%	16,3%	22,5%	21,3%
[0,75;1[	59,5%	60,1%	24,8%	37,0%
>1	7,8%	14,9%	6,8%	8,6%

Figure IV-2: Répartition des indices de ciel clair pour les différents jeux de données

Si toutes les journées étaient gouvernées par une nébulosité très faible (ciel clair) ou très forte (ciel couvert), nous obtiendrions un pourcentage élevé de données dans la dernière ou première classe, et donc serait représentatif d'un climat à faible variabilité. A l'opposé, un climat très variable verrait une répartition de l'indice de ciel clair plus homogène dans les différentes classes. Il est donc possible de relier le classement par proportion des indices de ciel clair avec la variabilité des données. Cette interprétation du lien entre la variabilité et répartition des indices de clarté est cependant à prendre avec précaution car on note, par exemple, que Nancy a une répartition plus homogène qu'Odeillo ce qui laisserait penser que sa variabilité est plus élevée que celle d'Odeillo alors que ce n'est pas le cas. Il conviendrait de regarder plus précisément la saisonnalité de ces répartitions, leur occurrence et l'influence de la taille des intervalles. L'homogénéité de la répartition ne permet à elle seule d'avoir une idée de la variabilité car il est nécessaire d'y adjoindre une étude d'occurrence nuageuse ; par exemple si chaque classe est représentative d'une saison la variabilité est très faible et il sera très facile de prévoir.

On note que pour Ajaccio et Tilos, une classe est prépondérante [0,75 ;1[regroupant pour les deux sites environ 60% des données ; Tilos présente une répartition dans les quatre autres classes moins homogènes qu'Ajaccio ce qui peut justifier le fait que la variabilité au sens du MALR est plus élevée que pour Ajaccio. Nancy et Odeillo ont une répartition des indices de ciel clair beaucoup plus homogène, en particulier Nancy qui voit une proportion équidistribuée de données dans les 4 principales classes. On constate donc que l'on peut rattacher, d'une manière plus ou moins évidente, la variabilité au sens du MALR et la répartition par classes d'indice de ciel clair.

Pour les 4 sites de mesures nous avons réalisé des prévisions de rayonnement global horizontal au pas de temps horaire et pour des horizons de 1 heure à 6 heures. Nous allons dans un premier temps présenter les résultats pour chaque site séparément en dégageant les meilleurs modèles en fonction des incertitudes obtenues, puis dans un second temps nous comparerons les résultats des différents sites entre eux.

Nous présenterons dans des tableaux les différents paramètres d'estimation de l'erreur de prévision relatifs et absolus (MAE, nMAE, RMSE, nRMSE) et graphiquement la MBE qui nous permettra non pas de juger de la précision des modèles mais de déterminer si globalement ces modèles prédictifs surestiment ou sous-estiment la réalité expérimentale et permet de valider la construction des modèles lorsqu'elle est proche de 0 (moins de 5%). Contrairement à la MAE (basée sur la norme 1), la RMSE (basée sur la norme Euclidienne) est plus sensible aux erreurs importantes de prévision (et pondère plus les grand écarts) et est reconnue pour être plus représentative de la qualité d'un modèle (*COST action ES1002 Weather intelligence for Renewable Energies (WIRE)*, 2012) ; elle est utilisée surtout lorsque les petites erreurs sont tolérables et que les grandes erreurs ont des conséquences néfastes. Les définitions des différentes incertitudes ont été présentées dans le Chapitre 2 de ce document.

Pour une meilleure lisibilité du document, nous comparerons ensuite les modèles entre eux et en particulier les performances en termes de nRMSE des modèles en fonction des sites. Ces modèles seront également comparés entre eux en utilisant le score d'aptitude du modèle plus connue sous le terme anglais « skill score » ou « forecast score » ; ce paramètre permet d'estimer l'amélioration (ou non) apportée par l'utilisation d'un modèle de prédiction (en général plus complexe) par rapport à un modèle de référence (souvent naïf et simple de mise en œuvre) ; ce modèle de référence est souvent la persistance simple, mais il nous a semblé plus pertinent de lui substituer la persistance intelligente.

2.2.1. Application aux données d'Ajaccio

Le calcul du nombre de données historiques, à utiliser en entrées des modèles, réalisé à partir de l'auto Information Mutuelle a donné les résultats suivants pour le site d'Ajaccio (*Tableau IV-3*) :

Tableau IV-3: Nombre d'entrées (i) utilisées dans les modèles d'apprentissage automatique pour Ajaccio (6h/1h)

Horizon (h)	*1*	*2*	*3*	*4*	*5*	*6*
Nombre d'entrées (i)	6	6	5	5	5	7

Il est difficile de tirer des conclusions du *Tableau IV-3* mais on note toutefois que le nombre de données d'entrée est relativement élevé. En effet, il faut au moins les 5 valeurs précédentes de l'irradiation pour prévoir ce que sera l'irradiation dans le futur. Mathématiquement, cela signifie que pour prévoir à l'instant t l'irradiation à l'instant t+h, (h horizon de prédiction), nous avons besoin de connaître la mesure de l'irradiation à l'instant t ainsi que les (i-1) mesures précédentes (t-(i-1), t-(i-2), t-(i-3), …t-1.

Les résultats des prévisions sont consignés dans le

Tableau IV-4, les trois meilleures valeurs sont surlignées, en vert pour la première, en jaune pour la seconde et en rouge pour la troisième.

Tableau IV-4: Erreurs de prévision en fonction de l'horizon pour Ajaccio

		h+1	*h+2*	*h+3*	*h+4*	*h+5*	*h+6*
nRMSE (%)	P	26,49	42,55	54,03	61,08	64,62	64,03
	PI	19,14	26,42	31,13	34,21	37,00	39,08
	PIK	18,94	25,68	30,93	33,61	36,00	38,08
	ARMA	18,25	29,16	31,20	32,37	32,98	33,72
	PMC	18,29	29,48	31,22	32,38	33,16	33,91
	PG	18,93	30,47	31,86	32,96	33,78	34,48
	MVS	18,51	27,69	31,89	32,09	32,50	33,51
	ARs	24,26	36,50	38,63	39,20	40,23	40,79
	ARel	18,58	30,84	32,23	33,40	34,34	34,48
	ARr	18,63	29,67	31,56	32,73	33,43	34,33
	ARen	18,47	29,98	31,54	32,49	33,44	33,85
	FA	18,96	30,14	31,35	32,47	33,18	33,92
RMSE (Wh.m^{-2})	P	113,57	182,43	231,68	261,90	277,05	274,56
	PI	82,05	113,30	133,49	146,69	158,65	167,58
	PIK	81,19	110,12	132,63	144,12	154,36	163,29
	ARMA	78,26	125,03	133,77	138,81	141,40	144,60
	PMC	78,42	126,38	133,87	138,84	142,19	145,41
	PG	81,18	130,64	136,60	141,32	144,85	147,82
	MVS	79,37	118,73	136,73	137,59	139,35	143,68
	ARs	104,00	156,52	165,65	168,07	172,48	174,88
	ARel	79,66	132,23	138,19	143,19	147,24	147,85
	ARr	79,90	127,23	135,33	140,33	143,35	147,21
	ARen	79,20	128,55	135,25	139,33	143,36	145,16
	FA	81,28	129,23	134,41	139,22	142,28	145,45
nMAE (%)	P	24,44	41,13	51,33	58,38	60,57	59,78
	PI	12,83	18,60	22,45	24,93	27,23	28,90
	PIK	12,37	16,97	21,75	24,24	26,76	28,43
	ARMA	14,12	20,63	22,03	22,80	23,29	23,79
	PMC	14,13	20,96	22,24	22,98	23,53	23,99
	PG	14,49	22,65	23,60	24,42	24,95	25,42
	MVS	12,75	29,45	30,75	31,65	31,52	31,42
	ARs	16,94	27,22	29,41	29,98	30,87	31,27
	ARel	14,19	22,15	23,25	24,33	25,05	25,22
	ARr	14,15	21,59	22,81	23,39	24,09	24,52
	ARen	14,21	21,79	23,05	23,72	24,32	24,72
	FA	14,30	22,05	22,93	23,84	24,38	25,01
MAE (Wh.m^{-2})	P	104,78	176,34	220,10	250,32	259,72	256,30
	PI	55,02	79,77	96,26	106,91	116,76	123,90
	PIK	53,02	72,77	93,26	103,91	114,76	121,90
	ARMA	60,54	88,45	94,47	97,77	99,84	102,01
	PMC	60,59	89,87	95,37	98,54	100,91	102,85
	PG	62,13	97,14	101,17	104,69	106,98	108,99
	MVS	54,67	126,29	131,86	135,72	135,16	134,70
	ARs	72,62	116,73	126,09	128,54	132,35	134,08
	ARel	60,85	94,96	99,70	104,33	107,41	108,14
	ARr	60,69	92,55	97,82	100,31	103,28	105,13
	ARen	60,94	93,43	98,84	101,70	104,26	106,01
	FA	61,31	94,56	98,33	102,22	104,54	107,22

On constate que les écarts entre les estimations de l'erreur de prévision, MAE entre elles ou RMSE entre elles sont faibles et que par conséquent le classement que nous avons réalisé n'est pas vraiment représentatif. Quand les modèles prédictifs donnent des résultats très proches, le choix du prédicteur peut se faire sur la base de la complexité de la mise en œuvre et du temps de calcul.

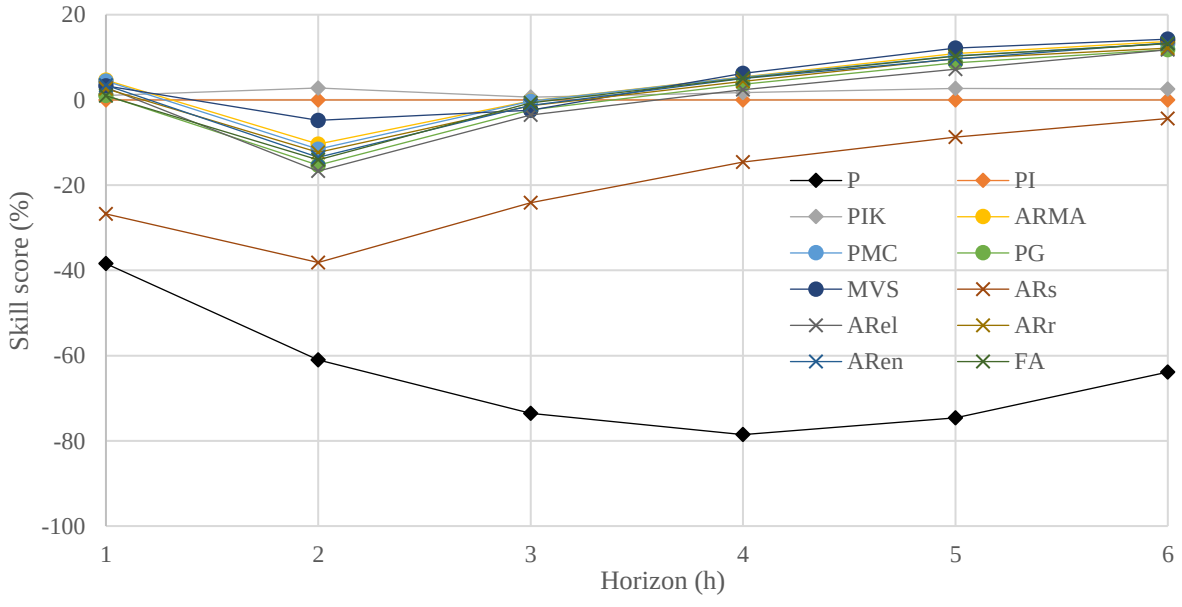
Si on compare le classement obtenu sur la base de la nMAE et celui réalisé sur la base de la nRMSE, on constate quelques différences, mais encore une fois l'écart entre les valeurs obtenues est si faible qu'il est difficile d'en tirer une quelconque conclusion.

Les résultats révèlent que pour les données d'Ajaccio, les deux modèles qui présentent globalement les meilleures performances sont les PMC et ARMA aussi bien d'un point de vue de la nMAE que de la nRMSE. Pour les troisième modèle, le choix est plus délicat avec des différences notées selon l'incertitude utilisée et toujours un faible écart au niveau du résultat.

L'écart entre le modèle le plus performant, qui est le modèle ARMA, et la PI (pris comme référence) varie de 0,9% pour h+1 à 6% pour h+6, cet écart est faible pour justifier l'utilisation d'un modèle très complexe. L'écart entre le « meilleur » modèle et le « plus mauvais », qui est le modèle ARs, varie quant à lui entre 7% et 11%, selon l'horizon, confirmant le faible écart de performance.

Les deux modèles les moins performants sont la persistance simple et les arbres de régression simples. La persistance simple connue pour donner de mauvais résultats est un estimateur naïf et ne sert en général que de modèle de référence pour une comparaison avec les autres prédicteurs. Le modèle basé sur les arbres de régression simples donne de mauvais résultats sans doute à cause du fait qu'il possède un nombre très important de ramifications, cet effet néfaste est bien connu et il est possible d'y remédier en donnant une contrainte sur la croissance de l'arbre (c'est l'élagage).

Nous avons en plus de ces incertitudes, utilisé le skill score qui est de plus en plus utilisé dans le domaine de la prévision et qui permet de comparer les performances des modèles par rapport à un modèle de référence (détaillé dans le Chapitre 2). Les résultats négatifs signifient que le modèle est moins bon que la référence et les résultats positifs, qu'il est meilleur. La *Figure IV-3* représente le skill score en prenant comme référence la persistance intelligente (PI). Etant donné que les courbes ont tendance à se confondre pour la plupart des modèles, nous avons représentés nos résultats sur deux figures la seconde étant un zoom sur la partie la plus intéressante pour un skill score compris entre -20% et +15%.



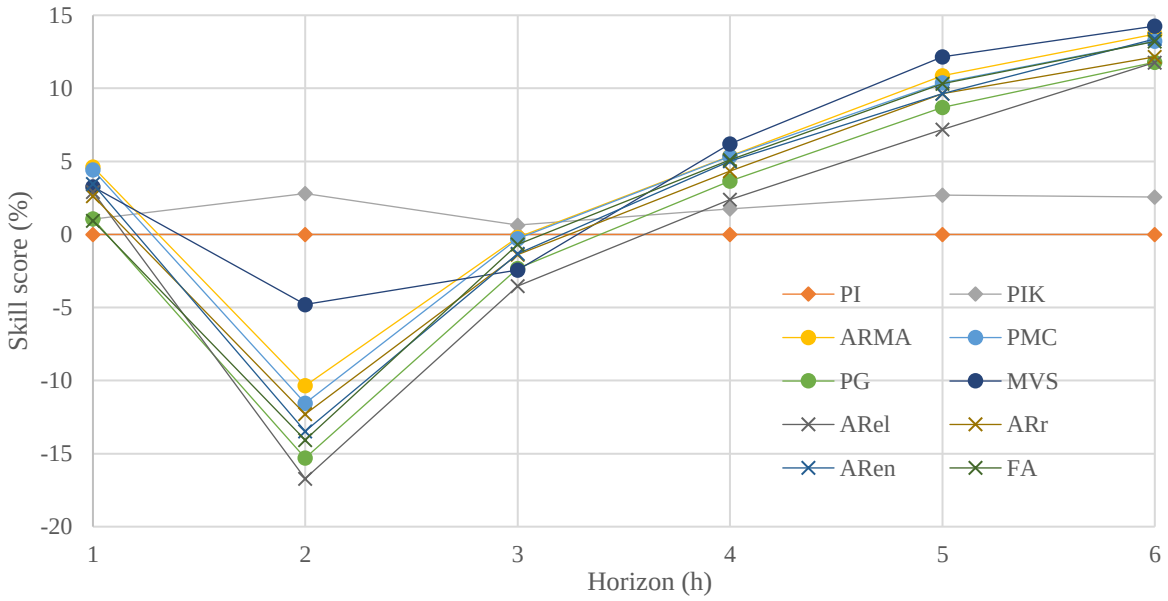


Figure IV-3: Skill score en fonction de l'horizon, en haut tous les modèles, en bas zoom sur les modèles les plus intéressants pour les données d'Ajaccio

On retrouve alors les deux modèles les moins performants, qui sont les arbres de régression simples et la persistance simple bien en dessous de la référence (PI). Tous les autres modèles ont un comportement semblable, ils sont meilleurs que la PI pour un horizon de 1 heure, moins bons pour un horizon de 2 heures et redeviennent meilleurs pour des horizons supérieurs à 3 heures. Le modèle de PI améliorée par les filtres de Kalman (PIK) présente toujours de meilleures performances que la PI simple. Ce type de représentation ne peut pas donner en détail les performances des modèles mais, il permet d'avoir une vision assez rapide des performances des modèles les uns par rapports aux autres et par rapport à une référence.

Nous avons représenté de façon graphique la MBE en fonction de l'horizon sur la *Figure IV-4*.

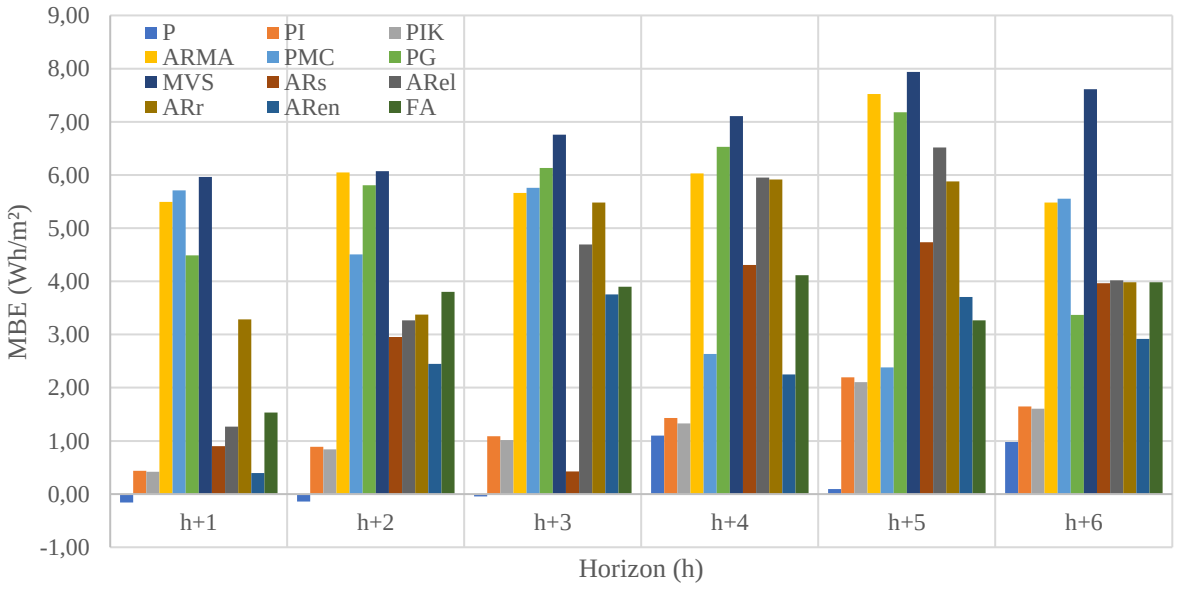


Figure IV-4: MBE en fonction de l'horizon pour les prévisions sur les données d'Ajaccio

La représentation de la MBE sous forme d'histogrammes nous permet de juger de la surestimation ou de la sous-estimation des modèles. En général, si le modèle est bien construit, sa MBE doit être très

proche de zéro. On constate que quel que soit l'horizon, tous les modèles présentent une MBE inférieure à 8 $Wh.m^{-2}$ soit environ 1,8%.

Tous les modèles ont tendance à surestimer les données. Les modèles basés sur la persistance quant à eux sont ceux qui présentent souvent le biais le moins élevé, cela est dû à leur nature même car ils sont construits avec les données elles-mêmes ils n'induisent donc pas de biais dans les prévisions.

Le

Tableau IV-4 nous a permis de déterminer le modèle le plus performant. Même s'il n'apparait pas de modèle qui soit « le plus performant » pour tous les horizons, notre choix du « meilleur » a été basé sur l'analyse des performances globales de chaque prédicteur ; c'est le modèle ARMA qui présente les meilleurs résultats pour Ajaccio. Pour illustrer les performances de ce meilleur modèle, il nous semble intéressant de représenter les données prévues en fonction des données mesurées pour trois horizons h+1, h+3 et h+6 sur la *Figure IV-5*.

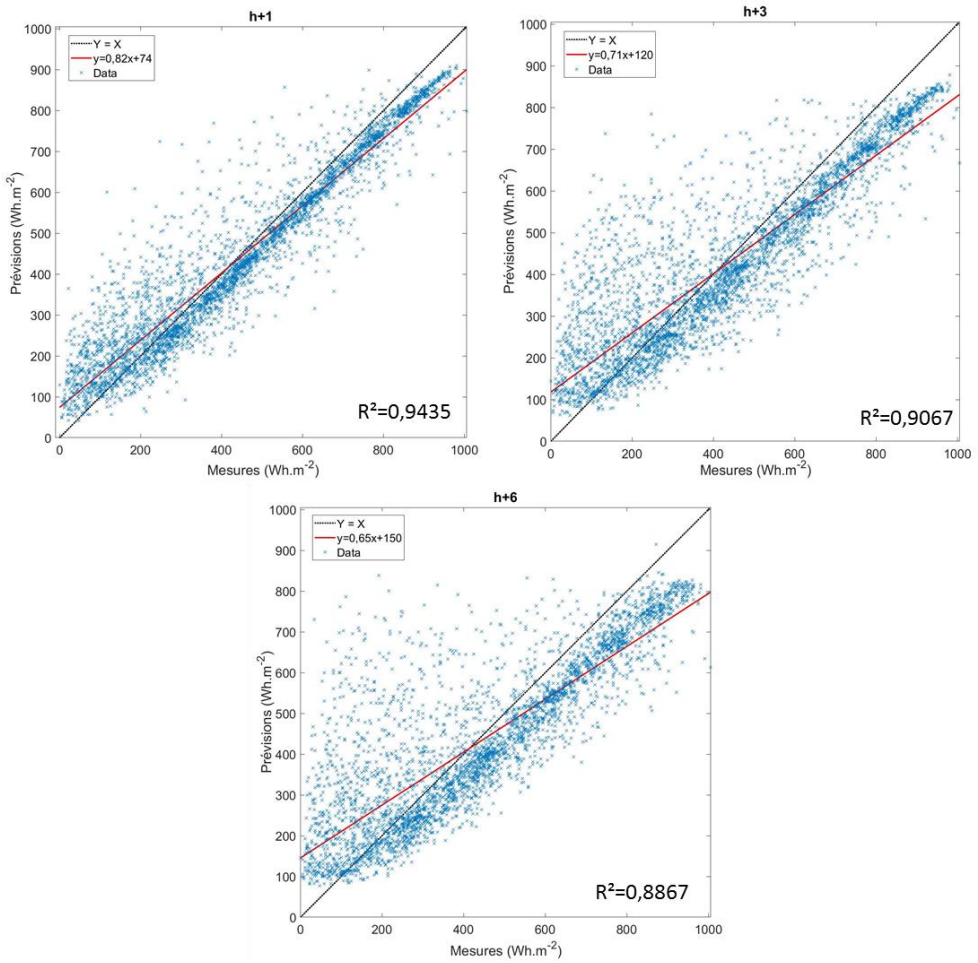


Figure IV-5: Représentation des valeurs prévues en fonction des valeurs mesurées pour trois horizons temporels pour le modèle ARMA

Nous avons présenté sur la *Figure IV-5* la droite Y=X en noir (prédicteur idéal) et la droite rouge qui est la droite de régression linéaire des points dont les équations ont été reportées sur les figures. On note que la dispersion des points augmente avec l'horizon temporel ce qui paraît logique ; cette constatation est également visible avec une augmentation de l'angle entre les droites noire et rouge avec l'horizon. Même s'il apparait des points dont la valeur prédite est très éloignée de celle mesurée, on constate cependant que la plus grande partie du nuage se situe aux alentours de la droite noire.

En conclusion de ce paragraphe sur les données d'Ajaccio, ville présentant, comme vu dans le Chapitre 2, la plus faible variabilité (*var* = 9,2%), on note :

- Aucun modèle ne se distingue réellement des autres et les incertitudes sont toujours très proches quel que soit le modèle choisi ;
- Deux modèles, en revanche, ne sont pas être adaptés, la persistance simple (comme nous nous y attendions) et les arbres de régression simples ;
- Bien que les performances soient proches, le modèle ARMA est le plus adapté pour ce site de mesure à faible variabilité, suivi des PMC ;
- Le modèle de PI présente des performances assez intéressantes et a l'avantage d'être simple à mettre en œuvre.

2.2.2. Application aux données de Tilos

Nous avons ensuite réalisé les prévisions sur les données en provenance de l'île de Tilos et obtenues dans le cadre du projet Horizon 2020, ces données ont une variabilité légèrement plus élevée que celles d'Ajaccio (11,0% vs 9,2%).

Tableau IV-5: nombre d'entrées (i) utilisées dans les modèles d'apprentissage automatique pour Tilos (6h/1h)

Horizon	*1*	*2*	*3*	*4*	*5*	*6*
Nombre d'entrées(i)	5	5	4	5	6	5

Comme précédemment, il est difficile de relier le nombre d'entrées à l'horizon temporel de prévision et à la variabilité des données ; on ne note pas de différence importante avec les valeurs obtenues pour le site d'Ajaccio.

Les résultats des simulations sont présentés dans le *Tableau IV-6*, les trois meilleures valeurs ont été surlignées, en vert pour la première, en jaune pour la seconde et en rouge pour la troisième.

Trois modèles se détachent de ces résultats, il s'agit des modèles processus gaussien (PG), arbres de régression ensachés (ARen) et forets aléatoires (FA).

Contrairement aux résultats obtenus pour Ajaccio, ce classement est valable quel que soit la métrique utilisée même si le rang peut varier entre la 1ère et la 3ème place. On note également que les différences de performances entre ces trois premiers modèles sont très faibles (moins de 1% d'écart en termes de nRMSE à h+6), il est donc difficile de donner un ordre de préférence entre ces trois modèles même si ARen est celui qui se classe le plus souvent en pole position et minimise l'erreur de prévision.

L'écart entre le meilleur modèle PG et la PI a augmenté par rapport à celui obtenu avec les données d'Ajaccio puisqu'il passe de 4% à h+1 (contre 0,9% pour Ajaccio) à 14% à h+6 (contre 6% pour Ajaccio). De même, on retrouve cette légère augmentation pour l'écart obtenu entre le « meilleur » et « plus mauvais » modèle ARs est de l'ordre de 10% (entre 6% et 13% contre 7% et 11% pour Ajaccio).

A l'exception de la persistance simple, les deux modèles les moins performants qui se détachent sont les Arbres de Régression simple (ARs) (déjà dernier dans le cas d'Ajaccio).

Tableau IV-6: Erreurs de prévision en fonction de l'horizon pour les données de Tilos

	Tilos	*h+1*	*h+2*	*h+3*	*h+4*	*h+5*	*h+6*
nRMSE (%)	P	30,90	49,14	62,44	70,61	72,72	71,17
	PI	23,58	33,69	40,82	44,85	46,89	46,65
	PIK	22,48	31,99	39,62	43,65	45,59	46,03
	ARMA	22,26	31,33	33,22	33,13	33,39	33,78
	PMC	20,57	31,18	31,62	32,08	33,07	34,11
	PG	19,34	29,46	30,57	32,30	32,09	33,39
	MVS	21,84	30,73	31,58	32,63	32,99	33,71
	ARs	25,74	36,71	37,51	38,32	39,16	39,66
	ARel	20,79	31,16	32,74	34,28	34,42	34,92
	ARr	21,12	30,35	32,09	32,41	33,82	33,53
	ARen	19,73	29,28	30,40	32,10	32,10	32,79
	FA	19,70	29,54	31,13	31,94	32,16	32,49
RMSE (Wh.m^{-2})	P	192,75	307,33	389,69	440,61	454,65	443,14
	PI	147,11	210,71	254,77	279,86	293,20	290,44
	PIK	145,10	209,37	252,64	275,66	289,65	290,03
	ARMA	138,83	195,93	207,35	206,69	208,76	210,30
	PMC	128,44	194,20	197,53	200,43	206,54	212,44
	PG	120,78	183,80	191,54	201,19	200,19	207,81
	MVS	136,16	264,98	264,47	263,38	264,47	262,43
	ARs	161,17	229,12	234,64	239,44	244,74	247,11
	ARel	130,10	194,66	204,26	213,71	214,96	218,25
	ARr	131,93	189,92	199,95	202,30	209,87	209,26
	ARen	122,94	183,34	189,84	199,31	200,55	205,22
	FA	122,96	184,00	193,86	198,82	200,33	202,55
nMAE (%)	P	25,52	41,64	52,96	59,84	61,13	59,24
	PIs	15,95	25,12	31,40	34,65	36,16	35,46
	PIK	15,63	24,79	31,14	34,36	35,81	35,37
	ARMA	16,38	21,73	22,83	22,79	22,88	22,93
	PMC	14,22	21,69	21,76	21,87	22,31	22,87
	PG	11,91	19,85	20,69	21,72	21,48	22,20
	MVS	14,83	21,86	22,26	22,53	22,75	23,02
	ARs	15,26	24,78	25,62	26,33	26,95	27,14
	ARel	14,02	21,08	22,20	23,10	23,22	23,57
	ARr	14,45	20,93	21,91	22,11	22,65	22,52
	ARen	12,33	19,50	20,24	21,34	21,42	21,87
	FA	11,92	19,56	20,61	21,26	21,32	21,63
MAE (Wh.m^{-2})	P	159,18	259,74	330,35	373,31	381,34	369,55
	PIs	99,50	156,69	195,89	216,15	225,56	221,22
	PIK	97,49	154,65	194,24	214,33	223,36	220,64
	ARMA	102,16	135,57	142,44	142,16	142,70	143,06
	PMC	88,68	135,32	135,76	136,46	139,20	142,70
	PG	74,29	123,82	129,09	135,49	134,03	138,50
	MVS	92,49	136,40	138,84	140,57	141,89	143,61
	ARs	95,22	154,61	159,81	164,23	168,11	169,33
	ARel	87,45	131,51	138,49	144,09	144,83	147,06
	ARr	90,14	130,57	136,70	137,94	141,27	140,50
	ARen	76,93	121,67	126,25	133,12	133,63	136,46
	FA	74,30	122,03	128,54	132,65	132,99	134,93

La *Figure IV-6* représente le skill score en prenant comme référence la persistance intelligente simple (PI). Tous les modèles ont un comportement semblable, à l'exception de ARs, ils sont tous meilleurs

que la PI quel que soit l'horizon. Le modèle de PI améliorée par les filtres de Kalman (PIK) présente toujours de meilleures performances que la PI simple mais l'écart se réduit avec l'horizon.

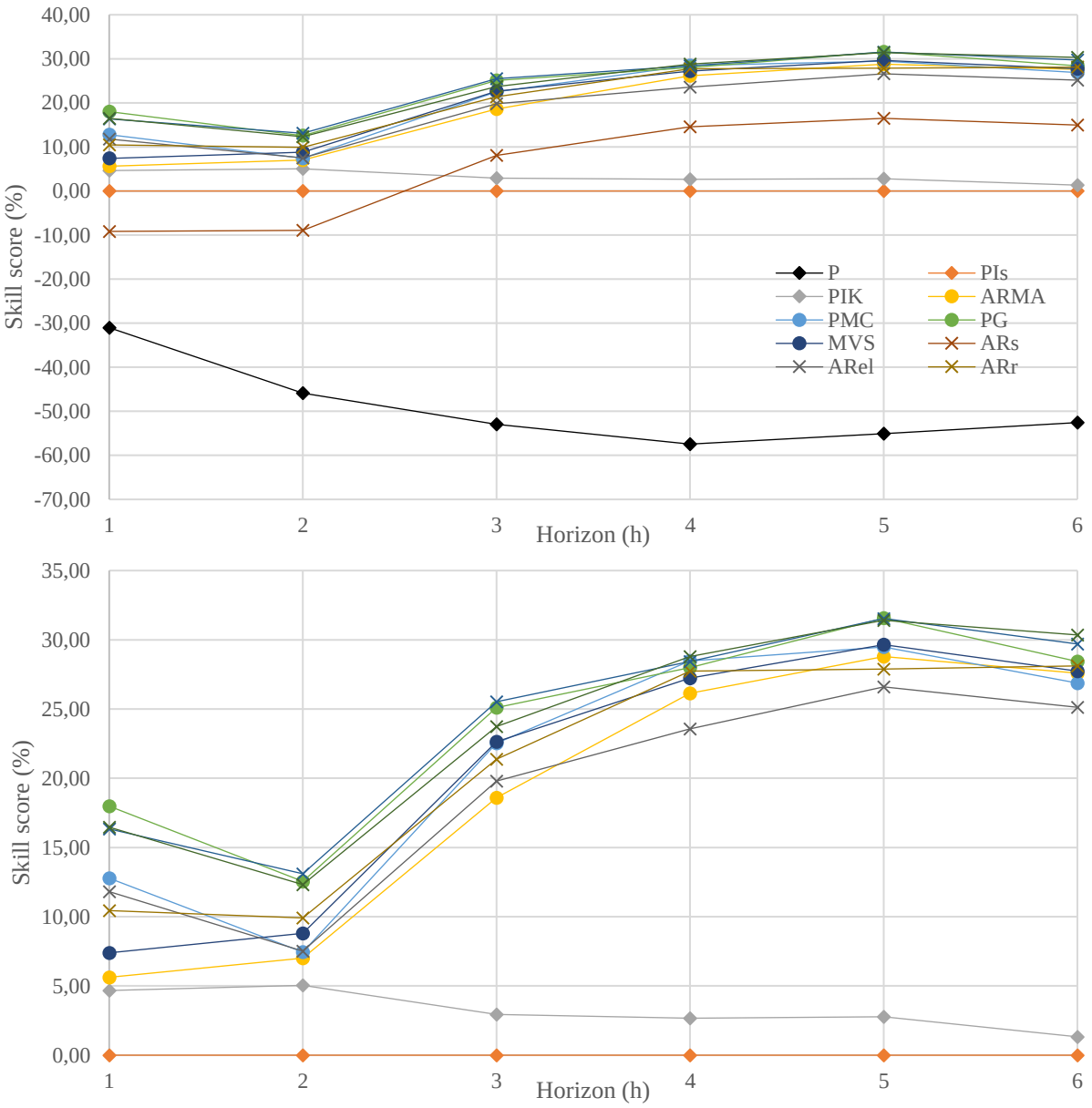


Figure IV-6: Skill score en fonction de l'horizon, en haut tous les modèles, en bas zoom sur les modèles les plus intéressants pour les données de Tilos

La *Figure IV-7* représente la MBE pour chaque modèle et chaque horizon temporel. On constate que quel que soit l'horizon, tous les modèles présentent une MBE inférieure à 15 $Wh.m^{-2}$, correspondant une fois normalisée à 2,9%. Globalement, tous nos modèles ont tendance à surestimer la réalité expérimentale mais l'erreur reste acceptable.

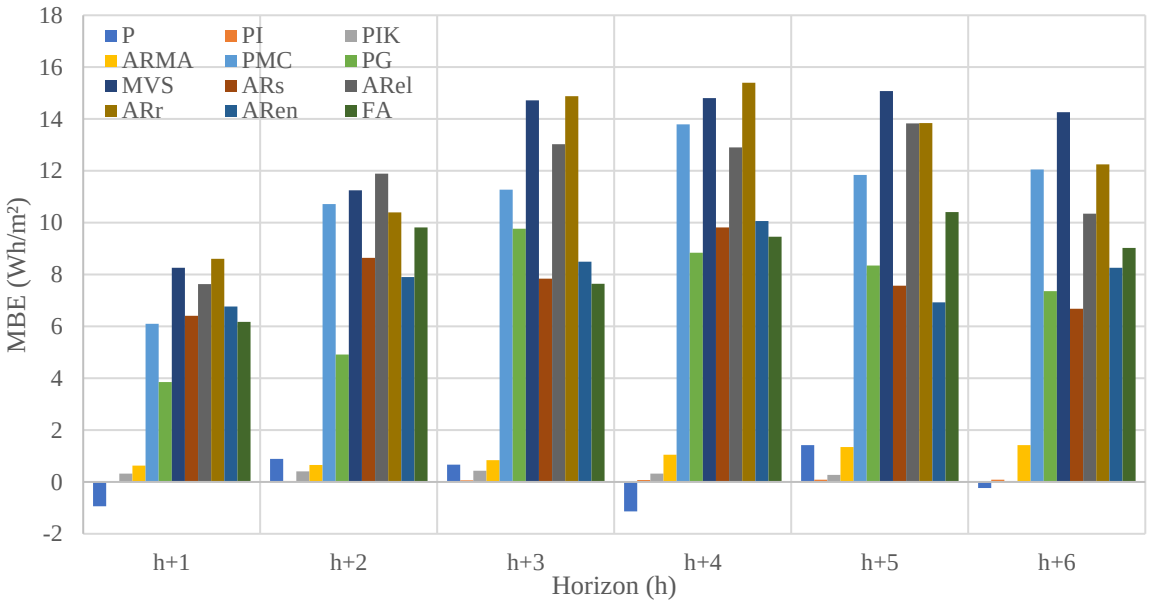


Figure IV-7: MBE en fonction de l'horizon pour les prévisions sur les données de Tilos

Nous avons considéré le « meilleur » modèle était le modèle ARen dont nous présentons les performances sur la *Figure IV-8* où sont représentées les données prévues en fonction des données mesurées.

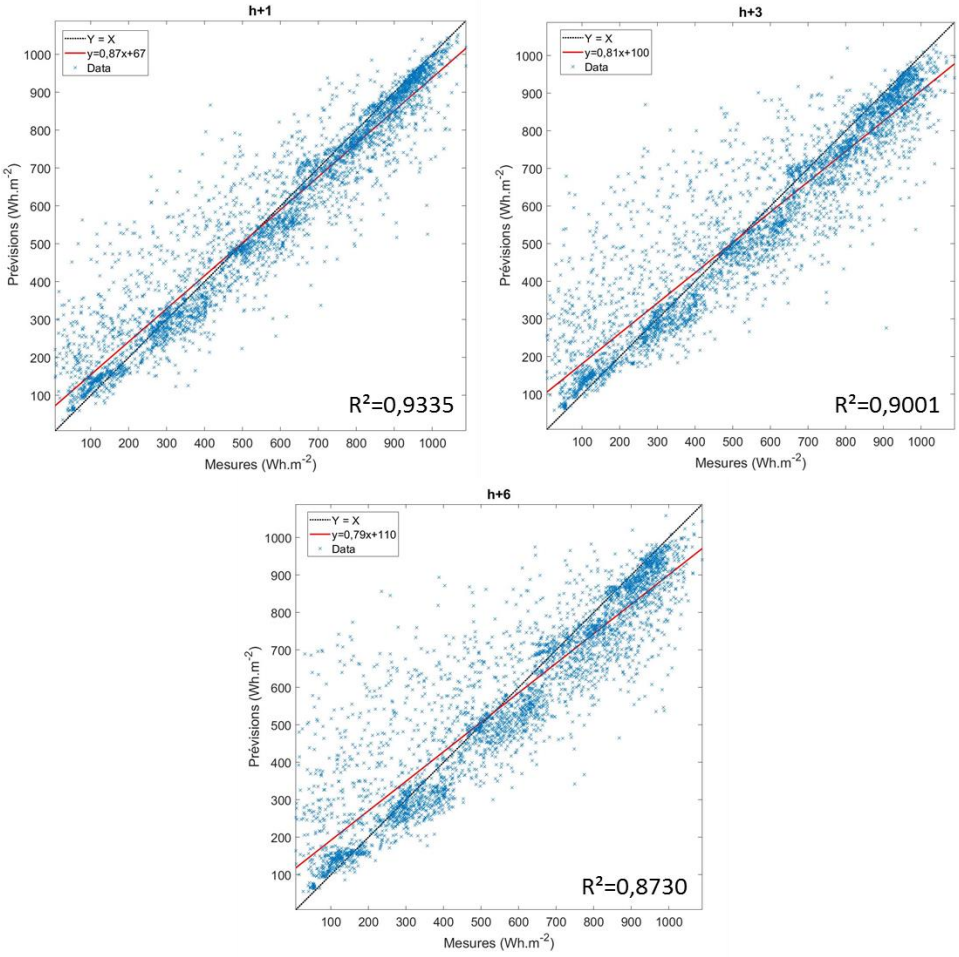


Figure IV-8: Représentation des valeurs prévues en fonction des valeurs mesurées pour trois horizons temporels pour le modèle ARen

On note que les pentes des droites rouges sont plus élevées que pour Ajaccio, tout en gardant à l'esprit qu'il ne s'agit pas du même modèle de prédiction, dénotant ainsi que la droite de régression est plus proche de la droite idéale, ce qui se retrouve également au niveau des ordonnées à l'origine plus faible que dans le cas du meilleur modèle pour Ajaccio.

En conclusion de ce paragraphe sur les données de Tilos, ville présentant une variabilité moyenne (*var* = 11,0%), légèrement supérieure à celle d'Ajaccio (*var* = 9,2%) on note :

- Trois modèles se distinguent des autres et les erreurs de prévision sont proches quel que soit l'horizon considéré ;
- Deux modèles ne sont pas adaptés, la persistance simple, les arbres de régression simples ;
- Bien que les performances soient proches, le modèle ARen est le plus adapté pour ce site de mesure à moyenne variabilité, suivi des processus gaussiens.

2.2.3. Application aux données de Nancy

Nous avons ensuite réalisé les prévisions sur les données en provenance de Nancy, données dont la variabilité est également moyenne, mais plus élevée que celles des sites d'Ajaccio et Tilos (voir Chapitre 2).

Tableau IV-7: nombre d'entrées (i) utilisées dans les modèles d'apprentissage automatique pour Nancy (6h/1h)

Horizon	*1*	*2*	*3*	*4*	*5*	*6*
Nombre d'entrées(i)	4	4	5	5	5	6

Comme pour Tilos et Ajaccio, il est difficile de commenter ces résultats mais le nombre maximal de données précédent l'instant de prédiction est toujours de 6 soit 5 heures plus tôt.

Les résultats des simulations sont présentés dans le *Tableau IV-8*, avec en couleur les trois meilleures valeurs selon le code couleur définie précédemment.

On note, contrairement aux deux autres stations, des valeurs de métriques beaucoup plus élevées et des écarts plus importants au niveau des résultats entre les modèles.

Si précédemment, les trois « premiers » modèles donnaient des résultats quasi identiques, pour Nancy, on constate que si les deux plus performants ont des résultats assez proches avec un écart en termes de nRMSE de l'ordre de 1% (Forêts aléatoires et Arbres de régression ensachés), le troisième (Arbres de régression élagués) a de bien moins bonnes performances avec un écart en nRMSE de près de 5%.

Si le classement est valable quel que soit la métrique utilisée pour les deux premiers modèles (FA et ARen), la troisième place n'est pas attribuée au même modèle si le classement est réalisé avec la nRMSE ou avec la nMAE : la troisième place est occupée par ARel d'un point de vue de la nRMSE mais par les Processus Gaussiens en considérant la nMAE. Encore une fois, on retrouve la difficulté de comparaison des modèles selon la métrique utilisée. Il est reconnu que la nRMSE est plus employée que la nMAE car elle prend mieux en compte les erreurs importantes.

L'écart entre le meilleur modèle, qui est ARen, et la PI est plus élevé que pour Ajaccio mais du même ordre de grandeur que pour Tilos puisqu'il passe de 3% à h+1 (contre 0,9% pour Ajaccio et 4% pour Tilos) à 13% à h+6 (contre 6% pour Ajaccio et 14% pour Tilos). Cependant, l'écart obtenu entre le « meilleur » et « plus mauvais » (ARs) modèle est plus important car il varie entre 8% pour h+1 et 29%

pour h+6 (entre 6% et 13% pour Tilos et entre 7% et 11% pour Ajaccio). A l'exception de la persistance simple, le modèle le moins performant qui se détache est celui basé sur les Arbres de régression simple, déjà derniers dans le cas de Tilos.

Tableau IV-8: Erreurs de prévision fonction de l'horizon pour les données de Nancy

Nancy		*h+1*	*h+2*	*h+3*	*h+4*	*h+5*	*h+6*
nRMSE (%)	P	35,92	53,21	65,32	74,03	77,43	77,16
	PI	28,92	37,15	41,95	47,02	49,44	51,58
	PIK	27,32	36,73	40,81	46,52	48,24	50,38
	ARMA	27,28	48,75	50,76	52,36	53,54	54,16
	PMC	27,38	49,00	50,31	52,29	53,24	54,36
	PG	27,46	49,52	51,05	51,88	52,98	53,83
	MVS	27,77	36,65	40,95	44,91	47,81	49,27
	ARs	33,70	52,59	64,12	66,34	66,99	67,60
	ARel	29,87	36,52	40,62	43,91	45,65	48,07
	ARr	29,65	36,86	40,74	44,34	46,51	47,79
	ARen	25,95	31,08	33,70	35,99	38,13	39,29
	FA	26,05	30,14	33,25	35,57	37,81	38,64
RMSE (Wh.m^{-2})	P	136,15	201,68	247,59	280,60	293,49	292,46
	PI	109,62	140,81	159,01	178,22	187,40	195,51
	PIK	103,55	139,22	154,68	176,33	182,85	190,96
	ARMA	103,40	184,78	192,40	198,46	202,94	205,29
	PMC	103,78	185,73	190,69	198,20	201,80	206,04
	PG	104,08	187,70	193,50	196,64	200,81	204,03
	MVS	105,25	138,91	155,23	170,24	181,20	186,75
	ARs	127,73	199,33	243,04	251,45	253,92	256,23
	ARel	113,22	138,42	153,96	166,43	173,03	182,20
	ARr	112,38	139,71	154,42	168,06	176,29	181,14
	ARen	98,36	117,80	127,73	136,41	144,53	148,92
	FA	98,74	114,24	126,03	134,82	143,31	146,46
nMAE (%)	P	27,49	41,73	51,48	57,86	59,82	59,30
	PI	14,57	20,03	23,45	26,59	28,57	30,10
	PIK	14,05	18,90	22,59	25,76	27,43	29,29
	ARMA	17,54	21,84	24,62	26,82	28,03	28,81
	PMC	17,46	22,54	24,73	26,69	27,57	28,64
	PG	14,03	18,69	21,28	21,57	22,53	23,08
	MVS	15,16	20,19	23,16	25,72	27,08	28,12
	ARs	17,26	18,71	20,79	21,75	25,70	28,79
	ARel	17,32	21,26	23,44	25,62	26,54	27,54
	ARr	17,16	21,35	24,08	25,93	27,19	27,77
	ARen	14,35	17,49	19,26	20,79	22,07	22,49
	FA	14,48	17,39	18,95	20,21	21,38	22,07
MAE (Wh.m^{-2})	P	104,20	158,17	195,14	219,29	226,74	224,76
	PI	55,22	75,94	88,90	100,80	108,29	114,08
	PIK	53,26	71,64	85,62	97,64	103,99	111,02
	ARMA	66,48	82,77	93,33	101,65	106,25	109,21
	PMC	66,19	85,42	93,74	101,18	104,49	108,54
	PG	53,19	70,83	78,64	81,77	85,45	87,49
	MVS	57,45	76,53	87,79	97,48	102,64	106,59
	ARs	65,44	70,93	80,68	82,42	97,32	109,11
	ARel	65,66	80,58	88,86	97,10	100,59	104,40
	ARr	65,05	80,93	91,29	98,29	103,08	105,27
	ARen	54,38	66,29	73,01	78,81	83,65	85,25
	FA	54,88	65,90	71,82	76,61	81,04	83,64

La *Figure IV-9* représente le skill score en prenant comme référence la persistance intelligente. On observe vraiment des groupes de modèles avec FA et ARen qui se détachent et sont toujours « meilleurs » quel que soit l'horizon considéré ; à l'exception de ARel et ARr qui ont des performances proches des modèles de persistance intelligente, tous les autres prédicteurs sont moins bons que la référence. Le modèle de PI améliorée par les filtres de Kalman (PIK) présente toujours de meilleures performances que la PI.

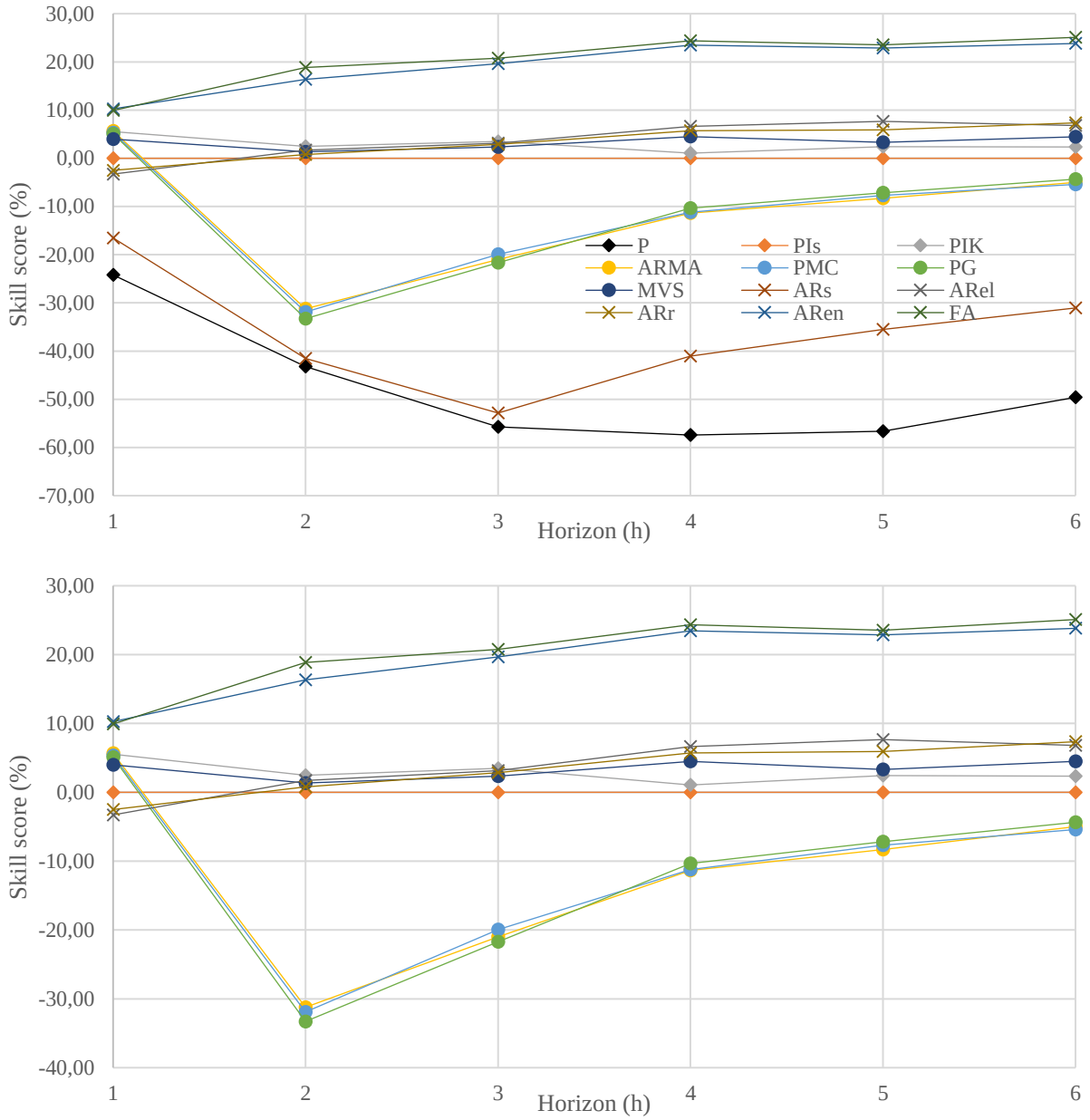


Figure IV-9: Skill score en fonction de l'horizon, en haut tous les modèles, en bas zoom sur les modèles les plus intéressants pour les données de Nancy

La *Figure IV-10* représente la MBE en fonction de l'horizon pour les données de Nancy. En ce qui concerne le biais, nous pouvons remarquer que tous les modèles ont un biais positif qui tend à augmenter à mesure que l'horizon grandit. Le biais maximum, pour le modèle de persistance intelligente simple à un horizon de 6 heures, est de l'ordre de + 16 $Wh.m^{-2}$ soit en valeur relative près de 4%.

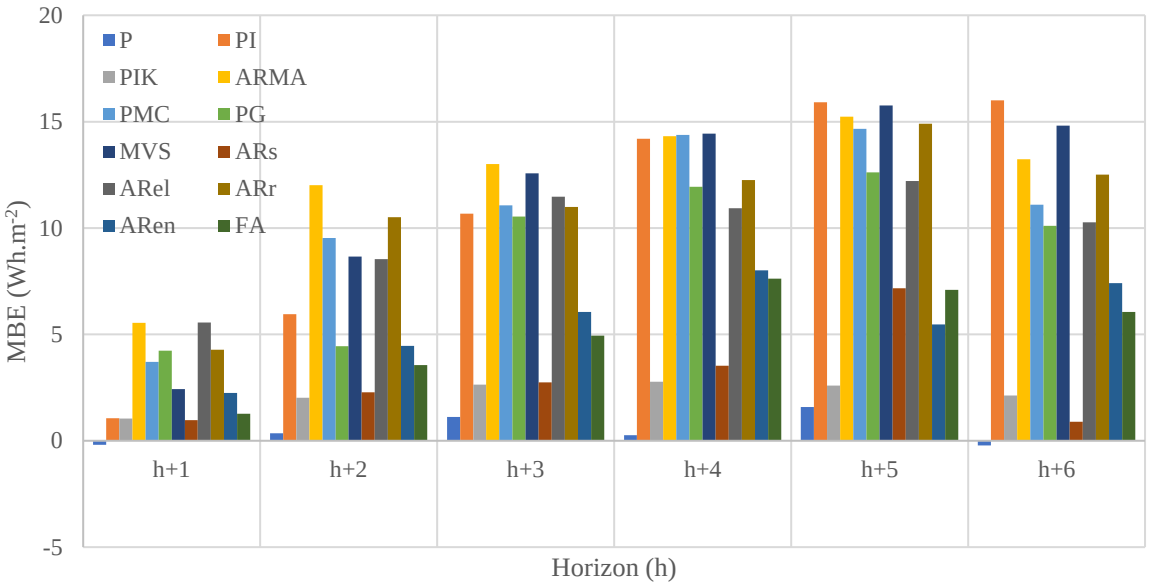


Figure IV-10: MBE en fonction de l'horizon pour les prévisions sur les données de Nancy

Même si les deux premiers modèles ont des performances proches, nous avons considéré que le « meilleur » était le modèle FA dont nous présentons les performances sur la *Figure IV-11*où sont représentées les données prévues en fonction des données mesurées pour les trois horizons h+1, h+3 et h+6.

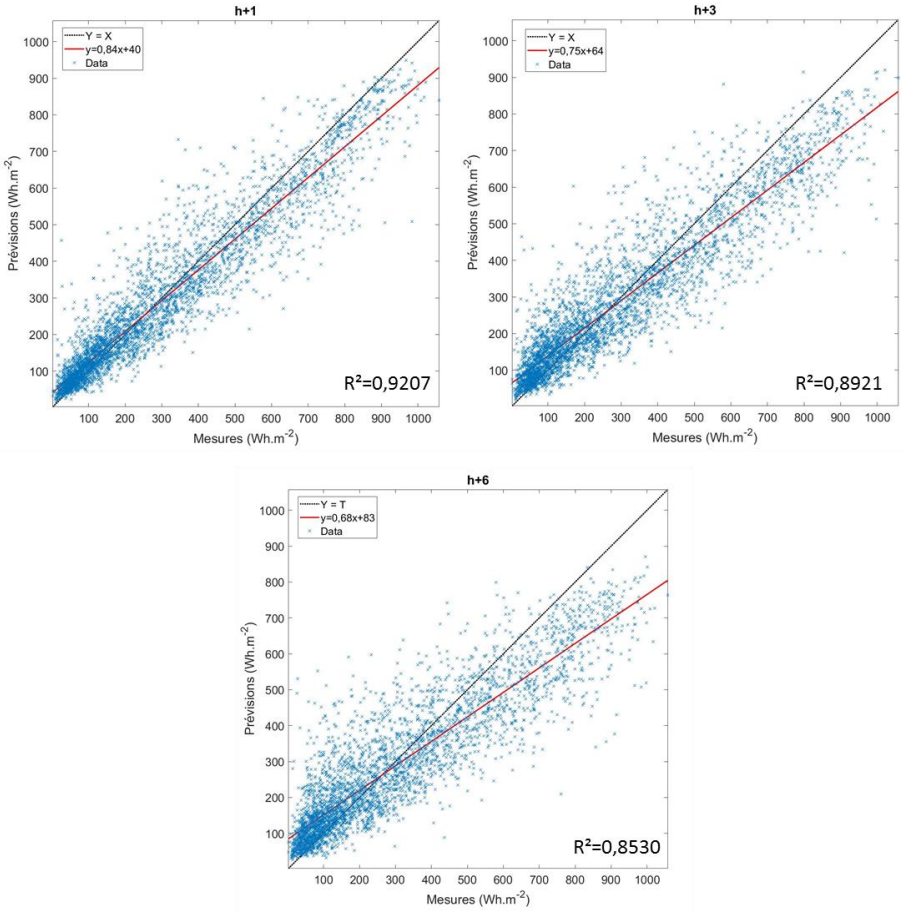


Figure IV-11: Représentation des valeurs prévues en fonction des valeurs mesurées pour trois horizons temporels pour le modèle FA

On note que les pentes des droites rouges sont plus élevées que pour Tilos et Ajaccio, tout en gardant à l'esprit qu'il ne s'agit pas du même modèle de prédiction, ce qui se retrouve également au niveau des ordonnées à l'origine (liées au biais) plus faible que dans le cas du meilleur modèle pour Ajaccio et Tilos.

En conclusion de ce paragraphe sur les données de Nancy, ville présentant une variabilité moyenne (16,9%), supérieure à celle d'Ajaccio (9,2%) et de Tilos (11,0%), on note :

- Les incertitudes sont plus élevées que pour les deux cas précédents et des groupes de prédicteurs apparaissent plus nettement ;
- Deux modèles se distinguent des autres quel que soit la métrique considérée, il s'agit de ARen et FA. Les écarts de nRMSE et de nMAE entre eux sont faibles quel que soit l'horizon considéré (de l'ordre de 1%) ;
- Trois modèles semblent ne pas être adaptés, en plus de la persistance simple, les PMC, PG et ARMA comme dans le cas de Tilos ;
- Bien que les performances soient proches, le modèle FA est le plus adapté pour ce site de mesure à moyenne variabilité, suivi des ARen.

2.2.4. Application aux données d'Odeillo

Nous avons poursuivi les prévisions sur les données en provenance d'Odeillo, données dont la variabilité peut être considérée comme élevée avec une valeur de 23,5%.

Tableau IV-9: nombre d'entrées (i) utilisées dans les modèles d'apprentissage automatique pour Odeillo (6h/1h)

Horizon	*1*	*2*	*3*	*4*	*5*	*6*
Nombre d'entrées (i)	6	6	7	6	7	7

On note que le nombre de données d'entrées est plus élevé (en général d'un point) que pour les trois stations précédentes.

Les résultats des simulations sont présentés dans le *Tableau IV-10*, avec en couleur les trois meilleures valeurs selon le code couleur définie précédemment.

Tableau IV-10: Erreurs de prévision en fonction de l'horizon pour les données d'Odeillo

	Odeillo	*h+1*	*h+2*	*h+3*	*h+4*	*h+5*	*h+6*
nRMSE (%)	P	40,91	61,81	76,18	84,53	88,35	86,25
	PI	37,00	54,59	66,68	74,64	77,77	77,76
	PIK	36,00	53,42	65,33	72,62	76,79	76,62
	ARMA	34,42	47,12	49,26	49,76	50,16	51,56
	PMC	29,88	43,51	45,86	47,87	48,44	48,72
	PG	28,80	42,35	45,37	46,67	48,42	48,54
	MVS	29,39	43,71	46,25	48,76	49,27	49,87
	ARs	36,73	51,94	54,95	56,76	57,86	57,45
	ARel	29,90	43,97	46,47	48,47	50,01	49,84
	ARr	30,20	43,24	45,59	47,71	48,57	48,78
	ARen	28,72	42,00	44,66	46,63	47,83	47,52
	FA	28,76	42,15	44,89	46,56	47,78	48,34

RMSE (Wh.m^{-2})	*P*	*197,60*	*298,54*	*367,95*	*408,28*	*426,73*	*416,59*
	PI	178,71	263,67	322,06	360,51	375,63	375,58
	PIK	173,88	258,02	315,56	350,78	370,88	370,07
	ARMA	166,25	227,59	237,93	240,34	242,27	249,03
	PMC	144,32	210,15	221,50	231,21	233,97	235,32
	PG	139,10	203,58	219,14	225,42	234,45	233,87
	MVS	141,95	211,12	223,39	235,51	237,97	240,87
	ARs	177,41	250,87	265,41	274,15	279,46	277,48
	ARel	144,42	212,38	224,45	234,11	241,55	240,73
	ARr	145,87	208,85	220,20	230,44	234,59	235,61
	ARen	138,72	202,86	215,71	225,22	231,02	229,52
	FA	138,91	204,55	216,82	224,88	230,78	233,48
nMAE (%)	P	24,33	43,19	55,19	62,44	64,13	62,20
	PI	20,89	37,18	47,50	53,76	55,11	53,97
	PIK	20,63	36,79	47,09	53,42	54,59	53,35
	ARMA	20,68	41,15	42,19	42,01	41,90	41,63
	PMC	18,26	39,40	40,77	41,03	41,37	41,58
	PG	17,05	39,80	41,11	42,03	40,95	41,29
	MVS	17,78	40,02	40,80	41,59	41,62	42,06
	ARs	19,99	43,35	46,27	47,18	47,26	47,14
	ARel	19,16	39,31	41,50	41,67	41,67	41,48
	ARr	19,48	39,71	40,92	41,91	41,45	41,51
	ARen	15,95	38,22	39,63	40,32	40,33	40,17
	FA	15,63	37,84	39,22	40,25	40,13	40,43
MAE (Wh.m^{-2})	P	117,53	208,62	266,59	301,60	309,76	300,41
	PI	100,91	179,58	229,42	259,65	266,18	260,68
	PIK	99,66	177,68	227,46	258,04	263,68	257,68
	ARMA	99,88	198,76	203,76	202,90	202,38	201,09
	PMC	88,19	190,29	196,90	198,18	199,84	200,83
	PG	82,35	192,23	198,54	203,02	197,78	199,44
	MVS	85,89	193,30	197,06	200,87	201,01	203,15
	ARs	96,56	209,36	223,47	227,90	228,27	227,69
	ARel	92,54	189,87	200,45	201,25	201,24	200,33
	ARr	94,10	191,78	197,63	202,45	200,22	200,48
	ARen	77,05	184,61	191,41	194,76	194,79	194,04
	FA	75,47	182,77	189,45	194,41	193,84	195,29

On remarque que les niveaux d'erreur sont bien plus élevés que pour les autres stations :

- S'il était possible pour Ajaccio de prévoir à h+1 avec une nRMSE de l'ordre de 18% pour Odeillo, cette même prévision ne peut se faire dans le meilleur des cas qu'avec une incertitude de 28%, soit avec 10 points de pourcentage de plus ;
- La prédiction à Ajaccio à h+6 pouvait être réalisée avec une nRMSE de 33% contre 47% pour Odeillo, soit 14 points de plus.

On ne retrouve pas, comme pour Nancy, deux modèles qui sortent du lot, mais de nouveau des modèles dont les performances sont proches. Les trois « premiers » modèles sont, par ordre décroissant de performance, les Arbres de régression ensaché, les Forêts aléatoires et les Processus Gaussien mais avec des écarts entre le premier et le troisième de moins de 1% quel que soit l'horizon considéré. Même les quatrième et cinquième modèles (PMC et ARel) permettent d'obtenir des nRMSE du même ordre de grandeur (1,2 à 1,6% d'écart avec les ARen).

Si le classement est valable quel que soit la métrique utilisée pour les deux premiers modèles FA et ARen, la troisième place n'est pas aussi clairement attribuée ce qui n'est pas surprenant étant donnée la faible différence de nRMSE notée entre les 5 meilleurs prédicteurs cités précédemment.

La PI qui présentait des résultats « convenables » pour les stations précédentes fait partie des plus mauvais modèles pour Odeillo. L'écart entre le meilleur modèle (ARen) et la PI est beaucoup plus élevé que pour les autres stations puisqu'il passe de 8% à h+1 (0,9% pour Ajaccio, 4% pour Tilos et 3% pour Nancy) à 29% à h+6 (6% pour Ajaccio, 14% pour Tilos et 13% pour Nancy). Le plus mauvais modèle étant PI, la comparaison entre le meilleur et plus mauvais modèle n'est pas présentée ici.

A l'exception de la persistance simple, les deux modèles les moins performants sont la PI et la PIK.

La *Figure IV-12* représente le skill score en prenant comme référence la PI. On observe trois groupes de modèles avec le groupe des « plus performants » qui se tiennent dans un mouchoir de poche (ARen, FA, PG, PMC, MVS, ARel et ARr), les modèles moyens (ARMA et ARs) et enfin les plus mauvais modèles basés sur la Persistance.

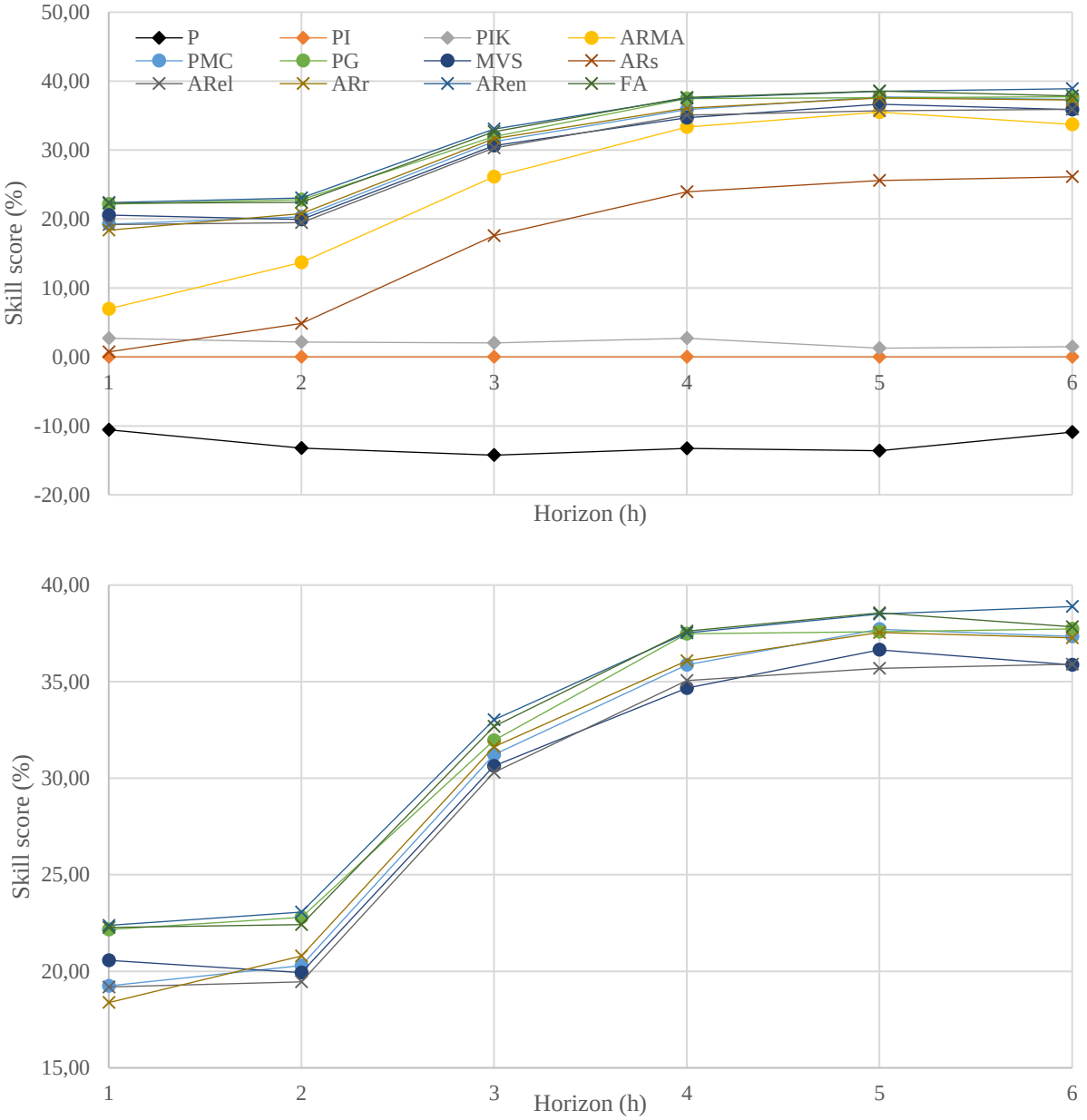


Figure IV-12: Skill score en fonction de l'horizon, en haut tous les modèles, en bas zoom sur les modèles les plus intéressants pour les données d'Odeillo

Dans le cas du zoom, nous avons enlevé les représentations graphiques des modèles les moins bons (P, Pi, PIK, ARMA et Ars) pour mieux voir les différences entre les autres modèles.

La *Figure IV-13* représente le calcul de la MBE en fonction de l'horizon. En ce qui concerne le biais, nous retrouvons les tendances des deux premières stations, à savoir que tous les prédicteurs ont tendance à surestimer. Le biais maximum est de l'ordre de 17,5 Wh.m-2 soit en valeur relative près de 3,5%.

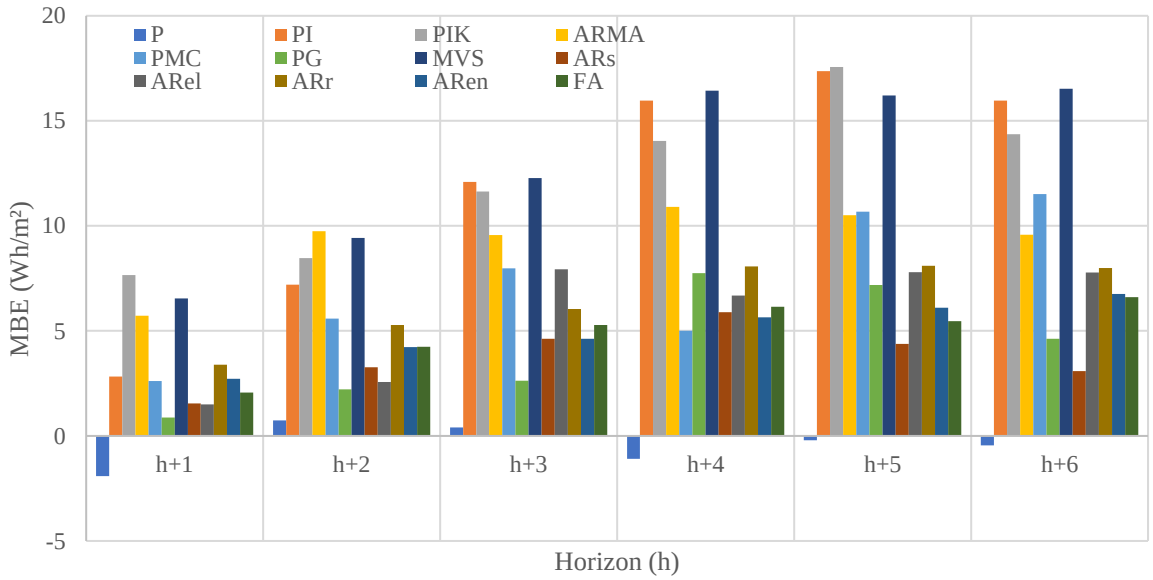


Figure IV-13: MBE en fonction de l'horizon pour les prévisions sur les données d'Odeillo

Même si les cinq premiers modèles ont des performances proches, nous avons considéré que le « meilleur » était le modèle ARen dont nous présentons les performances sur la Figure IV-14 où sont représentées les données prévues en fonction des données mesurées pour les trois horizons h+1, h+3 et h+6.

La dispersion des points est plus importante et les pentes des droites rouges sont de nouveau plus faibles que pour les trois autres stations étudiées. Cette augmentation de la dispersion est liée à l'augmentation des incertitudes notée précédemment.

En conclusion de ce paragraphe sur les données d'Odeillo, ville présentant une forte variabilité (23,5%), on note :

- Les incertitudes sont bien plus élevées que pour les trois cas précédents ;
- Trois groupes de prédicteurs apparaissent, les meilleurs ARen, FA, ARr, ARel, PMC, GP, MVS, ensuite on retrouve ARMA et ARs enfin les modèles basés sur la persistance PIK, PI et P ;
- Les persistances, qu'elles soient intelligentes ou non, ne sont pas adaptées à ce site à forte variabilité ;
- Un groupe de cinq modèles se distinguent des autres, mais l'écart des incertitudes entre eux est tels qu'il est difficile de déterminer le meilleur d'entre eux (de l'ordre de 1% en termes de nRMSE entre les 5 premiers) ;
- Bien que les performances soient proches, le modèle ARen est le plus adapté pour ce site de mesure à haute variabilité.

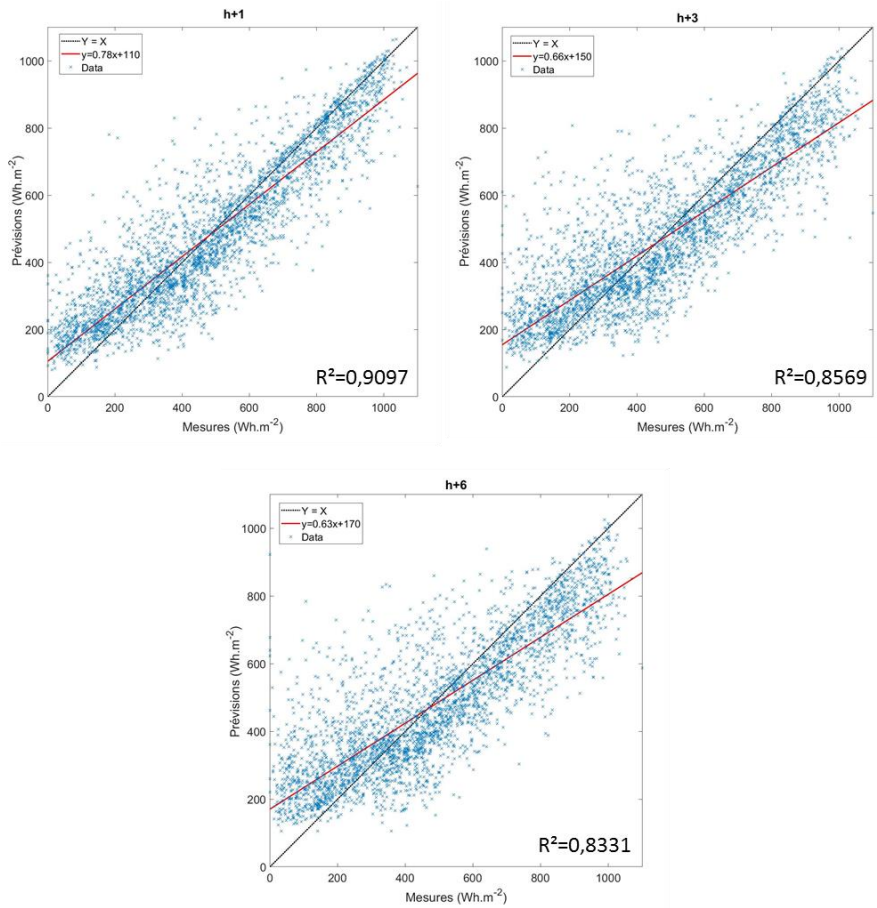


Figure IV-14: Représentation des valeurs prévues en fonction des valeurs mesurées pour trois horizons temporels pour le modèle ARen

2.3. Synthèse sur la prévision horaire pour les 4 sites de mesure

L'objectif que nous nous étions fixés est de déterminer s'il existe un lien entre variabilité des données et efficacité des modèles prédictifs et ainsi de faire ressortir les « meilleurs » modèles pour chaque site. Même s'il s'avère difficile de donner sans équivoque un classement des modèles en fonction de la variabilité des données, nous pouvons tout de même dégager des tendances. Il est apparu de manière générale que plus les données ont une forte variabilité, plus les modèles prédictifs adéquats deviennent complexes à la fois dans leur structure et dans leur mise en œuvre. Comme nous l'avons vu dans les paragraphes précédents, il y avait très peu de différences entre le classement effectué sur la base de la nMAE et de la nRMSE ; par conséquent, il nous est apparu pertinent de commenter nos résultats dans ce paragraphe sur la base de la nRMSE, métrique la plus répandue dans la littérature sur la prévision. De même, la persistance simple donnant, par nature, de mauvais résultats, nous ne commenterons plus ses performances.

Nous présentons un classement des modèles pour chaque jeu de données dans le Tableau IV-11 ; nous avons calculé la « nRMSE moyenne » sur l'ensemble des horizons pour chaque modèle puis, à titre indicatif, la médiane des nRMSE.

Tableau IV-11: Classement des modèles pour les différents sites de mesure en fonction de la nRMSE globale calculée pour tous les horizons

	Ajaccio		Tilos		Nancy		Odeillo	
	<nRMSE>	Classement	<nRMSE>	Classement	<nRMSE>	Classement	<nRMSE>	Classement
P	52,13	12	59,50	12	63,84	12	73,01	12
PI	31,16	10	39,41	11	42,68	7	64,74	11
PIK	30,54	8	38,23	10	41,53	6	63,46	10
ARMA	29,61	1	31,18	7	47,81	10	47,05	8
PMC	29,74	2	30,44	4	47,77	8	44,05	5
PG	30,41	7	29,53	3	47,79	9	43,33	3
MVS	29,37	5	30,58	6	41,23	5	44,54	6
ARs	36,60	11	36,18	9	58,55	11	52,62	9
ARel	30,64	9	31,39	8	40,77	3	44,78	7
ARr	30,06	6	30,55	5	40,98	4	44,02	4
ARen	29,96	3	29,40	1	34,02	2	42,89	1
FA	30,00	4	29,49	2	33,58	1	43,11	2
Médiane	30,06		30,58		41,53		44,54	

De façon globale, on peut dire que :

- Pour les données d'Ajaccio, à faible variabilité (*var* < 10%), tous les modèles présentent des performances comparables. Il n'est alors pas nécessaire d'utiliser un modèle très complexe pour avoir une prévision de qualité. Typiquement, un modèle simple comme la persistance intelligente est largement utilisable pour des horizons de 1 à 3 heures. Au-delà de ces horizons, un modèle ARMA ou PMC sera préférable ;
- Pour les données de Tilos, à variabilité moyenne (10% < *var* < 15%), les modèles simples et les arbres de régression simples ont plus de mal à donner de bons résultats. Les meilleurs modèles sont ceux qui sont construits sur les apprentissages d'ensemble, cependant ils ne surclassent les autres modèles de machine learning que faiblement. Pour ce type de données, il est donc tout à fait possible de se servir des modèles que nous avons appelés « classiques », à savoir ARMA, PMC, processus Gaussien. Pour améliorer encore les prévisions, il faudra se tourner vers les modèles basés sur les arbres de régression dans leurs versions améliorées : élaguées, renforcées, ensachées et les forets aléatoires ;
- Pour les données de Nancy, qui ont une variabilité supérieure (15% < *var* < 20%) les modèles sans apprentissage sont à bannir car peu efficaces, les modèles dits « classiques » ont certes de meilleures performances mais restent peu satisfaisants. Si dans le cas de Tilos, les modèles basés sur l'apprentissage d'ensemble étaient légèrement meilleurs que les autres, dans le cas de Nancy, ils dominent beaucoup plus les autres modèles ;
- Enfin, pour les données d'Odeillo, à variabilité forte (20% < *var*), les modèles les plus simples, basés sur la persistance ou de type ARMA ne sont pas efficaces. Les modèles d'apprentissage automatique classiques comme PMC, PG ou MVS ont des performances médiocres. Les modèles les plus complexes sont quant à eux les meilleurs, même s'il faut souligner que les niveaux de nRMSE sont quand même bien supérieurs à ceux que l'on obtient pour les autres sites.

Il pourrait être intéressant de coupler les résultats présentés dans *Tableau IV-11* avec les informations sur la complexité des modèles du *Tableau IV-1*ce qui permettrait de réaliser un choix plus pertinent lorsque deux modèles donnent des résultats proches d'un point de vue de leur performance de prédiction.

3. Prévision probabiliste, application au rayonnement global horizontal

La méthode de prévision probabiliste que nous avons choisie et qui permet d'obtenir un encadrement de la prévision a été détaillée dans le Chapitre 3. Nous avons développé une approche d'ensemble d'arbres de régression mixant « bootstrap », quantiles et fonctions de répartition. Le bootstrap est basé

sur la réplication multiple des données selon les techniques de rééchantillonnage. Les apprentissages statistiques sont réalisés sur chacun des sous-échantillons et permettent d'obtenir les fonctions de répartition et les distributions des prévisions. Nous avons appliqué cette méthode à la prévision du rayonnement global horizontal sur les mesures réalisées à Ajaccio avec une granularité horaire.

Nous avons choisi d'utiliser les arbres de régression durant cette étude, car à la suite de l'analyse bibliographique que nous avons réalisée (Voyant *et al.*, 2017), ces modèles étaient ressortis comme étant de plus en plus utilisés en estimation de la ressource solaire et possédaient de ce fait un fort potentiel prédictif. Nous n'avons pas souhaité intégrer d'autres prédicteurs durant cette série de simulations, car comme nous l'avons montré dans la section précédente, quels que soient le site d'étude et l'horizon considéré, tous les résultats issus de l'apprentissage automatique sont sensiblement identiques et qu'il est très difficile d'établir un classement objectif. Il est toutefois intéressant de remarquer que les arbres de régression se positionnent systématiquement comme les meilleurs modèles prédictifs. De ce fait, pour ne pas alourdir le manuscrit nous nous sommes focalisés sur les modèles ARs, ARel, ARr et ARen. Notons, qu'une de nos motivations résidait aussi dans le fait que ce type de modèles n'avait que très rarement été dédié à la prévision probabiliste du rayonnement solaire.

Il faut garder à l'esprit qu'au moment de la réalisation de ces simulations, peu d'études (traitant de prévisions probabilistes de l'irradiation solaire) avaient été réalisées et qu'il nous était très difficile de valider nos résultats en nous appuyant sur la littérature. De nouveaux critères de validation commençaient à émerger (RPS, CRPS, CRPSS, etc.) mais aucun consensus n'était clairement formulé. L'intérêt de cette étude résidait plus sur le fait d'être capable de proposer un encadrement de prévision et un outil de validation (ad-hoc) que de valider l'approche sur de nombreux sites en utilisant de nombreuses méthodes prédictives. Si la méthode fonctionne à Ajaccio, pour des données horaires, avec des arbres de régression et que l'on arrive à estimer la validité de l'encadrement, alors la méthodologie est très facilement transposable à d'autres configurations, sites, prédicteurs, etc. Il est très important de noter que la méthodologie « bootstrap » non paramétrique développée ici n'est basée sur aucune hypothèse de normalité, de linéarité, ou de stationnarité. Les seules contraintes sont de disposer d'un modèle de prévision déterministe basé sur l'apprentissage automatique et d'assez de données (au moins deux ans) pour que le processus d'échantillonnage soit cohérent. Comme nous allons le voir, outre l'encadrement de la prévision, nous proposerons aussi un outil d'estimation de sa pertinence (métriques spécifiques) que chaque prévisionniste pourra utiliser avec son propre modèle de prévision. Notre méthode permet de passer facilement d'une prévision déterministe à une prévision probabiliste par le biais d'échantillonnage et de la méthode suivie durant ces simulations.

Ce paragraphe sera structuré en en trois parties : nous commencerons par étudier l'horizon 1 heure, puis viendront les horizons plus lointains (2-6 heures) et enfin nous montrerons l'intérêt d'utiliser un modèle de connaissance pour borner l'intervalle de prévision et ainsi améliorer la méthodologie d'encadrement.

3.1. Application à un horizon de 1 heure et comparaison des performances des modèles

Avant d'envisager la prévision probabiliste, il est nécessaire de mettre en place une méthodologie de prévision déterministe robuste. Dans cette étude de l'horizon 1 heure, nous avons utilisé la méthodologie décrite précédemment (pré-traitement, stationnarisation, entrainement des modèles et prévision) pour les modèles ARs, ARel, ARr, ARen, PI et P. Puis pour chaque modèle nous avons appliqué la méthode d'échantillonnage pour générer une prévision probabiliste. Un exemple de cette dernière est représenté sur la *Figure IV-15*.

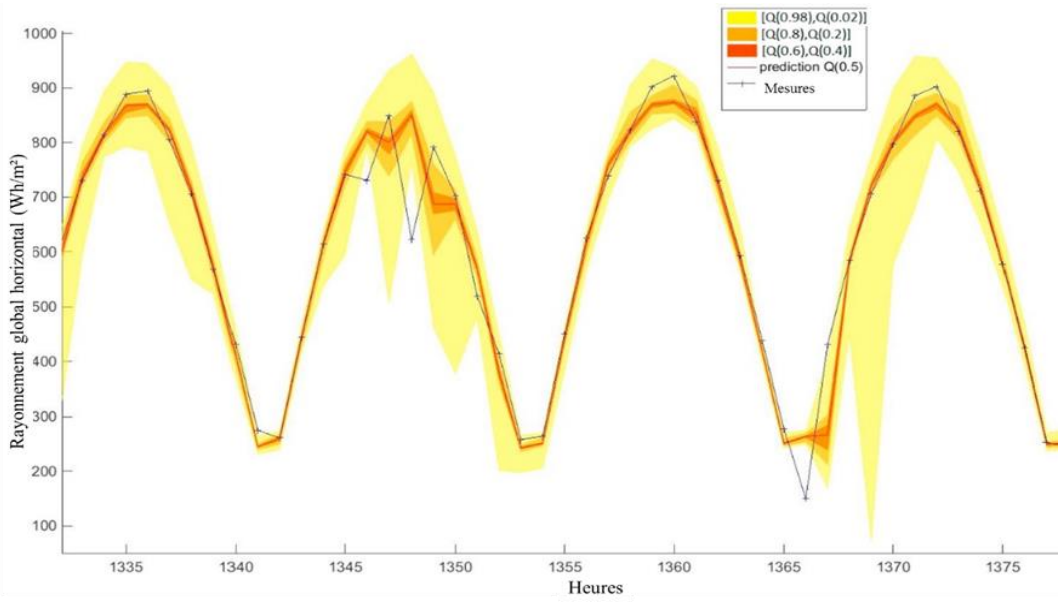


Figure IV-15: Encadrement de la prévision pour un horizon de 1 heure sur les données d'Ajaccio pour le modèle ARs

Pour chaque modèle, les courbes obtenues sont équivalentes, nous ne les exposerons pas ici étant donné qu'elles ne procurent aucune valeur ajoutée concernant les conclusions que l'on peut tirer. Sur cette courbe, nous voyons que suivant les quantiles considérés, la taille de l'encadrement va différer mais aussi que les prévisions seront plus ou moins pertinentes. L'intervalle le plus grand montré sur cette figure (compris entre le $2^{ème}$ et le $98^{ème}$ centiles) est bien trop large et sans pour autant coller systématiquement avec la mesure (Cf $1369^{ème}$ heure). L'intervalle le plus petit (compris entre le $4^{ème}$ et le $6^{ème}$ décile) est bien trop étroit et ne colle pour ainsi dire quasiment jamais avec la mesure. De ces deux constats, nait la notion de risque ! En effet, le prévisionniste propose un encadrement suivant un risque qu'il aura au préalablement quantifié et de cela dépendra la construction de l'intervalle de prévision. Pour estimer ce risque et de ce fait le nombre de quantiles à considérer au moment de la construction de l'intervalle, il est nécessaire de mettre en place des paramètres objectifs de validation. Si actuellement il existe de nombreux outils de comparaison, au moment de notre étude peu étaient présents dans la littérature. Nous sommes partis des paramètres plus généraux, qui sont le MIL et le PICP, pour établir un nouveau critère de jugement de la pertinence de l'intervalle de prévision (Γ_{min}).

Il convient de garder à l'esprit, que quel que soit le critère de validation ou la méthode prédictive retenus, un bon modèle de prévision minimise son MIL et maximise son PICP. C'est fort de ce constat, que nous avons mis en place le paramètre Γ_{min} ainsi que le test statistique du gamma qui permet de valider l'encadrement. Dans le *Tableau IV-12* les valeurs calculées de ces différents paramètres sont rassemblées pour six modèles différents, la persistance et la persistance intelligente ainsi que quatre modèles basés sur les arbres de régression, simples, élagués, renforcés et ensachés. Pour chaque arbre de régression, nous montrons les résultats liés aux prévisions déterministes (notées « classique » dans le tableau et issues de la section précédente) ainsi que ceux liés aux prévisions basées sur l'estimation du quantile Q (0,5) de l'approche probabiliste.

Tableau IV-12: Comparaison des modèles de prévision pour un horizon de 1 heure (en gras les meilleures valeurs)

		nRMSE	***Γ_{min}***	***MIL (Wh.m^{-2})***	***PICP (%)***
Persistance	Simple	0,3040	-	-	-
	Intelligente	0,1914	-	-	-
ARs	Classique	0,2426	-	-	-
	Q(0,5)	0,1858	0,91	148,50	**69,77%**
ARel	Classique	0,1858	-	-	-
	Q(0,5)	0,1843	1,01	159,71	65,55%
ARr	Classique	0,1863	-	-	-
	Q(0,5)	0,1832	1,01	160,17	65,37%
ARen	Classique	0,1847	-	-	-
	Q(0,5)	**0,1815**	**0,88**	**134,24**	68,82%

On note que lorsque l'on réalise une prévision ponctuelle à partir d'une prévision probabiliste par le biais du quantile médian Q (0,5), on améliore systématiquement la nRMSE du modèle, la diminution la plus importante étant pour le modèle ARs pour lequel on gagne 5,7 points d'erreur. Pour les autres modèles, l'amélioration n'est pas aussi importante mais on gagne néanmoins entre 0,15 à 0,30 point de pourcentage. De plus, l'approche probabiliste rend les méthodes d'apprentissage automatique systématiquement meilleures que la persistance intelligente avec une amélioration de la nRMSE comprise entre 0,5 et 1 point de pourcentage. Le MIL, autrement dit la largeur de la bande, doit être la plus faible possible tout en garantissant une bonne fiabilité de l'intervalle de prévision (PICP élevé). ARen possède le MIL le plus faible (134,2 Wh.m^{-2}) alors que ARel possède le PICP le plus élevé (69,8%). Il faut tout de même modérer cette conclusion compte tenu du faible écart entre les PICP de tous les modèles (écart maximal inférieur à 5 points de pourcentage). Le paramètre Γ_{min} est un outil de comparaison complémentaire qui permet de valider ARen comme étant le meilleur modèle. En effet, ce dernier permet d'obtenir un Γ_{min} le plus faible possible, ce qui signifie qu'il permet le meilleur compromis entre largeur de bande de prévision (faible MIL) et fiabilité de l'encadrement (fort PICP). Pour chaque modèle, cinquante intervalles sont considérés, chacun correspondant à un nombre de centiles différent (n). La grandeur Γ_{min} est liée au nombre de centiles (n_0) qui minimise Γ_n comme le montre l'équation IV-1 (voir Chapitre 3, partie 3.2).

$$\begin{cases} n_0 = argmin_n(\Gamma_n) \\ \Gamma_{min} = \Gamma_{n_0} \end{cases} \quad \text{IV-1}$$

Dans la *Figure IV-16* sont présentés pour les 50 valeurs d'intervalle possibles (une marque pour chaque intervalle et chaque modèle), les MIL et PICP correspondant. Les paramètres Γ_n peuvent facilement être interprétés à partir de cette figure. En effet, ils peuvent être exprimés comme étant la distance qui relie chaque marque à l'origine du repère. De ce point de vue Γ_{min} n'est autre que la marque située le plus près de l'origine. L'ellipse qui sert de base à la construction du test d'hypothèse du gamma est représentée sur cette figure, ainsi seules les marques situées à l'intérieur de l'ellipse correspondent à un « bon » encadrement au sens du test du gamma et des tolérances que l'on s'est fixées.

Les interprétations du *Tableau IV-12* et de la *Figure IV-16* nous permettent de conclure que seuls deux modèles (basés sur les arbres de régression) réussissent le test du gamma : il s'agit de ARs et ARen pour lesquels les valeurs minimales de Γ sont respectivement de 0,91 et 0,88. Les marques les plus proches de l'origine du repère défini par le MIL et (100-PICP) (*Figure IV-16*) indiquent les meilleurs intervalles à considérer (fort PICP et faible MIL). Pour les deux meilleurs modèles (Ars et ARen), les meilleures configurations sont obtenues pour des n compris entre 35 et 45.

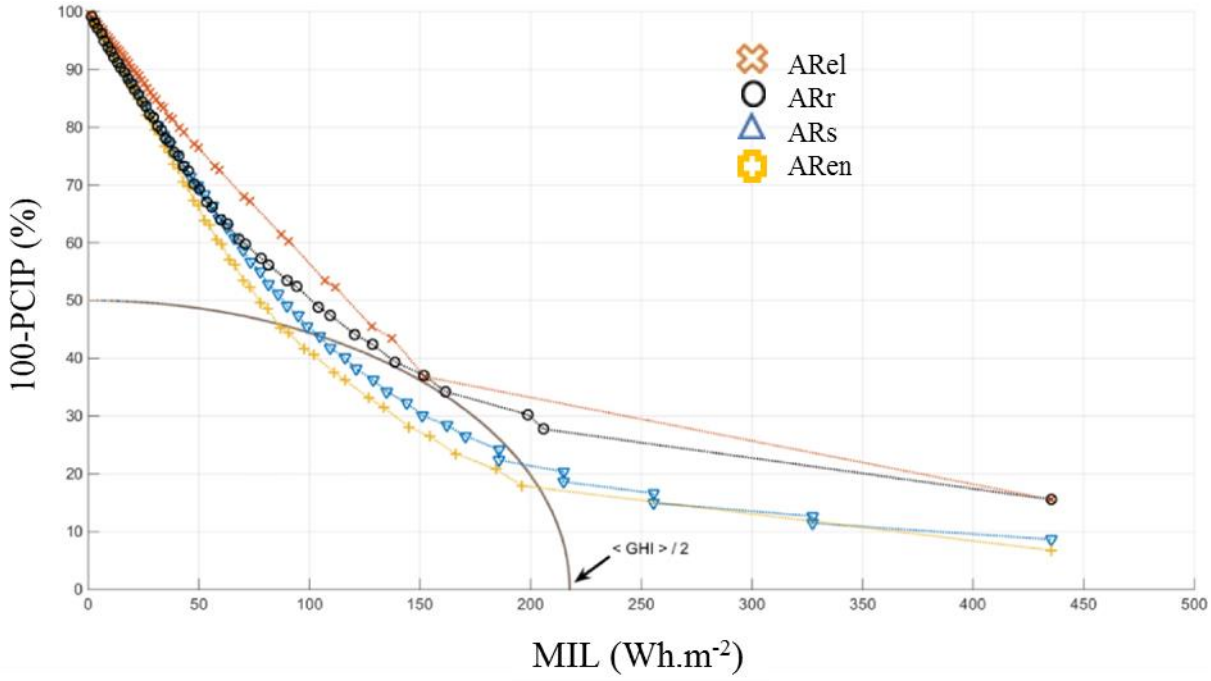


Figure IV-16: MIL et PICP pour chaque modèles et chaque type d'encadrement des prévisions

Nous allons maintenant focaliser notre analyse sur les horizons de prévision compris entre 1h et 6h.

3.2. Résultats sur l'influence de l'horizon de prévision

Dans cette partie nous allons nous intéresser particulièrement au modèle d'arbres de régression ensachés car dans la section précédente concernant l'horizon 1 heure, ce modèle proposait les plus faibles nRMSE, Γ et MIL. Concernant la valeur du PICP, même si elle n'est pas la plus élevée, elle est très proche du meilleur résultat. Le *Tableau IV-13* et la Figure IV-17 présentent les valeurs de l'indice gamma, du MIL, du PICP et du skill score pour les arbres de régression ensachés et concernant les horizons de prévision inférieurs à 6 heures (6h/1h). Le MIL normalisé (nMIL) correspond à la division du MIL par la valeur moyenne du rayonnement global (nMIL= MIL / < GHI >)

Tableau IV-13: Evolution des différents paramètres de l'intervalle en fonction de l'horizon de prévision pour le modèle d'arbres de régression ensachés

Horizon	Γ_{min}	*MIL (Wh.m^{-2})*	*nMIL*	*PICP (%)*	*Skill Score*	*nRMSE*
h+1	0,88	134,2	0,3136	68,82%	0,03	0,1847
h+2	1,38	102,2	0,2388	34,60%	-0,13	0,2998
h+3	1,33	105,9	0,2474	37,67%	-0,01	0,3154
h+4	1,31	100,8	0,2355	38,76%	0,05	0,3249
h+5	1,31	104,5	0,2442	38,82%	0,10	0,3344
h+6	1,27	98,7	0,2306	40,28%	0,13	0,3385

Pour les horizons supérieurs à 2 heures, le meilleur Γ_{min} correspond à des MIL proches de 100Wh/m² alors que pour un horizon h+1, il correspond à 134 Wh.m^{-2}. Le MIL a tendance à diminuer avec l'horizon, tout comme le PICP. En effet, ce dernier diminue presque de moitié à l'horizon h+2. C'est une conséquence de la variabilité de l'indice de ciel clair (et du rayonnement global horizontal) et de la diminution de la prédictibilité pour les horizons lointains : la météorologie revêt un caractère fortement chaotique ! Notons également que le skill score est, pour h+2 et h+3, inférieur à 0, ce qui signifie que, pour ces horizons, la persistance intelligente est meilleure que les arbres de régression renforcés. Mais il ne faut pas oublier que la persistance intelligente ne permet pas la génération d'intervalles de prévision.

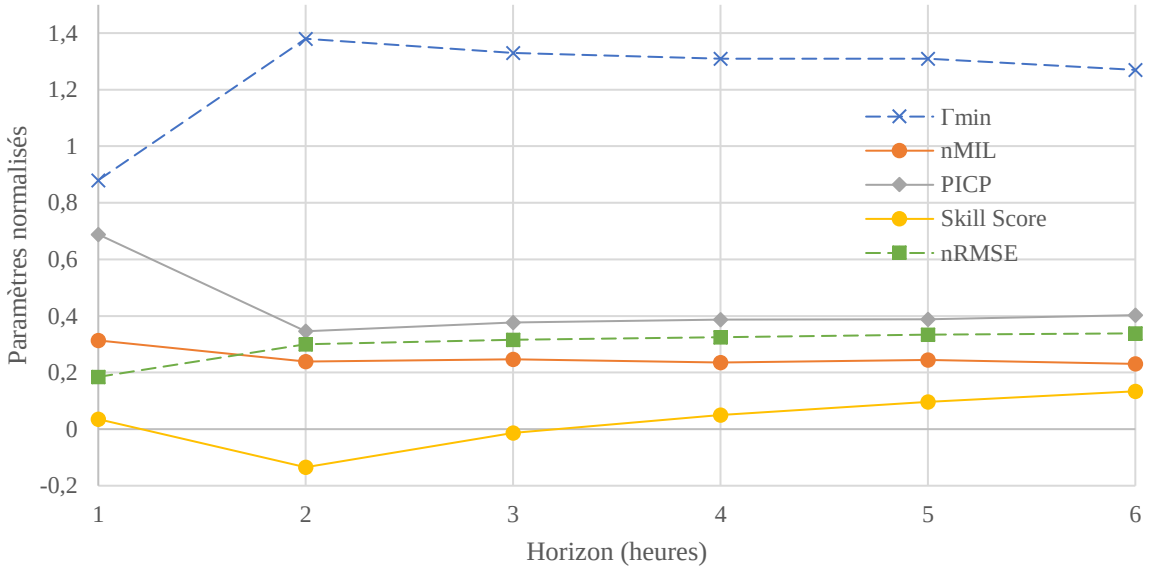


Figure IV-17: Représentation graphique des différents paramètres des intervalles de prévision (nMIL = MIL / < GHI >).

La Figure IV-17 montre de façon claire l'influence de l'horizon de prévision sur la taille de l'intervalle, le PICP, le skill score et le Γ_{min}. Pour l'horizon 1 heure ces paramètres présentent presque tous les valeurs les meilleures sauf pour le nMIL. En effet, ce dernier diminue dés h+1 (0,31) pour se stabiliser ensuite pour les autres horizons (~0,24), montrant ainsi que même pour des horizons lointains la méthodologie de prévision probabiliste ne donne pas des résultats aberrants et inexploitables. L'horizon h+2 est l'horizon de prévision le moins bien modélisé comme on peut le voir dans la figure IV-17 et le tableau IV-13, le skill score est le plus faible (-0,13) et le Γ_{min} (1,38) le plus élevé. Il n'y a pas réellement d'explication à ce phénomène, sans doute lié à la dynamique des occurrences nuageuses sur le site d'Ajaccio et à son orographie particulière. En résumé, cette étude montre qu'il est nécessaire de redéfinir les intervalles de prévision et le nombre de quantiles à considérer au moment de leur génération dès lors que l'on veut augmenter l'horizon de prévision. Ce n'est qu'avec cette étape supplémentaire que les bandes de prévision deviendront pertinentes.

3.3. Utilisation du modèle de connaissance (ciel clair) pour borner les intervalles de prévision

Il est possible d'améliorer les limites des prévisions probabilistes, en utilisant le modèle de ciel clair Solis (Mueller *et al.*, 2004). En effet, le rayonnement global horizontal mesuré est la plupart du temps inférieur à l'irradiation globale par ciel clair (I_g^{cs}) et supérieur à la composante diffuse de cette dernière (I_d^{cs}). Notons que cette amélioration peut s'appliquer quel que soit la méthodologie de génération des bandes de prévision. Le principe est simple, dès lors que la limite supérieure de l'encadrement dépasse le rayonnement global Solis, il suffit de la remplacer par (I_g^{cs}) et si la limite inférieure est en dessous de la composante diffuse du modèle Solis, on la remplacera par (I_d^{cs}). Ce bornage devrait en théorie permettre de diminuer le MIL tout en garantissant un PICP constant. Les valeurs physiquement impossibles à atteindre n'apportent qu'une augmentation de MIL et donc un manque de robustesse dans la méthodologie.

En théorie, cette modification de la prévision probabiliste (qui est une amélioration objective assez simple à mettre en œuvre) devrait permettre de diminuer l'indice gamma. En pratique, le fait d'utiliser le modèle Solis est très intéressant mais les incertitudes numériques du modèle Solis (Ineichen, 2006)

modifient légèrement le PICP (moins de 3% quel que soit l'horizon considéré). Le résultat de cette amélioration est présenté dans le *Tableau IV-14* pour le modèle basé sur les arbres de régression ensachés et un horizon allant de 1 heure à 6 heures.

Tableau IV-14: Impact de l'amélioration par le modèle Solis sur les intervalles de prévision en fonction de l'horizon

	Sans le modèle Solis		**Avec le modèle Solis**			
Horizon	Γmin	**MIL (Wh.m^{-2})**	**Γmin**	**Gain (%)**	**MIL (Wh.m^{-2})**	**Gain (%)**
h+1	0,88	134,2	0,74	15,9	113,5	15,4
h+2	1,38	102,2	1,14	17,4	84,0	17,8
h+3	1,33	105,9	1,07	19,2	86,7	18,1
h+4	1,31	100,8	1,04	20,6	80,7	19,9
h+5	1,31	104,5	1,03	21,2	82,2	21,3
h+6	1,27	98,7	0,95	24,7	72,5	26,5

Pour l'horizon 1 heure et une même valeur du PICP, le MIL est réduit de 15,4% en utilisant le modèle Solis comme limite des bandes de prévision. La valeur du gamma est fortement modifiée passant de 0,88 à 0,74, elle est donc améliorée de presque 16%. Pour l'horizon 6 heures, la valeur du MIL est améliorée de 26,5% et celle de Γmin de 24,7%. On constate que le gain généré par l'utilisation du modèle de connaissance augmente avec l'horizon de prévision (de ~15% à ~25% pour Γmin et MIL). Nous pouvons alors souligner que même si le modèle SOLIS apporte une petite incertitude, celle-ci est largement acceptable état donné le gain observé sur la largeur de bandes et le gamma. Les conclusions sont identiques quel que soit l'horizon considéré, ce qui permet de les généraliser et de valider les effets de l'utilisation du modèle de connaissance dans la méthodologie de prévision probabiliste.

Notons que pour l'horizon 1 heure, si l'on considère une large bande définie uniquement avec les modélisations par ciel clair (modèle Solis) du rayonnement global (I_g^{cs}) et diffus (I_d^{cs}), le MIL est égal à 408,87 Wh.m^{-2} (nMIL = 95,5%), le PICP est proche de 100% (toutes les valeurs mesurées sont comprises entre les limites définies par le global et le diffus par ciel clair) et l'indice gamma vaut 0,94.

3.4. Synthèse sur la prévision probabiliste

La prévision probabiliste est devenue en météorologie et plus particulièrement en prévision de la ressource solaire, un domaine dont l'intérêt est de plus en plus plébiscité. Contrairement à la prévision déterministe, le mode probabiliste permet la mise à disposition de bandes et de niveaux de confiance de prévision qui est largement appréciée par les prévisionnistes. Les nouveaux enjeux et les difficultés liées à la gestion de plus en plus poussée des flux énergétiques conduisent les gestionnaires de réseau à adopter ce type d'approche.

Au cours de ces travaux nous avons voulu développer une méthode robuste et objective de jugement de la pertinence des intervalles de prévision, et l'appliquer sur des modèles d'apprentissage automatique basés sur les arbres de régression. La particularité de notre méthode réside dans le fait qu'elle est

applicable à n'importe quel modèle de prévision basé sur l'apprentissage automatique, permettant ainsi de passer d'une prévision déterministe à une prévision probabiliste. Les modèles que nous avons choisi d'évaluer sont ceux qui nous ont paru les plus intéressants du fait de leurs bons résultats dans nos premiers travaux (CF Chapitre 4 Paragraphe 2). Ces modèles, construits à partir des arbres de régression (ARs, ARel, ARr et ARen), ont pu être validés par comparaison croisée avec le modèle naïf P et le modèle de référence PI.

Dans un premier temps nous avons comparé les performances des modèles concernant l'horizon 1 heure. Cette partie nous a notamment permis de nous familiariser avec cette nouvelle approche ainsi qu'avec les différentes métriques qui lui sont liées (MIL, PICP et Γ). Durant l'étude sur la prévision ponctuelle, on a pu noter que l'utilisation de prévisions probabilistes Q (0,5) améliorait les nRMSE de 5,7 points pour le modèle ARs et entre 0,15 et 0,30 point pour les trois autres modèles basés sur les arbres de régression (ARel, ARr et ARen). Les premiers résultats concernant les métriques MIL et Γ montrent que le meilleur modèle est celui basé sur les arbres de régression ensachés (ARen). Pour tous les modèles, les PICP sont comparables avec des différences inférieures à 5%. Nous avons ensuite réalisé une représentation graphique des différentes bandes de prévisions générées à partir de chaque prédicteur, dans un repère défini par (100-PICP) en ordonnée et le MIL en abscisse. Cette représentation est un outil assez pratique car elle permet de voir assez simplement quels sont les modèles qui réussissent le test du gamma. En effet, après avoir matérialisé une ellipse qui sépare le plan en deux zones puis en fonction de l'emplacement de chaque marque relative à chaque bande de prévision (intérieur ou extérieur de l'ellipse), il est simple de statuer quant aux résultats du test. Ici nous avons pu remarquer que deux modèles répondent à ces contraintes, ARs et ARen avec un nombre de centiles compris entre 35 et 45.

Nous avons voulu ensuite observer l'influence de l'horizon de prévision sur la taille des intervalles. Pour ce faire, nous avons gardé le meilleur modèle pour l'horizon de 1 heure (ARen) et nous avons calculé les valeurs de différentes métriques en fonction de l'horizon : Γ_{min} MIL, PICP, skill score et nRMSE. Concernant le skill score et la nRMSE, les tendances sont celles que nous avons remarquées jusqu'à maintenant, la nRMSE augmente avec l'horizon et le skill score diminue entre 1 et 2 heures puis remonte jusqu'à 6 heures. Les autres métriques propres à la prévision probabiliste diminuent entre 1 et 2 heures et se stabilisent pour les horizons plus lointains. Cela signifie en d'autres termes qu'il est nécessaire de redéfinir les intervalles de prévision par le biais du nombre de quantiles à considérer dès lors que l'horizon de prévision est modifié.

Enfin nous avons voulu montrer le bénéfice de la limitation des intervalles de prévision par l'utilisation de la modélisation par ciel clair de l'irradiation solaire. Il apparait que les gains pour Γ et MIL sont significatifs et augmentent avec l'horizon. Pour le Γ le gain est de 15,9% à 1 heure et 24,7% à 6 heures et pour le MIL il est de 15,4% à 1 heure et 26,5% à 6 heures. Il ressort que même si la modélisation par ciel clair possède une erreur intrinsèque liée au site sur lequel elle est appliquée, celle-ci est largement acceptable vu la diminution du MIL constatée.

4. Prévision horaire pour les composantes directe et diffuse

4.1. Introduction

Pour certaines applications solaires, il est souhaitable, voire parfois nécessaire, de connaître les différentes composantes du rayonnement solaire. Il nous est donc apparu intéressant de regarder les performances de certaines méthodes de prévision développées au préalable aux données de rayonnement direct et diffus sur le site d'Odeillo pour lequel nous disposions des données nécessaires et qui nous

semblait le plus intéressant du fait de la plus grande difficulté de prévision. Cette partie de nos travaux a été réalisé en collaboration avec deux chercheurs de l'Université Houari Boumédiène d'Alger, Lamara Benali et Rabah Dizene.

Si les méthodes de prévision ont largement été étudiées et appliquées au rayonnement global, elles l'ont été très peu pour le rayonnement direct normal et pratiquement jamais pour le rayonnement diffus horizontal.

Pour les systèmes à concentration solaire thermiques (appelé CSP pour « concentrated solar powerplant ») ou photovoltaïques (appelé CPV pour « concentrated photovoltaic »), c'est le rayonnement direct qui est converti en chaleur ou électricité. La prédiction de la composante directe est particulièrement importante pour la gestion et le contrôle des centrales thermiques (Chaturvedi, 2016). Les CSP ont un avantage considérable par rapport aux centrales photovoltaïques et éoliennes de pouvoir utiliser un stockage d'énergie thermique (Kraas *et al.*, 2013; Lara-Fanego *et al.*, 2012). Il a été souligné dans le chapitre 3 que très peu d'articles scientifiques traitaient de la prédiction du rayonnement direct.

L'éclairement diffus horizontal est quant à lui moins utilisé dans la modélisation des systèmes solaires, mais sa connaissance est nécessaire dans la modélisation der certains phénomènes thermiques dans le bâtiment et pour calculer le rayonnement solaire reçu par les systèmes thermiques ou photovoltaïques qui modifient leur azimut et leur inclinaison au cours de la journée (systèmes de suivi); l'éclairement global reçu par ces capteurs solaires doit être calculée à partir des composantes directe et diffuse en utilisant divers modèles d'éclairement global dans des plans inclinés.

4.2. Données et modèles utilisés

Les données utilisées ici proviennent du laboratoire PROMES CNRS (UPR 8521 Odeillo, Pyrénées Orientales, France). De la même manière que nous avons essayé de quantifier la variabilité de la série temporelle de rayonnement global (Cf Chapitre 4), celle des séries de rayonnement direct et diffus a été estimée en utilisant le MALR. Les résultats sont consignés dans le Tableau IV-15.

Tableau IV-15: Variabilité des composantes globale, directe et diffuse du rayonnement pour les données d'Odeillo

Composantes	*Globale*	*Directe*	*Diffuse*
MALR	0,5028	0,9945	0,4732
var	23,5%	46,6%	22,2%

On constate que les données de rayonnement solaire direct ont une variabilité largement supérieure à ce que nous avons pu observer pour le rayonnement global et le rayonnement diffus ce qui laisse à penser qu'elles devraient être plus difficiles à prévoir.

Le protocole expérimental que nous avons suivi pour réaliser ces prévisions est le même que pour les prévisions d'énergie globale horizontale (décrit dans le Chapitre 2) mais en effectuant une stationnarisation après avoir préalablement calculé les composantes diffuses et directes par ciel clair. Pour illustrer la validation des modèles d'éclairement direct et diffus par ciel clair, nous présentons sur la *Figure IV-18* pour une journée d'Avril et de Septembre l'éclairement solaire mesurée et modélisé par le modèle SOLIS. L'éclairement diffus par ciel clair se situe toujours au-dessous de l'éclairement mesuré car la composante diffuse est toujours minimum lors de belles journées et maximale par journées à ciel nuageux.

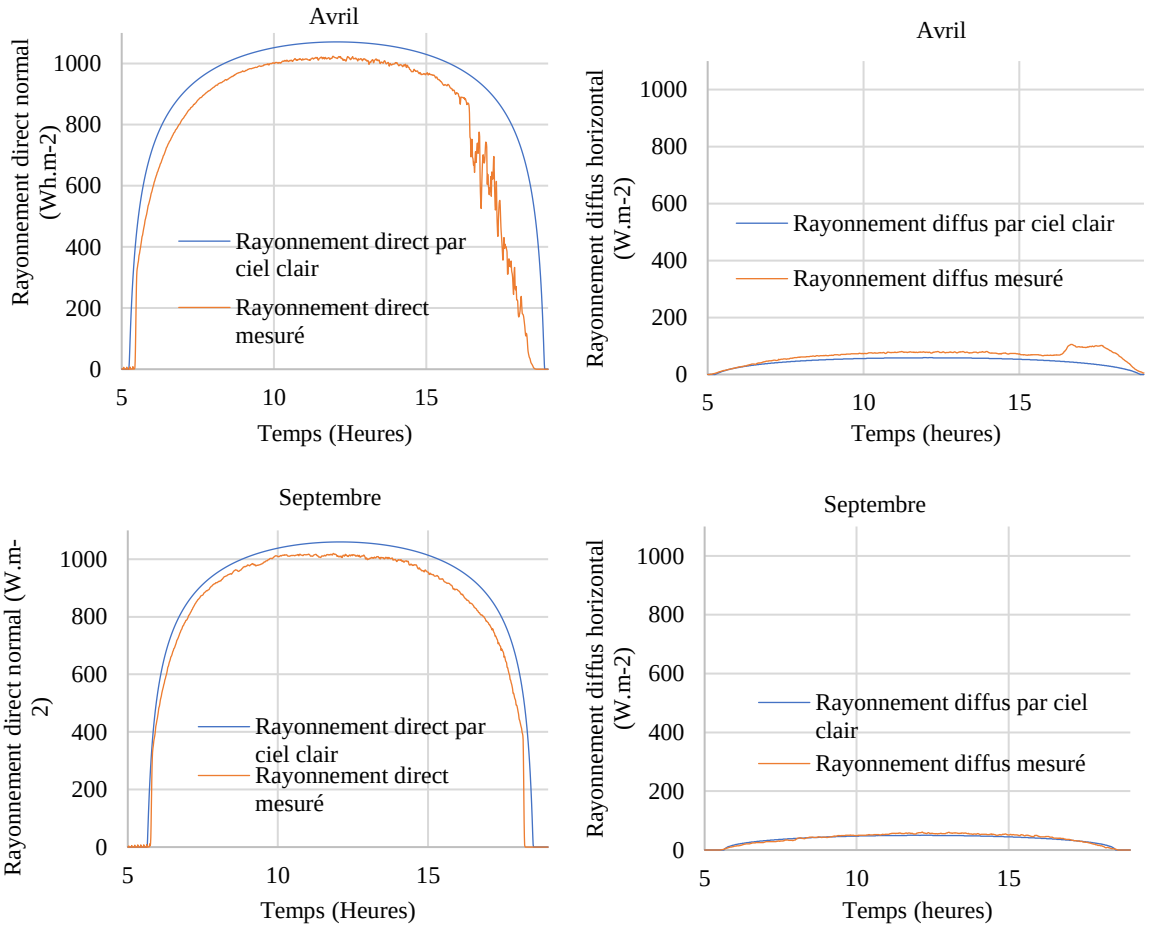


Figure IV-18: Représentation des modélisations par ciel clair et des données mesurés pour deux journées, rayonnement direct et diffus

On note que le modèle SOLIS, validé sur plusieurs mois de données, donne des résultats satisfaisants également pour les composantes directes et diffuses, même si nous avions noté des variations importantes des paramètres météorologiques qui entrent dans le modèle au cours d'une même journée et d'un mois à l'autre.

Le nombre d'entrées des modèles est déterminé de la même façon que pour la prévision du rayonnement global par le calcul de l'information mutuelle et les valeurs obtenues sont présentées dans le Tableau IV-16.

Tableau IV-16: nombre de données d'entrées pour les rayonnements direct et diffus

Horizon (h)		*h+1*	*h+2*	*h+3*	*h+4*	*h+5*	*h+6*
Nombre D'entrées (i)	Rayonnement direct	3	3	5	5	6	7
	Rayonnement diffus	6	7	9	10	9	10

Nous avons choisi d'utiliser trois modèles de prédiction différents lors de ces simulations. Notre choix s'est porté sur le modèle de persistance intelligente (PI) qui malgré sa simplicité donne des résultats souvent corrects et nous servira de référence, les réseaux de neurones artificiels avec le perceptron multicouche (PMC) et les forets aléatoires (FA). Ce choix nous permettra d'avoir une comparaison entre trois grands types de modèles, un modèle naïf, un modèle d'apprentissage automatique « classique » et un modèle d'apprentissage d'ensemble basé sur les arbres de régression.

4.3. Résultats des prévisions

Cette partie présente les résultats des prévisions pour des horizons de 1 à 6 heures des énergies solaires horaires directes et diffuses pour le site d'Odeillo obtenus sur la base annuelle puis pour chaque saison.

4.3.1. Performances annuelles

Nous présenterons successivement les résultats obtenus pour les prévisions du rayonnement direct et diffus.

- Prévision du rayonnement direct normal (var=46,6%)

Le Tableau IV-17 rassemble les résultats des prévisions du rayonnement direct normal, pour des horizons de 1 à 6 heures par pas de temps horaire. Nous avons surligné en vert les meilleurs résultats pour chaque indice d'erreur.

Tableau IV-17: Résultats des performances annuelles pour la prévision de l'éclairement direct normal

RMSE (Wh.m^{-2})	*h+1*	*h+2*	*h+3*	*h+4*	*h+5*	*h+6*
PI	207,86	287,64	338,86	378,19	412,67	434,51
PMC	212,33	270,1	297,73	321,59	336,98	344
FA	189,5	223,68	242,52	254,07	265,38	272,84
nRMSE						
PI	37,42%	51,77%	60,98%	68,04%	74,23%	78,16%
PMC	38,23%	48,62%	53,58%	57,85%	60,61%	61,88%
FA	34,12%	40,26%	43,64%	45,71%	47,73%	49,08%
MAE (Wh.m^{-2})						
PI	125,24	187,12	230,34	266,01	298,23	317,71
PMC	168,14	223,27	244,91	274,6	283,68	299,16
FA	141,65	175,16	194,23	207,24	216,71	226,19
nMAE						
PI	22,55%	33,68%	41,45%	47,85%	53,64%	57,15%
PMC	30,27%	40,19%	44,07%	49,40%	51,02%	53,82%
FA	25,50%	31,53%	34,95%	37,28%	38,98%	40,69%

La prévision de l'énergie solaire directs est difficile en particulier parce qu'il est plus sensible aux conditions météorologiques que l'énergie solaire globale et que les variations de l'éclairement direct fluctuent plus rapidement et avec une plus grande amplitude (pouvant passer de 1000 W.m^{-2} à 0 W.m^{-2} en quelques secondes lors d'un passage nuageux). Le modèle FA est le plus performant quel que soit l'horizon temporel et l'écart en termes de nRMSE entre le perceptron multicouche et les forets aléatoires passe de 4,11% à h+1 et 12,8% à h+6. Ce résultat confirme le fait que les modèles basés sur l'apprentissage d'ensemble sont plus performants dès lors que la variabilité est importante. Pour avoir une vue plus globale des performances, nous avons tracé sous forme d'histogrammes la nRMSE et la RMSE pour chaque modèle en fonction de l'horizon de prévision. Ce graphique est représenté sur la Figure IV-19 les trois histogrammes de gauche pour la RMSE et les trois de droite pour la nRMSE.

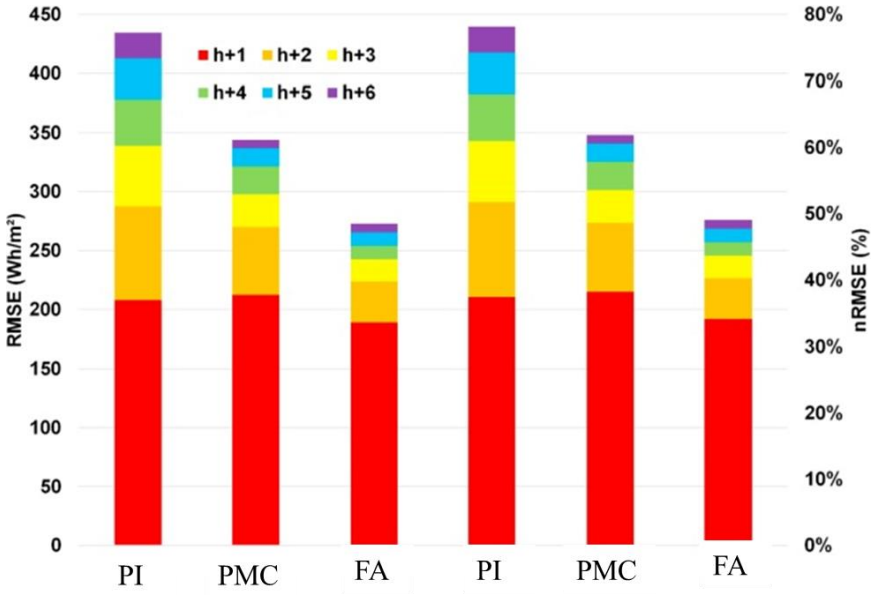


Figure IV-19: RMSE et nRMSE pour la prévision de l'éclairement direct normal

Même si modèle FA est toujours le meilleur quel que soit l'horizon, les niveaux d'erreur restent élevés par rapport à ceux obtenus pour le rayonnement global et confirment la difficulté de prévoir ce type de données.

- Prévision de l'éclairement diffus horizontal (var=22,2%)

Le Tableau IV-18 consigne les résultats des calculs d'erreur pour la prévision de l'éclairement diffus horizontal pour lesquels les meilleures valeurs ont été surlignées en vert.

Tableau IV-18: Performances annuelles pour la prévision de l'éclairement diffus horizontal

RMSE (Wh.m^{-2})	*h+1*	*h+2*	*h+3*	*h+4*	*h+5*	*h+6*
PI	87,34	96,51	104,02	110,3	114,16	115,35
PMC	56,77	70,28	79,71	84,58	86,19	88,49
FA	48,52	57,97	63,12	66,02	67,59	68,12
nRMSE						
PI	63,64%	69,69%	75,10%	79,62%	82,41%	83,27%
PMC	40,99%	50,75%	57,55%	61,06%	62,22%	63,88%
FA	35,08%	41,86%	45,57%	47,66%	48,79%	49,14%
MAE (Wh.m^{-2})						
PI	76,29	80,55	84,09	86,95	88,87	89,6
PMC	40,61	52,28	60,52	64,28	65,06	66,41
FA	33,64	41,00	44,75	47,73	48,86	50,11
nMAE						
PI	55,08%	58,16%	60,71%	62,77%	64,15%	64,68%
PMC	29,32%	37,75%	43,70%	46,40%	46,97%	47,94%
FA	24,29%	29,60%	32,31%	34,45%	35,27%	35,72%

Comme pour le rayonnement direct, le modèle FA permet d'obtenir les meilleures performances. L'ordre de grandeur des nRMSE et nMAE est identique à celui obtenu pour le rayonnement direct à l'exception de celles relatives à la persistance intelligente (cette comparaison ne peut pas se faire sur les valeurs absolues de ces erreurs car l'amplitude maximale du rayonnement direct et diffus est très différente).

Les résultats en termes de RMSE et nRMSE sont présentés sur la Figure IV-20.

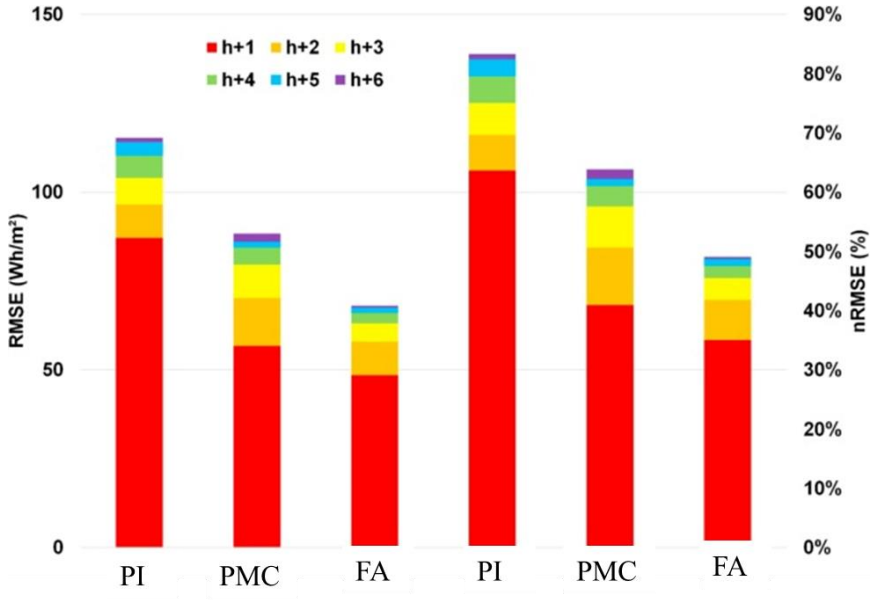


Figure IV-20: RMSE et nRMSE pour la prévision de l'éclairement diffus horizontal

La PI présente de mauvais résultats car le profil quotidien de l'éclairement diffus horizontal par ciel clair (pris en compte dans ce modèle) n'est pas aussi bien défini que pour les autres composantes (Ineichen, 2008). Comme on l'a vu pour l'éclairement direct normal, l'écart en termes de performances entre les forêts aléatoires et le perceptron multicouche augmente avec l'horizon de prévision, de 5,91% pour h+1 à 14,74% pour h+6. L'utilisation d'un modèle utilisant les FA pour l'éclairement diffus horizontal donne des résultats satisfaisants.

Nous avons tracé sur la Figure IV-21 les performances en termes de d'indice normalisé (nRMSE) pour les trois modèles et les deux composantes. La précision obtenue pour l'éclairement direct normal et pour l'éclairement diffus horizontal en utilisant les modèles basés sur le PMC et les FA est du même ordre de grandeur ; en revanche, la persistance intelligente ne semble pas adaptée à la prévision de ces deux composantes.

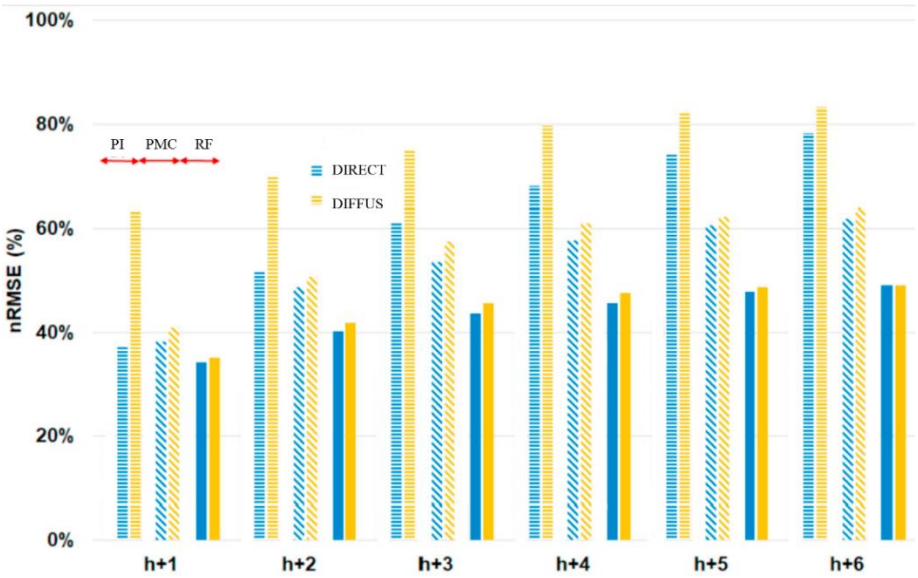


Figure IV-21: nRMSE pour les deux composantes de l'éclairement et pour les trois modèles étudiés (PI, PMC et FA) en fonction de l'horizon

Nous allons maintenant détailler les résultats des performances saison par saison pour les deux composantes.

4.3.2. Performances saisonnières

Il est intéressant d'observer les performances des modèles saison par saison afin de quantifier l'impact des conditions météorologiques (et de la variabilité) sur la précision des prévisions. Nous allons commencer par présenter les résultats liés à la prévision de l'éclairement direct normal puis nous présenterons les résultats concernant l'éclairement diffus horizontal.

- Rayonnement direct normal :

Le Tableau IV-19 présente les résultats des calculs d'erreur pour les différents modèles, pour chaque saison en fonction de l'horizon de prévision. Les valeurs du meilleur modèle pour chaque horizon ont été surlignées en vert.

Tableau IV-19: Erreurs pour la prévision du rayonnement direct normal en fonction de l'horizon de prévision pour chaque saison

		Hiver					
		h+1	h+2	h+3	h+4	h+5	h+6
RMSE (Wh.m^{-2})	PI	284,81	330,15	357,33	379,88	397,13	416,98
	PMC	203,59	261,11	293,39	320,65	339,72	349,55
	FA	163,79	203,59	225,05	250,11	253,51	260,60
nRMSE	PI	44,43%	51,48%	55,66%	59,10%	61,72%	64,78%
	PMC	31,76%	40,72%	45,70%	49,89%	52,79%	54,31%
	FA	25,55%	31,75%	35,06%	38,91%	39,40%	40,49%
MAE (Wh.m^{-2})	PI	185,72	227,74	248,42	264,73	278,48	296,34
	PMC	161,27	213,47	250,53	274,81	302,95	308,99
	FA	116,57	152,32	174,28	195,05	203,79	214,76
nMAE	PI	28,97%	35,51%	38,70%	41,19%	43,28%	46,04%
	PMC	25,16%	33,29%	39,02%	42,76%	47,08%	48,00%
	FA	18,19%	23,75%	27,15%	30,35%	31,67%	33,36%
		Printemps					
		h+1	h+2	h+3	h+4	h+5	h+6
RMSE (Wh.m-2)	PI	249,67	325,68	373,53	409,92	445	456,27
	PMC	219,97	282,03	312,08	335,08	353,58	363,10
	FA	200,92	239,69	256,26	272,37	284,34	287,05
nRMSE	PI	54,80%	71,44%	81,97%	89,98%	97,84%	100,48%
	PMC	48,28%	61,87%	68,49%	73,55%	77,74%	79,96%
	FA	44,10%	52,58%	56,24%	59,78%	62,52%	63,22%
MAE (Wh.m^{-2})	PI	168,81	236,38	278,9	312,54	344,78	353,16
	PMC	179,68	235,69	264,84	289,73	312,47	319,02
	FA	156,28	192,25	208,85	223,84	233,16	236,25
nMAE	PI	37,05%	51,85%	61,21%	68,60%	75,81%	77,78%
	PMC	39,44%	51,70%	58,12%	63,59%	68,70%	70,26%
	FA	34,30%	42,17%	45,83%	49,13%	51,26%	52,03%

		Eté					
		h+1	h+2	h+3	h+4	h+5	h+6
RMSE (Wh.m^{-2})	PI	231,14	297,74	339,43	371,09	390,02	396,63
	PMC	188,66	238,38	267,45	279,5	291,43	298,09
	FA	179,84	204,7	223,56	232,59	241,67	242,62
nRMSE	PI	36,55%	47,07%	53,71%	58,69%	61,60%	62,60%
	PMC	29,84%	37,68%	42,32%	44,20%	46,03%	47,05%
	FA	28,44%	32,36%	35,37%	36,79%	38,17%	38,29%
MAE (Wh.m^{-2})	PI	152,68	214,92	256,03	288,86	306,29	308,10
	PMC	145,93	190,11	218,15	229	249,3	250,21
	FA	136,37	159,86	177,36	182,9	192,89	195,14
nMAE	PI	24,15%	33,97%	40,51%	45,68%	48,38%	48,63%
	PMC	23,08%	30,05%	34,52%	36,22%	39,38%	39,49%
	FA	21,57%	25,27%	28,06%	28,93%	30,47%	30,80%
		Automne					
		h+1	h+2	h+3	h+4	h+5	h+6
RMSE (Wh.m-2)	PI	262,56	335,21	372,43	400,09	429,12	449,61
	PMC	233,60	296,86	321,12	347,35	363,17	371,82
	FA	207,66	240,33	261,45	274,00	288,86	295,78
nRMSE	PI	53,26%	68,06%	75,47%	81,15%	87,12%	91,33%
	PMC	47,38%	60,27%	65,07%	70,45%	73,73%	75,53%
	FA	42,12%	48,79%	52,98%	55,58%	58,65%	60,08%
MAE (Wh.m^{-2})	PI	168,81	236,38	278,90	312,54	344,78	353,16
	PMC	179,68	235,69	264,84	289,73	312,47	319,02
	FA	156,28	192,25	208,85	223,84	233,16	236,25
nMAE	PI	34,46%	47,39%	54,04%	59,32%	64,29%	67,54%
	PMC	38,75%	50,68%	56,13%	61,09%	65,80%	67,56%
	FA	32,19%	38,79%	43,78%	46,86%	50,27%	51,97%

Beaucoup d'informations sont données dans ce tableau, cependant nous constatons que le modèle basé sur les FA est le meilleur dans tous les cas et que le modèle de PI donne de mauvaises performances. Nous avons représenté dans la Figure IV-22 les résultats du calcul de la nRMSE par des histogrammes.

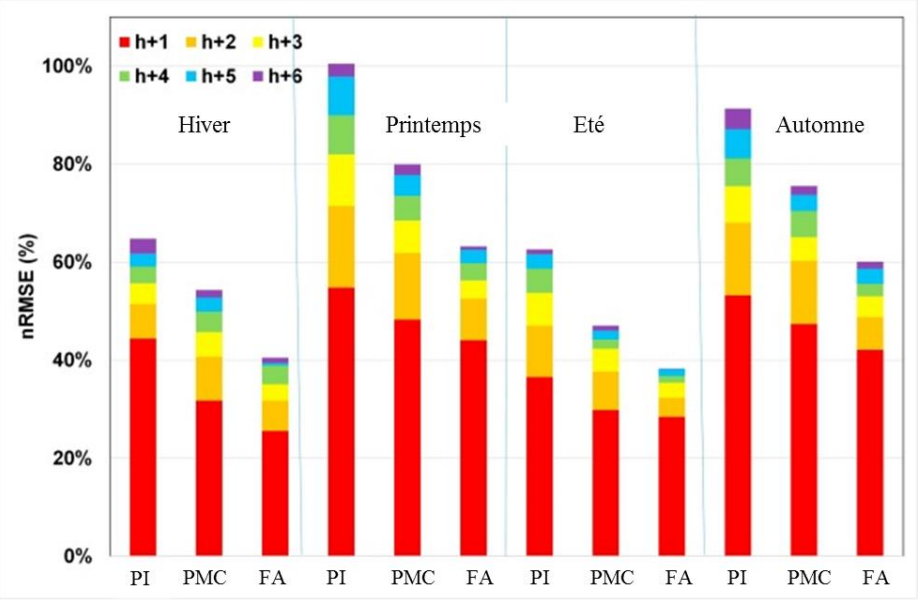


Figure IV-22: nRMSE pour la prévision de l'énergie solaire directe normale en fonction de l'horizon de prévision et des saisons

On note, en plus du fait que le modèle basé sur les forêts aléatoires est le meilleur modèle pour la prévision de l'énergie solaire directe, que la prévision est réalisée avec une plus grande précision en hiver et en été par rapport au printemps et à l'automne. Les saisons printemps et automne sont des périodes transitoires d'un point de vue climatique et constituent des moments de l'année pour lesquelles les conditions météorologiques sont très perturbées et variables : pour simplifier, en été il fait toujours beau et en hiver toujours mauvais alors qu'au printemps et en automne la succession du type de journée est beaucoup plus aléatoire.

- Eclairement diffus horizontal :

Le *Tableau IV-20* présente les résultats des calculs d'erreur pour les différents modèles, pour chaque saison en fonction de l'horizon de prévision. Les meilleurs résultats ont été surlignés en vert pour chaque horizon.

Tableau IV-20: Erreurs pour la prévision de l'énergie diffuse horizontale en fonction de l'horizon de prévision pour chaque saison

		Hiver					
		h+1	h+2	h+3	h+4	h+5	h+6
RMSE (Wh.m^{-2})	PI	74,37	82,28	88,58	93,93	96,64	98,04
	PMC	45,60	59,31	68,46	72,95	73,45	75,04
	FA	39,45	47,20	54,11	57,84	58,03	58,62
nRMSE	PI	68,17%	75,51%	81,32%	86,24%	88,75%	90,09%
	PMC	41,80%	54,43%	62,85%	66,98%	67,45%	68,96%
	FA	36,16%	43,31%	49,68%	53,10%	54,36%	55,01%
MAE (Wh.m^{-2})	PI	65,38	68,53	71,19	73,53	74,90	75,59
	PMC	33,36	44,85	53,42	56,94	57,23	58,00
	FA	26,57	33,31	37,55	41,39	42,01	43,16
nMAE	PI	59,93%	62,89%	65,37%	67,51%	68,78%	69,46%
	PMC	30,58%	41,16%	49,04%	52,28%	52,56%	53,29%
	FA	24,35%	30,57%	34,48%	38,00%	38,58%	39,66%
		Printemps					
		h+1	h+2	h+3	h+4	h+5	h+6
RMSE (Wh.m-2)	PI	102,49	114,88	127,25	137,95	143,89	145,72
	PMC	74,95	91,29	103,49	110,03	112,79	115,78
	FA	64,16	75,93	83,64	87,54	90,55	91,04
nRMSE	PI	55,00%	61,57%	68,12%	73,74%	76,97%	77,86%
	PMC	40,22%	48,93%	55,40%	58,81%	60,33%	61,86%
	FA	34,43%	40,70%	44,77%	46,79%	48,44%	48,64%
MAE (Wh.m^{-2})	PI	88,71	95,13	101,55	107,04	110,53	112,19
	PMC	54,65	67,85	77,57	82,92	84,94	86,33
	FA	46,16	54,84	60,14	64,19	65,00	65,28
nMAE	PI	47,60%	50,99%	54,36%	57,22%	59,12%	59,95%
	PMC	29,32%	36,36%	41,52%	44,32%	45,43%	46,13%
	FA	24,77%	29,39%	32,19%	34,31%	34,77%	34,88%

		Eté					
		h+1	h+2	h+3	h+4	h+5	h+6
RMSE (Wh.m^{-2})	PI	92,60	99,88	104,77	108,29	111,52	112,01
	PMC	51,52	63,06	70,43	74,01	75,25	77,74
	FA	44,22	53,39	55,98	56,72	57,71	60,89
nRMSE	PI	65,34%	70,56%	74,10%	76,66%	78,86%	79,33%
	PMC	36,35%	44,55%	49,82%	52,39%	53,21%	55,06%
	FA	31,20%	37,71%	39,60%	40,15%	40,81%	43,12%
MAE (Wh.m^{-2})	PI	83,82	87,25	89,53	91,11	92,51	92,81
	PMC	35,92	47,15	53,61	56,22	56,47	58,29
	FA	30,08	37,62	40,05	41,16	41,85	43,85
nMAE	PI	59,15%	61,64%	63,32%	64,50%	65,42%	65,73%
	PMC	25,34%	33,31%	37,92%	39,80%	39,93%	41,29%
	FA	21,22%	26,57%	28,33%	29,14%	29,60%	31,06%
		Automne					
		h+1	h+2	h+3	h+4	h+5	h+6
RMSE (Wh.m-2)	PI	76,84	85,56	90,95	95,31	98,30	99,19
	PMC	50,43	62,72	71,13	75,56	77,08	79,02
	FA	42,63	50,98	53,67	56,67	59,33	60,12
nRMSE	PI	65,74%	73,17%	77,77%	81,56%	84,07%	84,82%
	PMC	43,14%	53,64%	60,82%	64,65%	65,92%	67,57%
	FA	36,46%	43,60%	45,90%	48,49%	50,74%	51,04%
MAE (Wh.m^{-2})	PI	67,26	71,31	74,10	76,13	77,56	77,83
	PMC	38,51	49,26	57,49	61,02	61,62	63,01
	FA	31,74	38,23	41,27	44,17	46,49	47,04
nMAE	PI	57,53%	60,99%	63,36%	65,14%	66,33%	66,56%
	PMC	32,94%	42,13%	49,16%	52,22%	52,70%	53,88%
	FA	27,15%	32,70%	35,29%	37,79%	39,76%	40,37%

Sur la Figure IV-23 nous avons représenté la nRMSE en fonction des saisons et de l'horizon de prévision pour chaque modèle afin de mieux se rendre compte des performances. Les différences de performances entre les saisons sont bien moins flagrantes que pour la composante directe normale et le niveau d'erreur pour le modèle de PI est particulièrement élevé.

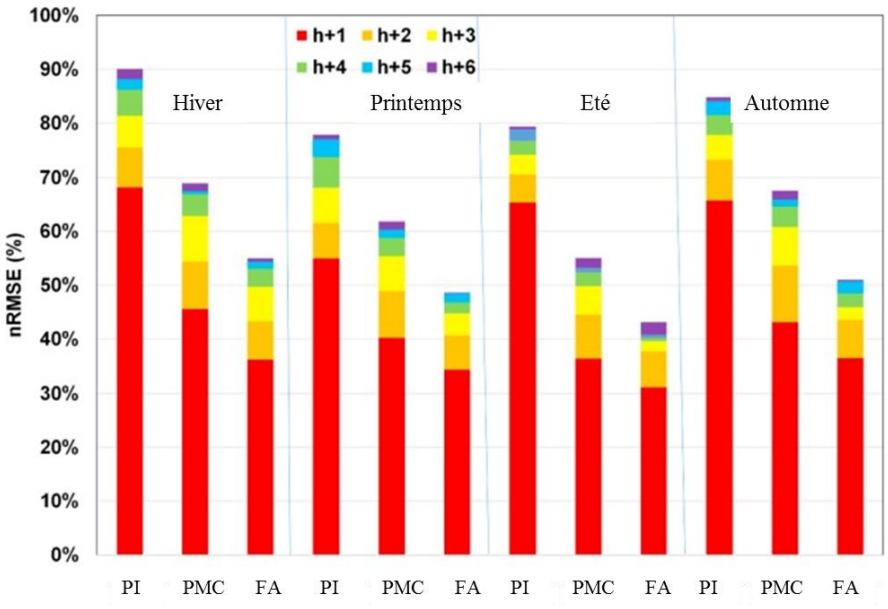


Figure IV-23: nRMSE pour la prévision de l'énergie directe normale en fonction de l'horizon de prévision et des saisons

4.4. Synthèse sur la prévision du rayonnement direct et diffus

Trois méthodes de prévision, la PI, le PMC et les FA, ont été comparées et testées sur des données solaires mesurées sur un site météorologique présentant une forte variabilité. L'objectif était de prévoir l'énergie solaire horaire pour un horizon temporel compris entre h+1 et h+6 ; ces méthodes ont été appliquées aux deux composantes : directe normale et diffuse horizontale.

La méthode FA permet d'effectuer des prévisions de ces deux composantes avec la meilleure précision des trois modèles testés :

- nRMSE de 34,11% pour h+1 et 49,08% pour h+6 pour l'énergie directe normale (var = 46,6%);
- nRMSE de 35,08% pour h+1 et 49,14% pour h+6 pour l'énergie diffuse horizontale (var = 22,2%).

L'amélioration engendrée par l'utilisation du modèle FA par rapport au PMC est d'autant plus importante que l'horizon de prévision augmente.

Le gain en termes de nRMSE lié à l'utilisation du modèle FA par rapport à l'utilisation de la PI est :

- Pour la composante directe normale de + 3,30% pour h+1 et +28,36% pour h+6 ;
- Pour la composante diffuse horizontale de + 28,56% pour h+1 et +34,13% pour h+6.

Une étude saisonnière a montré que les prévisions printanières et automnales sont plus difficiles à réaliser qu'en hiver et en été en raison d'une variabilité plus élevée du rayonnement solaire sur ces périodes.

Les composantes directes normales et diffuses horizontales sont plus difficiles à prévoir que le rayonnement global : ces deux composantes sont plus sensibles aux occurrences nuageuses. Le rayonnement global est une combinaison des composantes directe et diffuse, ainsi le fait d'additionner deux rayonnements qui ont tendance à varier en sens inverse, crée un effet de compensation et lisse les variations du rayonnement global résultant.

Le rayonnement direct normal est très variable, avec des vitesses et des amplitudes de variation très élevées ce qui complique encore plus sa prévision. Pour le rayonnement diffus horizontal, l'indice de ciel clair est toujours supérieur à 1 (le rayonnement diffus par ciel clair ayant une valeur inférieure à celle du diffus mesuré dans les autres conditions de ciel) et peut atteindre des valeurs élevées (contrairement à l'éclairement global et direct pour lesquels il reste compris entre 0 et 1), ce qui pourrait être la cause des résultats médiocres obtenus.

Il serait intéressant de valider ces conclusions sur d'autres sites soumis à diverses conditions météorologiques, mais il est difficile de trouver des séries chronologiques fiables d'irradiations solaires horaires, en particulier des composantes directes et diffuses.

5. Prévision du rayonnement global incliné : projet TILOS

5.1. Introduction

Dans cette partie, nous allons présenter les résultats obtenus lors les simulations effectuées dans le cadre des travaux du projet H2020 TILOS (détaillé dans le Chapitre 1). Ces travaux ont été menés dans le but de contribuer au développement du logiciel de pilotage et de gestion de la centrale hybride de l'île de Tilos. Chronologiquement, le cahier des charges ayant plusieurs fois changé nous avons dû répéter nos simulations et présenter de nouveaux résultats tout au long du projet :

- Au départ, il avait été convenu de prédire les énergies solaires globales inclinées horaires pour des horizons de 1 à 6 heures, la prédiction devant être réalisée chaque heure (6h/1h).
- À la demande du leader du groupe de travail sur la prédiction, il nous a été demandé de développer des modèles identiques pour une prévision de l'irradiance solaire au pas de temps de la minute pour un horizon de 1 à 10 minutes prédites chaque 10 minutes (10min/1min) et de prédire les puissances moyennes sur 15 minutes pour des horizons de 15 minutes à 120 minutes (soit 8 données 15 minutes prédites chaque 2 heures ; 2h/15min).
- Enfin, le pas de temps de 15 minutes pour la prédiction de 15 minutes à 2 heures a été remplacé par un pas de temps de 10 minutes (2h/10min) : il convenait donc de prédire chaque deux heures, les 12 puissances moyennes solaires sur 10 minutes suivantes.

Pour le gestionnaire du micro-réseau, il n'est pas possible de réajuster ces commandes toutes les minutes, voire tous les dix minutes, il a besoin de connaître suffisamment à l'avance les contraintes liées aux différents moyens de production. En effet, le temps de démarrage d'un groupe électrogène n'est pas instantané (environ 20 minutes) et donner l'ordre de démarrage trop tard peut nuire à la continuité de la production. Ces deux prédicteurs ont été introduits dans le système de gestion de l'énergie (EMS : Energy Management System) présenté sur la Figure IV-24.

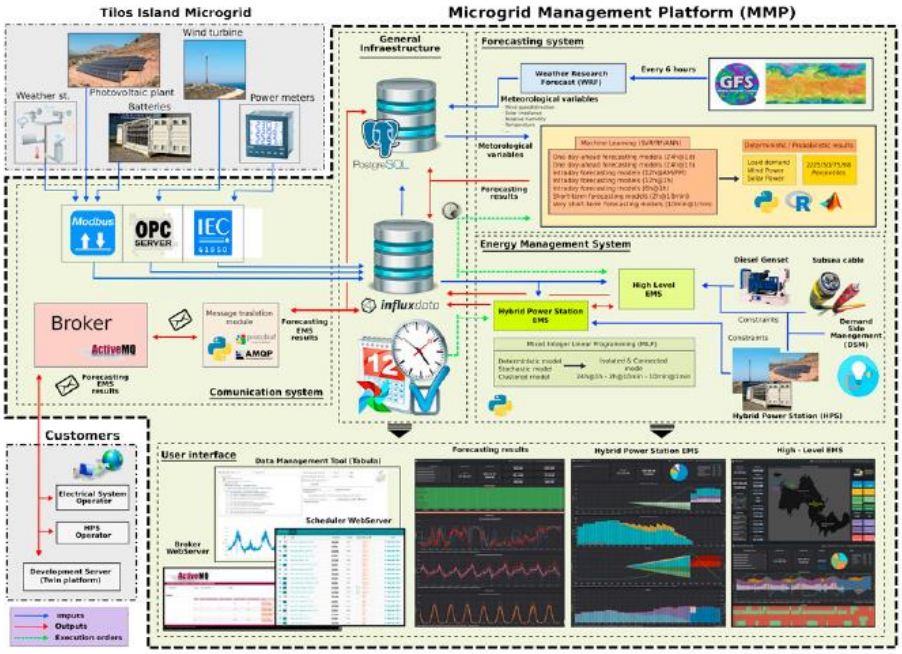


Figure IV-24: Description de l'architecture de la plateforme de gestion du micro-réseau.

Contrairement aux travaux présentés précédemment dans ce manuscrit, ces modèles ont été développés sur des mesures d'éclairement global incliné à 30° qui correspond à l'angle d'inclinaison des panneaux PV sur l'Ile de Tilos.

Nous avons donc dans un premier temps, suivi le même mode opératoire que celui présenté pour les données horaires de rayonnement global sur les 4 sites mais pour des horizons différents et des pas de temps différents. La seule différence au niveau du développement des modèles a été la nécessité d'adapter la stationnarisation aux nouvelles données en générant le rayonnement solaire global incliné par ciel clair. Une fois la sélection réalisée, ces modèles ont été adaptés aux contraintes de fonctionnement opérationnel en temps réel que nous détaillerons ensuite dans ce chapitre.

L'originalité de cette partie des travaux réside dans le fait qu'il a fallu développer et rendre opérationnel les modèles prédictifs en faisant en sorte qu'ils soient directement exploitables par le gestionnaire de réseau. Cette particularité nous a permis de mettre en évidence un grand nombre de contraintes auxquelles nous n'avions, auparavant, pas été confrontés :

- La nécessité que le modèle prédictif fonctionne en totale autonomie. En effet sur Tilos, les logiciels vont fonctionner en mode « online », ce qui implique que l'EMS doit recevoir les prévisions en temps réel et en permanence, même la nuit ;
- Pour améliorer les performances, les modèles d'apprentissage automatique doivent être « ré-entrainés » régulièrement (tous les mois environ) et ce réapprentissage doit se faire sans intervention humaine.

A l'heure ou ces lignes sont écrites, les modèles ont été implémentés et remplissent leur rôle au sein du logiciel de gestion de la centrale hybride.

Les résultats seront présentés dans l'ordre chronologique des demandes des partenaires : prévisions pour des horizons de 1 à 10 minutes par pas de temps de 1 minute (10min/1min), puis prévisions pour des horizons de 15 minutes à 2 heures par pas de temps de 15 minutes (2h/15min) et enfin les résultats des prévisions pour des horizons allant 10 minutes à 2 heures par pas de temps de 10 minutes (2h/10min).

5.2. Prévision pour un horizon de 1 à 10 minutes par pas de temps de 1 minute (10min/1min)

Ces horizons très courts ont été étudiés à la demande des partenaires du projet, même s'ils sont bien trop courts pour être réellement utilisables. En effet, les temps de réponse des différents dispositifs de production ne sont pas assez courts pour que l'on puisse réagir aussi rapidement même si l'on dispose de prévisions fiables. Cela n'enlève en rien à l'intérêt que nous avons eu à tester nos modèles et à confronter notre protocole expérimental à ce type de prévisions.

Nous présenterons nos résultats selon la même structure que celle utilisée dans la partie relative aux irradiations horaires horizontales.

Le Tableau IV-21 présente les résultats de l'auto information mutuelle pour cette granularité temporelle et permet de déterminer le nombre de données d'entrée pour chaque prédicteur.

Tableau IV-21: nombre de données d'entrée des modèles d'apprentissage automatique pour les prévisions du rayonnement global incliné sur Tilos (10min/1min)

Horizon (min)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Nombre d'entrées	3	4	4	6	6	5	6	9	10	10

La Figure IV-25 présente les valeurs de nRMSE obtenues pour chaque type de prédicteur et chaque horizon temporel.

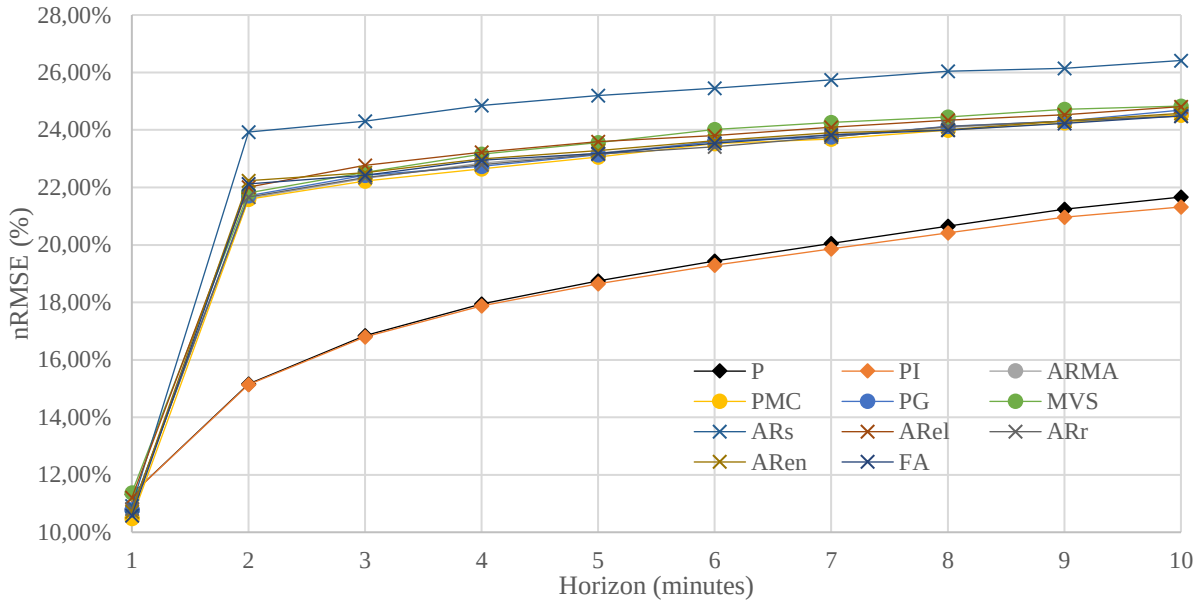


Figure IV-25: nRMSE en fonction de l'horizon pour les prévisions du rayonnement global incliné à Tilos (10min/1min)

Ce qui est particulièrement remarquable dans le cas de la prévision très court terme comme ici, c'est que les deux modèles naïfs (la persistance et la persistance intelligente) sont les meilleurs pour tous les horizons. Tous les modèles d'apprentissage automatique ont sensiblement les mêmes performances. Seul, le modèle d'arbres de régression simples se distingue par des performances moins élevées.

Les modèles P et PI surclassent largement les autres, sauf pour l'horizon de 1 minute pour lequel tous les prédicteurs se valent (moins de 1% de différence). Pour un horizon de 2 minutes l'écart entre le meilleur modèle (PI) et le « moins bon » (ARen), hormis Ars, est de 7%. Pour un horizon de 10 minutes, l'écart entre le meilleur et plus mauvais prédicteur, PI et ARel, atteint 3,5%.

Les résultats du skill score ne seront pas présentés, car cet indice est calculé par rapport à un modèle de référence, PI, et les résultats de nRMSE suffisent pour voir que tous les modèles sont ici moins performants que PI.

Les MBE obtenues pour la prédiction de l'irradiance solaire (plus précisément la densité de puissance exprimée en $W.m^{-2}$) chaque minute sont présentées sur la Figure IV-26.

Les résultats de la MBE sont assez similaires à ce que l'on a pu remarquer jusqu'à présent, les modèles P et PI sont très peu générateurs de biais et les autres ont une petite tendance à la surestimation, moins de 6 W.m-2, ce qui représente environ 1% d'erreur relative.

Ces résultats sont finalement assez logiques, en effet, l'éclairement global varie peu de façon instantanée (contrairement à l'éclairement direct), et ainsi les modèles simples peuvent donner des résultats satisfaisants à très court terme, contrairement à des horizons plus lointains comme constaté précédemment. De plus, la dynamique de l'occurrence nuageuse est largement supérieure à 1 minute, il suffit de regarder le ciel pour se rendre compte que les nuages bougent très peu en si peu de temps.

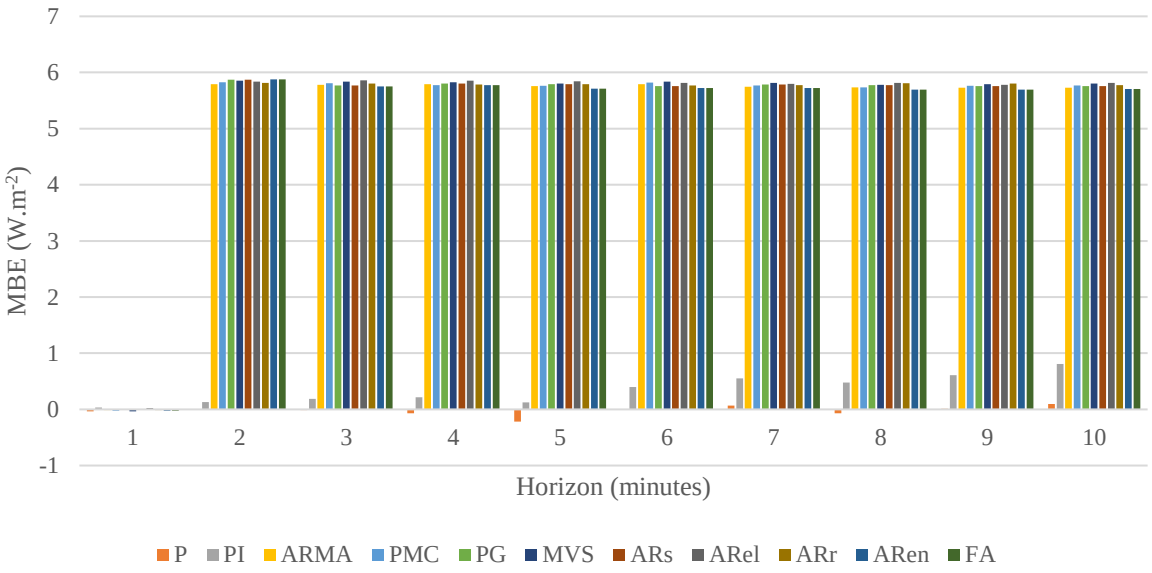


Figure IV-26: MBE en fonction de l'horizon pour les prévisions du rayonnement global incliné à Tilos (10min/1min)

D'un point de vue opérationnel, il est évident que nous ne pouvons pas changer de modèle pour chaque horizon ; étant donné que les performances de tous ces modèles sont très proches, il conviendra de choisir le prédicteur qui sera implémenté sur la base d'autres caractéristiques dont nous parlerons plus loin dans ce chapitre.

5.3. Prévision pour un horizon de 15 minutes à 2 heures par pas de temps de 15 minutes (2h/15min)

Le second prédicteur que nous avons eu à créer concerne les prévisions pour un horizon allant de 15 minutes à 2 heures par pas de temps de 15 minutes. Ces travaux nous ont été demandés par les autres partenaires car ils correspondaient à un besoin particulier à un moment précis du projet. Ces horizons et granularités semblaient compatibles avec la gestion du micro réseau de Tilos et nous avons donc dû tester nos modèles dans ces conditions expérimentales.

Le Tableau IV-22 présente les résultats de l'auto information mutuelle pour cette granularité temporelle :

Tableau IV-22:nombre de données d'entrée pour les prévisions du rayonnement global incliné à Tilos (2h/15min).

Horizon (min)	*15*	*30*	*45*	*60*	*75*	*90*	*105*	*120*
Nombre d'entrées	5	6	6	6	7	8	8	8

Les résultats des simulations sont présentés dans le Tableau IV-23 qui contient les nRMSE en fonction de l'horizon, nous avons surligné en vert la meilleure valeur, en jaune la seconde et en rouge la troisième.

Tableau IV-23: nRMSE en fonction de l'horizon pour les prévisions à Tilos (2h/15min)

nRMSE (%)	*15*	*30*	*45*	*60*	*75*	*90*	*105*	*120*
P	19,20	27,60	34,20	39,80	45,00	50,10	54,50	58,70
PI	16,60	22,30	25,90	28,20	30,30	32,30	33,60	35,00
ARMA	15,70	25,60	27,10	28,75	29,20	30,00	30,60	31,00
PMC	15,60	25,30	26,80	28,10	29,10	29,50	30,40	30,70
PG	16,20	26,60	28,80	28,41	29,70	29,70	30,80	32,10
MVS	15,65	25,70	27,50	28,50	29,20	30,90	31,10	31,20
ARs	20,20	32,20	33,30	33,80	34,50	35,00	35,00	35,80
ARel	16,40	26,00	28,40	29,30	29,45	29,75	31,10	31,10
ARr	16,00	26,60	27,20	29,00	29,55	30,28	30,30	31,15
ARen	15,81	27,30	28,20	29,30	29,90	30,10	30,50	30,40
FA	15,80	25,80	27,00	28,40	29,25	29,20	29,40	29,50

Les modèles P et ARs sont les moins bons pour tous les horizons. Encore une fois, il est difficile d'établir un classement des modèles du fait des résultats très proches, un modèle se détache néanmoins, le PMC dont les performances se classent toujours parmi les trois meilleures. Le modèle FA a lui aussi des performances intéressantes et se détache notamment pour les horizons supérieurs à 45 minutes. L'écart entre le PMC, meilleur modèle à horizon de 15 minutes, et la PI est de 1% tout au plus. Pour l'horizon maximal de 120 minutes (2 heures), l'écart entre le meilleur modèle (FA) et la PI est de 5,5%.

La Figure IV-27 présente les résultats du skill score des différents modèles en fonction de l'horizon de prévision.

Comme nous l'avons souligné précédemment, le modèle ARs est en permanence moins bon que la PI. En ce qui concerne les autres modèles, ils sont tous sont meilleurs que la PI pour un horizon de 15 minutes, puis deviennent moins bon à horizon de 30 minutes et enfin redeviennent meilleurs pour les horizons supérieurs à 75 minutes. Dans l'ensemble ils ont tous un comportement comparable et il n'y en a aucun qui se détache particulièrement du lot.

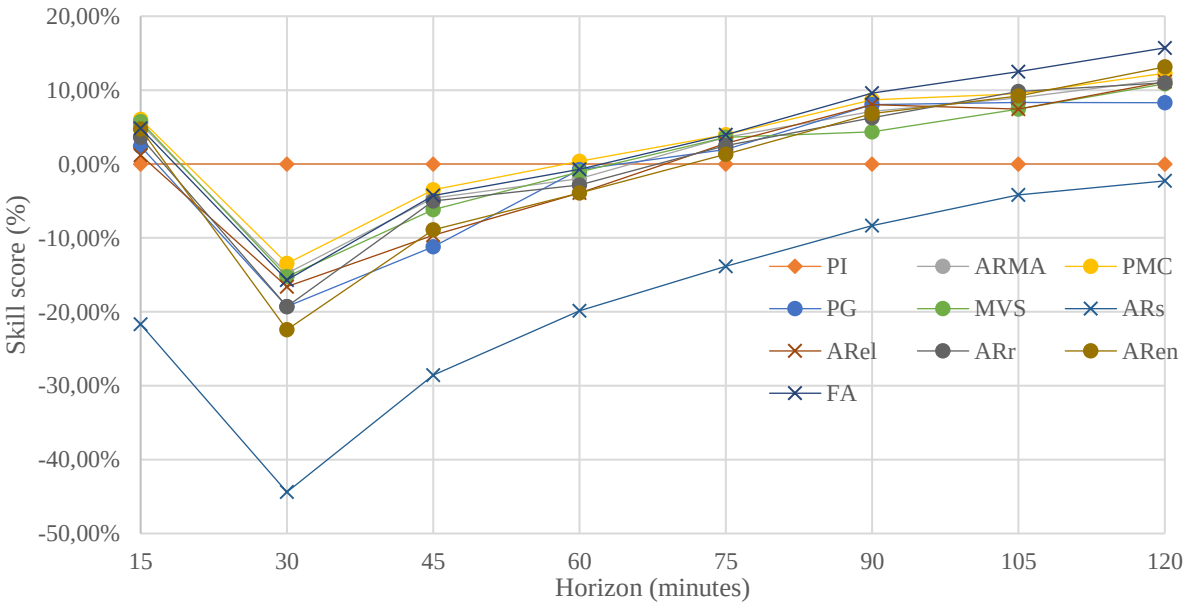


Figure IV-27: Skill score par rapport à la PI en fonction de l'horizon pour les prévisions à Tilos (2h/15min)

La Figure IV-28 présente les résultats du calcul de la MBE en fonction de l'horizon.

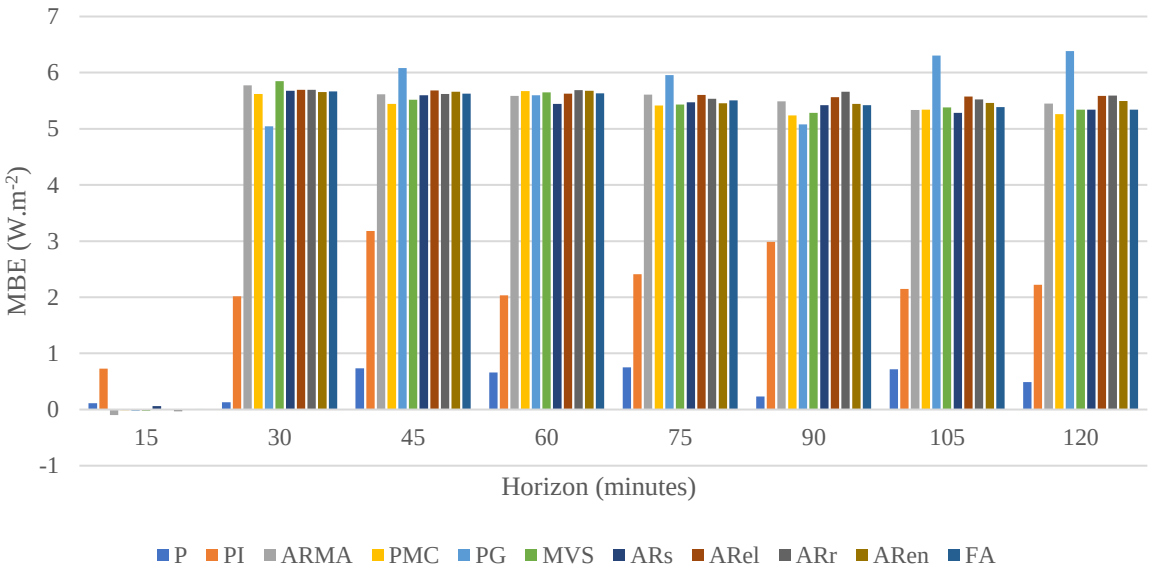


Figure IV-28: MBE en fonction de l'horizon pour les prévisions à Tilos (2h/15min)

On note que les modèles ont tendance à surestimer lorsqu'on augmente l'horizon, les modèles qui induisent le moins de biais, comme d'habitude, sont ceux basés sur la persistance. Pour tous les autres prédicteurs, les valeurs sont assez proches, autour de 5,6 $W.m^{-2}$ ce qui représente moins de 1% en valeur relative. La valeur la plus élevée de MBE est obtenue pour le modèle PG à un horizon de 120 minutes, elle est de 6,38 $W.m^{-2}$, soit 1,1% en valeur relative.

5.4. Prévision pour un horizon de 10 minutes à 2 heures par pas de temps de 10 minutes (2h/10min)

La dernière demande des partenaires du projet concerne le développement des modèles pour un horizon de 2h et une granularité de 10 minutes. Ce type de prévisions est bien plus intéressant pour les applications de gestion des flux énergétiques dans le micro-réseau hybride. En effet, cette échelle temporelle permet au gestionnaire du réseau intelligent (smart grid) de réagir rapidement et d'adapter sa production. Les modèles que nous avons développés pour cet horizon et cette granularité temporelle sont ceux qui ont finalement été implémentés dans la gestion du micro-réseau intelligent de l'île de Tilos actuellement en service.

Le Tableau IV-24 présente les résultats de l'auto information mutuelle pour cette granularité temporelle :

Tableau IV-24: nombre de données d'entrée pour les prévisions du rayonnement global incliné à Tilos (2h/10min)

Horizon (min)	*10*	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*	*100*	*110*	*120*
Nombre d'entrées	8	8	9	9	8	9	10	10	11	10	11	12

Même si les résultats de ces calculs sont difficilement interprétables on peut souligner que le nombre de données historiques nécessaire à la confection des modèles est plus grand que ce que l'on avait observé jusqu'ici. Pour l'horizon le plus lointain (2h), il faut prendre en compte 12 données historiques.

Les résultats des calculs de la nRMSE sont consignés dans le Tableau IV-25, pour chaque horizon la meilleure valeur est surlignée en vert, la seconde meilleure en jaune et la troisième en rouge.

Tableau IV-25: nRMSE en fonction de l'horizon pour les prévisions de rayonnement global incliné à Tilos (2h/10min)

nRMSE (%)	*10*	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*	*100*	*110*	*120*
P	13,93	19,51	23,42	26,59	29,87	32,81	35,72	38,65	41,18	43,94	46,41	48,85
PI	13,37	17,94	20,53	22,11	23,64	24,87	25,99	27,07	27,74	28,77	29,51	30,25
ARMA	12,90	22,76	24,21	24,97	25,61	26,38	26,77	27,42	27,58	28,08	28,09	28,62
PMC	12,80	22,68	24,22	24,64	25,51	26,25	26,53	26,95	27,26	27,78	28,19	28,33
PG	12,89	22,75	24,27	24,73	25,57	26,32	26,64	27,00	27,37	27,89	28,30	28,52
MVS	13,15	23,67	25,32	26,36	26,93	27,22	27,36	27,88	28,18	28,34	28,91	29,03
ARs	15,16	28,44	29,34	30,10	30,70	31,03	31,66	32,24	32,28	32,53	32,86	33,38
ARel	13,87	23,44	24,56	25,31	26,06	26,52	26,97	27,26	27,67	27,88	28,44	28,61
ARr	12,99	22,97	24,15	25,10	25,52	26,21	26,65	26,91	27,61	27,82	28,07	28,38
ARen	12,83	23,32	23,82	24,57	25,42	25,89	26,10	26,31	26,70	26,95	27,43	27,82
FA	12,91	23,14	23,94	24,82	25,29	25,85	26,14	26,56	26,80	27,26	27,35	27,70

On note une fois de plus qu'aucun modèle ne se détache réellement des autres au niveau des performances. On remarque que le modèle P a des performances intéressantes jusqu'à un horizon de 50 minutes, il se classe même second par deux fois, pour les horizons de 20 et 30 minutes. Pour un horizon de 10 minutes, tous les modèles ont des performances similaires hormis pour le modèle ARs qui est toujours le moins intéressant ; pour ce même horizon, l'écart entre le meilleur prédicteur (PMC), et le moins bon (P) est de 1,13% et l'écart avec la PI (qui reste notre référence) est de 0,6%. Pour l'horizon le plus éloigné, de 120 minutes, le classement des modèles change, le meilleur prédicteur devient FA et l'écart respectivement avec P (alors le plus mauvais) et PI est de 11,1% et 2,5%, ce qui rend difficile le classement des prédicteurs avec des performances qui varient très peu (écart maximal de 2,55%).

La Figure IV-29 représente les résultats du skill score en fonction de l'horizon.

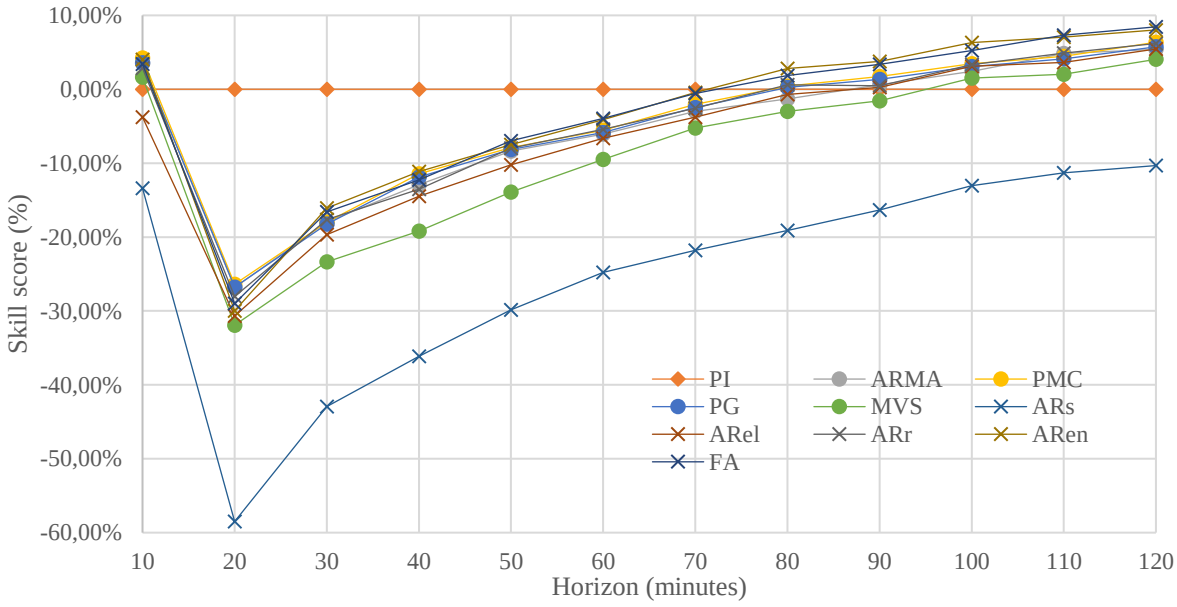


Figure IV-29: Skill score en fonction de l'horizon pour les prévisions à Tilos (2h/10min)

La persistance simple n'est pas représentée sur la Figure IV-29 car elle apporte peu d'informations. En effet, si l'on se réfère au Tableau MM on voit qu'elle est toujours moins performante que la PI. Tous

les modèles ont quasiment tous les mêmes performances par rapport à la PI, excepté ARs qui est toujours moins bon.

La Figure IV-30 présente les résultats en MBE pour ces prévisions.

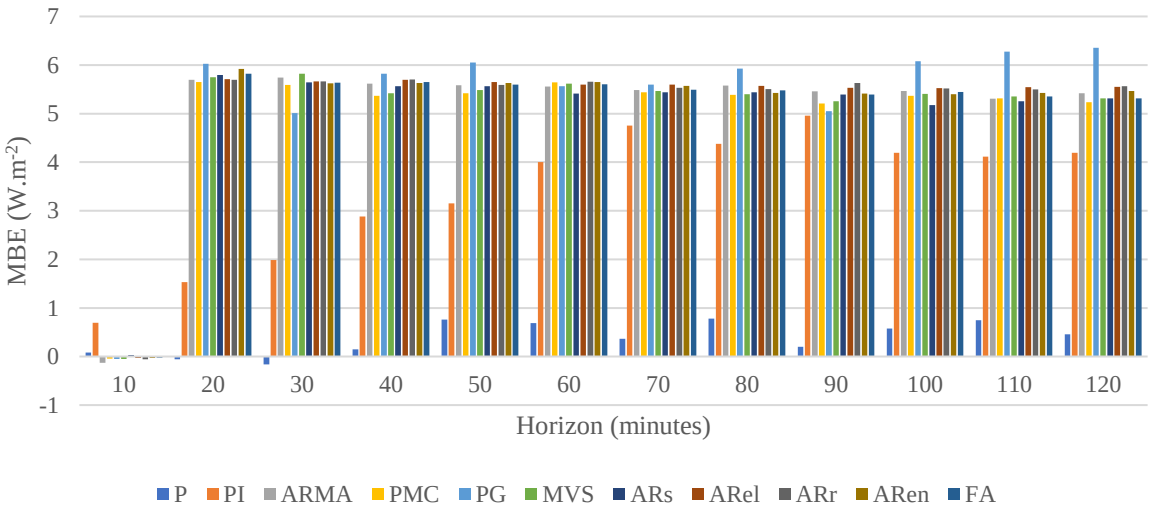


Figure IV-30: MBE en fonction de l'horizon pour les prévisions à Tilos (2h/10min)

Les résultats du calcul de la MBE sont semblables à ceux que nous avons pu observer jusqu'à maintenant, nos modèles ont une petite tendance à la surestimation, autour de 5,5 $W.m^{-2}$ en moyenne ce qui représente 0,9% en valeur relative. Le biais maximal pour le modèle PG est obtenu pour l'horizon 120 minutes, et vaut 6,35 $W.m^{-2}$ ce qui représente 1,1% en valeur relative.

Comme dans le cas du prédicteur 10min/1min, le modèle qui sera implémenté dans le Système de Gestion de l'Energie sera choisi sur la base d'autres paramètres.

5.5. Particularités de la prévision opérationnelle, contraintes et solutions

Comme nous l'avons mentionné en introduction du paragraphe 5, nous avons rencontré des contraintes auxquelles nous n'avions jamais été confrontés auparavant du fait de la nécessité d'avoir à rendre ces modèles opérationnels et « pratiques » pour notre gestionnaire de réseau. La mise en œuvre des modèles développés dans le paragraphe 2, n'a pas pris en compte les contraintes liées à une utilisation opérationnelle comme c'est le cas dans la majeure partie des travaux scientifiques qui ont été cités en références. Le fait que ces modèles soient implémentés au sein d'une unité de gestion de l'énergie nous a forcé à revoir leur mode opératoire et notamment à prendre en compte l'automatisation de leur exécution et de leur actualisation. En effet, l'opérateur n'a pas le temps de générer « manuellement » la prévision quand il en a besoin, de regarder les résultats et de prendre alors la décision ; il est totalement inenvisageable de lui demander chaque 10 minutes par exemple, les résultats générés par le prédicteur. Si l'on reprend la Figure TT vue en introduction du paragraphe 5, on constate que tous les constituants de la plateforme de gestion du micro-réseau sont liés, et doivent donc s'exécuter automatiquement et de manière indépendante pour fournir les données au centre de décision. C'est cette subtilité qui engendre une grande partie des modifications que nous avons dû réaliser sur la façon de coder nos modèles. Ces contraintes de codage ont eu des conséquences sur toutes les étapes de mise en œuvre des modèles, à savoir sur le traitement préalable des données, sur les apprentissages des modèles et sur les prévisions.

Les premières modifications viennent de la collecte des données : dans l'élaboration de nos modèles (paragraphe 2), le set de données total était déjà à notre disposition. Au besoin ce set était agrandi au fur et à mesure de l'arrivée de nouvelles données mais de manière ponctuelle, c'est-à-dire que l'on prenait en compte ces nouvelles données que lorsqu'elles étaient suffisamment nombreuses pour avoir une conséquence sur nos résultats. Dans le cas d'un fonctionnement opérationnel cette collecte des données est réalisée en temps réel, ce qui complique de manière significative le fonctionnement. Les données sont collectées et stockées dans une base de données SQL, (Structured Query Language : SQL est un langage informatique normalisé servant à exploiter des bases de données relationnelles) ce qui signifie que la plateforme doit interroger la base de données à chaque exécution afin d'avoir en entrée les dernières données disponibles. Cette base de données étant alimentée en permanence, se pose alors le deuxième problème auquel nous avons dû nous adapter, la gestion des heures de nuit. En effet en mode expérimental (lors du développement de nos modèles), nous avons appliqué une filtration sur les données et nous supprimions les heures de nuit, or dans le cas de la gestion du système installé à Tilos, les données de rayonnement solaire mesurées pendant la nuit sont présentes et « notre » prédicteur doit les détecter avant de les envoyer au logiciel de gestion de l'énergie. Comment détecte-t-on la nuit ? Les informations sur les heures de coucher et de lever du soleil sont connues et directement disponibles en particulier lorsque l'on génère le ciel clair, il a donc fallu rajouter à nos modèles une étape de prise en compte de ces heures de nuit. Même si elles n'apportent pas d'informations les données de nuit doivent être envoyées à la plateforme de gestion qui continue à fonctionner.

Ensuite, il a fallu modifier la phase d'apprentissages des modèles afin qu'elle soit adaptée à la manière dont étaient collectées les données. Dans le cas expérimental de développement et du test des modèles, nous réalisions les apprentissages à chaque exécution de modèles. Dans le cas du fonctionnement opérationnel « online » il est bien entendu impossible de réaliser cet apprentissage à chaque exécution, pour des raisons techniques et de temps d'exécution. Il est alors nécessaire de réaliser ces apprentissages régulièrement (environ chaque mois) en amont et de conserver les nouveaux modèles entrainés et vérifiés pour les implémenter ensuite dans la plateforme de gestion de l'énergie. C'est pour cela qu'il a été nécessaire de fractionner nos codes pour avoir des fonctions qui prennent en charge chaque étape, de la collecte des données à la prévision. Ainsi il est plus pratique d'utiliser un code maître qui « appelle » chaque fonction au moment où il en a besoin. Une fois les apprentissages réalisés pour chaque pas de temps, (à chaque pas de temps correspond un modèle entraîné), ces modèles sont sauvegardés au sein d'une fonction et sont utilisés ensuite pour la prévision en remplacement des modèles utilisés auparavant. Il a semblé après discussion avec nos partenaires du projet et gestionnaire de la plateforme de prévision qu'une bonne fréquence pour le renouvellement de l'apprentissage serait d'une fois par mois, le nombre de données ayant augmenté de manière significative pour avoir un effet bénéfique sur l'amélioration de la prévision.

Enfin, de nouvelles contraintes portaient sur la prévision elle-même. En effet, lors du développement et du test des modèles, les prévisions ont été effectuées à chaque instant t pour les horizons suivants t+1, t+2, … jusqu'à atteindre l'horizon maximal désiré, on utilise ainsi une fenêtre glissante. Dans le cas opérationnel, on utilise également une fenêtre glissante mais elle est définie différemment. Dans le cas expérimental (développement du modèle), cette fenêtre est décalée d'un pas de temps à chaque itération pour réaliser la prévision. Dans le cas de la prévision opérationnelle, l'opérateur ne peut pas réagir à chaque fois que la prévision est relancée (toutes les 10 minutes dans notre cas) car il n'a pas le temps de donner des ordres à son système de gestion de l'énergie et d'adapter la production si rapidement. La fenêtre glissante doit alors glisser d'autant de pas de temps que nécessaire pour couvrir l'horizon désiré.

Même si cette modification peut paraitre minime elle a toute son importance car elle signifie que les prévisions doivent être réalisées de manière bien différente lorsque l'on veut se servir des résultats d'un modèle dans une situation opérationnelle de gestion d'un réseau énergétique.

Pour une meilleure compréhension, il nous a paru plus judicieux de donner un exemple concret afin de comparer ce qui se passe dans une phase de développement classique d'un modèle de prévision et dans une phase opérationnelle. Considérons une prévision réalisée à midi dans le cas du troisième prédicteur (2h/10min). La Figure IV-31 présente les deux manières de réaliser la prévision de manière concrète avec les heures des données historiques et prévues.

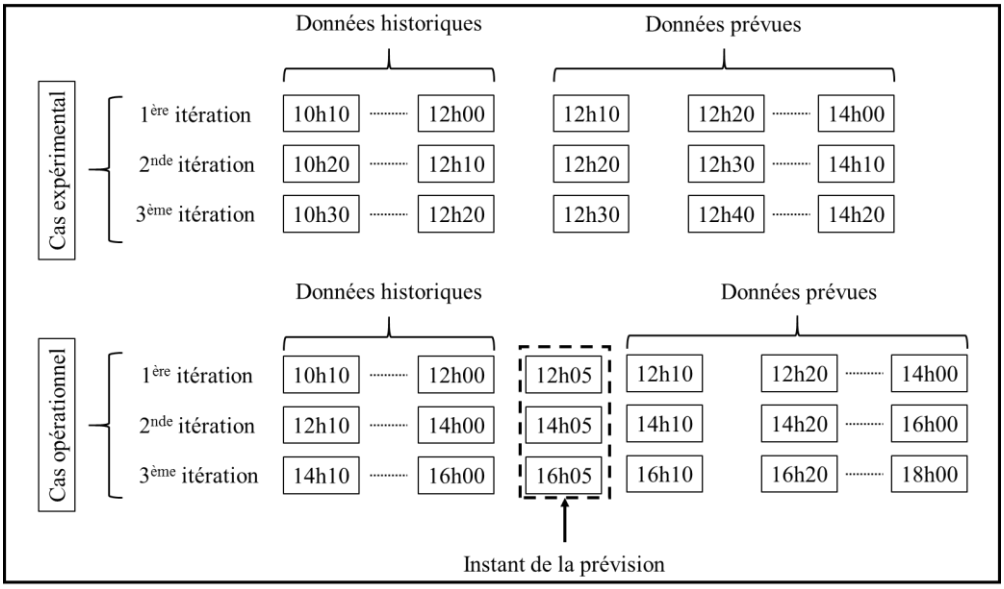


Figure IV-31: Cas concret de prévision pour le cas (2h/10min), en haut la méthode expérimentale et en bas la méthode opérationnelle

Dans la phase de développement, à midi on prédira les valeurs de 12h10, 12h20, …. 14h00. Puis à 12h10 on prédira les données de 12h20 à 14h10 et ainsi de suite. Cette manière de réaliser les prévisions n'est possible que lorsque l'on dispose des données d'entrée.

Dans la phase opérationnelle, à 12h05 on prédira les données de 12h10, 12h20…14h00. A 14h05 on prédira les données de 14h10, 14h20, … 16h00. L'exécution du modèle de prévision doit être réalisée 5 minutes avant la première prévision car il faut attendre d'avoir la dernière donnée précédente à disposition.

Comme nous l'avons évoqué dans le Chapitre 2, le manque de données est aussi un problème à prendre en compte lorsqu'on désire mettre en œuvre les modèles de prévision. Il peut arriver durant l'acquisition en temps réel des données qu'il y ait des problèmes de connexion ou de transfert. Bien sûr l'importance des données conduit à une surveillance accrue et une maintenance rigoureuse de la chaîne d'acquisition, toutefois nous avons intégré une fonctionnalité qui permet de pallier des manques de données peu étendus dans le temps pour ne pas gêner le fonctionnement de la plateforme logicielles.

Les données brutes utilisées ayant une granularité de 1 minute, nous avons mis en place le protocole suivant :

- Si l'intervalle de manque des données est inférieur à 5 minutes, les données sont remplacées par une interpolation linéaire entre les deux bornes de l'intervalle ;

- Si l'intervalle de manque des données est compris entre 5 et 30 minutes, les données sont remplacées par des données prévues par un modèle de persistance intelligente.

Ces deux manières de remplacer les données manquantes peuvent paraitres simplistes et génératrices d'erreurs, certes mais néanmoins acceptable pour que le fonctionnement de la plateforme logicielle ne soit pas interrompu. Il s'agit alors d'un bon compromis pour continuer à fournir des prévisions au logiciel de gestion du micro-réseau.

5.6. Synthèse sur les prévisions de rayonnement global incliné

Cette partie rassemble les différents résultats obtenus au cours des simulations que nous avons réalisées dans le cadre de la participation du projet H2020 TILOS. Ces simulations nous ont permis de confronter notre protocole expérimental avec des données différentes, des données de rayonnement global horizontal dans le paragraphe 2 et des mesures du rayonnement global incliné utilisées pour la gestion du micro-réseau de Tilos.

Les différents changements d'horizon et de pas de temps intervenus au cours du projet sont directement liés à la demande du partenaire en charge du développement de la plateforme de gestion du micro-réseau. Au fur et à mesure nous avons donc réalisé les simulations en répondant au cahier des charges que nous donnait le porteur du groupe de travail.

Les principaux résultats sont les suivants :

- Pour l'horizon de 1 à 10 minutes par pas de temps de 1 minute (10min/1min) : il s'agit de prévision à très court terme, les meilleurs modèles parmi ceux que nous avons mis en œuvre sont les modèles dits « naïfs » ce qui est une première dans le cadre de nos travaux, les niveaux de nRMSE sont de 10% pour l'horizon de 1 minute et 21% pour l'horizon de 10 minutes. Ces résultats ne sont pas inattendus, en effet comme nous l'avons souligné, la dynamique des nuages répond à une échelle temporelle supérieure à 1 minute. De plus après discussion avec les autres membres du projet il est apparu que ces résolutions temporelles étaient bien trop faibles et n'avait que très peu d'utilité pour les taches de gestion des flux énergétiques au sein du micro réseau intelligent.
- Pour l'horizon de 15 minutes à 2 heures par pas de temps de 15 minutes (2h/15min) : Nus avons pu dégager des tendances plus intéressantes et marquées. Il s'avère que nos modèles d'apprentissage automatique ont de bonnes performances, le classement des 3 meilleurs modèles était le suivant : PMC suivi de FA et de PI. Il faut cependant garder à l'esprit que tous les modèles basés sur l'apprentissage automatique donnaient des résultats assez proches (hormis le modèle d'arbres de régression simples). Les niveaux de nRMSE sont autour de 15% pour l'horizon de 15 minutes et autour de 30% pour l'horizon de 2 heures. Ces horizons et granularités semblent bien plus appropriés aux fonctions de gestion des flux énergétiques du micro réseau, cependant après discussion avec les autres membres du groupe de travail il est apparu nécessaire d'affiner encore un peu la fréquence des prévisions, ce qui nous a conduit à réaliser les dernières simulations ;
- Pour l'horizon de 10 minutes à 2 heures par pas de temps de 10 minutes (2h/10min) : le classement des modèles en fonction de leur performance a été difficile car on constate des résultats comparables pour tous les modèles basés sur l'apprentissage automatique, le classement que nous ayons pu faire donne comme meilleur modèle ARen suivis par FA et PMC. A souligner aussi que pour des horizons inférieurs à 60 minutes la persistance intelligente donnait des résultats très intéressants. Les niveaux de nRMSE sont plus bas, autour de 13% pour un horizon de 10 minutes et autour de 28% pour un horizon de 2 heures. On note que les modèles ont tous

des performances comparables ce qui signifie que le choix du modèle à implémenter sur la plateforme est difficile et dépend plus des contraintes techniques que des performances elles-mêmes. Dans ce cas c'est le modèle PMC qui a été choisi, car au niveau des versions logicielles utilisées qui ne prenaient pas en charge les modèles plus complexes basés sur les arbres de régression. De plus le modèle PMC fait preuve de robustesse et de facilité d'exécution, il est utilisé par une grosse communauté et les apprentissages sont rapides et éprouvés.

6. Synthèse des simulations

Nous présentons ici une brève synthèse finale car quatre synthèses relatives à chaque sous partie ont été déjà présentée.

Dans ce chapitre nous avons réalisé quatre études :

- Prévision des irradiations solaires horaires globales pour des horizons de 1 à 6 heures ;
- Prévision probabiliste afin d'obtenir un encadrement de chaque donnée prédite ;
- Application de trois prédicteurs (persistance intelligente, PMC et Forêts aléatoires) aux composantes diffuse horizontale et directe normale sur le site d'Odeillo pour des horizons de 1 à 6 heures ;
- Enfin, applications des 12 modèles prédicteurs pour des horizons/pas de temps de 2h/10min, 2h/15min et 10min/1min par suite de la demande du partenaire du projet Tilos en charge du Système de Gestion de l'Energie.

Il est ressorti de cette étude qu'il était difficile de dégager une tendance quelle que soit la variabilité du site étudié ; certes plus cette variabilité est élevée, plus les « meilleurs » prédicteurs sont complexes mais l'écart au niveau des performances n'est pas aussi important que celui auquel nous aurions pu nous attendre.

V. Conclusion générale et perspectives

1. Conclusion générale

La prévision de la ressource solaire est au cœur de la problématique de l'augmentation de l'intégration des systèmes de production photovoltaïques dans le mix énergétique. L'anticipation sur la production et les informations sur l'évolution de la ressource dans un futur proche permettent aux gestionnaires de réseaux électriques d'adapter leurs actions et de maintenir à tout instant la qualité et la sécurité du réseau électrique. L'émergence des micro-réseaux intelligents et des nouvelles méthodes de consommation de l'énergie conduisent eux aussi à une augmentation des besoins en prévisions fiables pour différents horizons temporels.

Il existe de nombreuses méthodes pour prévoir la ressource solaire, les différentes techniques et modèles associés sont mis en œuvre pour différents horizons. Il est bien connu que ce sont les modèles d'apprentissage automatique basés sur l'utilisation de séries temporelles qui sont les plus adaptés pour de la prévision à court terme, de quelques minutes à quelques heures.

La mise en œuvre de ce type de modèles nécessite de disposer de données de mesures in situ de rayonnement solaire ; nous avons pu obtenir des données en provenance de plusieurs sites de mesures avec des caractéristiques météorologiques variées : Ajaccio, Nancy, Odeillo, Tilos. Chacun des sites a été caractérisé par sa variabilité, propriété des données à varier plus ou moins au cours du temps, en utilisant le paramètre appelé « mean absolute log return ».

La « propreté » des séries temporelles qui doivent être modélisées nécessite de passer par une série de prétraitements rigoureux. Pour cela nous avons mis en place un protocole précis et adapté à notre utilisation, qui consiste à collecter les données, contrôler leur qualité par la détection des valeurs aberrantes et des périodes de données manquantes. Les méthodes utilisées nécessitaient que les données prédites soient stationnarisées afin de s'affranchir des tendances connues qu'elles contiennent et ainsi prédire uniquement leur composante stochastique. Cette stationnarisation a été réalisée avec l'utilisation d'un modèle de connaissance appelé modèle ciel clair.

Une fois stationnarisées, les données ont été organisées pour alimenter les modèles de prévision. Les modèles basés sur l'apprentissage automatique nécessitent de passer par une phase d'apprentissage au cours de laquelle les différents paramètres du modèle sont déterminés. Ainsi, la série temporelle est partitionnée en deux groupes, un pour la phase d'apprentissage (80%) et un pour la phase de test (20%). Pour éviter que la phase d'apprentissage ne soit soumise à des tendances saisonnières et pour renforcer la validation de nos modèles, une validation croisée a utilisé par méthode des k-fold.

Ces modèles admettent en entrée des données ordonnées sous forme matricielle. La dimension de ces matrices d'apprentissage, liée au nombre de données historiques nécessaires pour de la prévision à horizon est déterminée par utilisation de la méthode d'auto information mutuelle. Les performances des modèles ont été estimées par l'utilisation de métriques d'erreurs admises et répandues dans le domaine de la prévision.

La réalisation d'un état de l'art de la prévision univariée du rayonnement solaire global horizontal nous a permis de dégager différentes pistes sur les modèles à mettre en œuvre. Trois grands types de modèles ont été choisis et comparés :

- Les modèles sans apprentissage basés sur la persistance des conditions météorologiques, la persistance, la persistance intelligente et la persistance intelligente améliorée par la filtration de Kalman ;

- Les modèles avec apprentissage dits « classiques » car ils ont assez présents dans la littérature : les processus autorégressifs à moyenne mobile, les réseaux de neurones artificiels, les processus gaussiens et les machines à vecteurs de support.
- Les modèles avec apprentissage basés sur les arbres de régression, bien moins utilisés dans le domaine de la prévision mais dont l'émergence ne fait aucun doute au vu de leur fréquence d'apparition dans les dernières études publiées. Ils comprennent les arbres de régression simples, ensuite déclinés sous plusieurs formes grâce à diverses améliorations, les arbres élagués, renforcés, ensachés et les forêts aléatoires.

La confrontation de ces modèles avec la réalité expérimentale a été réalisée sur les différents jeux de données, d'une part sur différents sites et d'autre part sur différentes composantes du rayonnement solaire, directe et diffuse sur le site d'Odeillo, cette dernière prévision ayant été très peu abordée dans la littérature.

Les travaux réalisés peuvent être classés en quatre grandes parties :

- Pour la prévision des irradiations horaires globales horizontales, les performances des modèles ont été comparées en fonction des différents sites de mesures pour des horizons de 1 à 6 heures. Le but consistait à dégager les « meilleures » modèles en fonction de la variabilité des données. Les principales constatations sont les suivantes :
    - Plus la variabilité augmente, plus les prévisions sont difficiles à réaliser, ce qui paraissait évident avant même de tester les modèles.
    - Il n'apparait pas de prédicteur nettement meilleur que les autres.
    - Pour une faible variabilité, l'écart des performances entre les modèles est faible voire même très faible ; il faut atteindre une variabilité plus élevée (Nancy et Odeillo) pour qu'apparaisse un prédicteur ou un groupe de prédicteurs plus efficient.
- Une méthodologie probabiliste a été mise en œuvre et a permis d'ajouter à la valeur prédite un encadrement, autrement dit une confiance dans cette prévision, absolument nécessaire pour que le gestionnaire de réseau puisse prendre une décision avec le plus de sécurité possible.
- Les composantes directe normale et diffuse horizontale ont été prédites sur Odeillo, dont la variabilité est la plus élevée ; la connaissance de ces composantes est importante dans certaines applications solaires telles que celles appliquées pour le bâtiment ou les systèmes à concentration ; il est apparu que les performances obtenues pour ces deux composantes étaient bien moins bonnes que pour le rayonnement global du fait de leur plus profondes et brusques variations.
- Enfin une application de nos modèles de prévision sur des données de rayonnement global incliné pour le projet TILOS a été réalisées, pour des résolutions et des granularités temporelles différentes. Cette partie nous a notamment permis de nous confronter à un cahier des charges opérationnel qui nécessite de grand changement dans la manière d'appréhender la prévision pour rendre les modèles utilisables en mode online et efficaces pour la gestion d'un micro réseau.

2. Perspectives

La réalisation de ces travaux de thèse a révélé des perspectives d'amélioration auxquelles il semble judicieux de s'intéresser.

La première perspective qui vient à l'esprit est d'améliorer la prévision des irradiations aux heures de lever du soleil ; il apparait alors utile d'utiliser des données exogènes qui permettraient de « savoir » ce qui s'est passé pendant la nuit d'un point de vue météorologique.

L’utilisation de la variabilité des données pour comparer les performances des modèles nous a conduit vers des pistes de réflexion intéressantes. Une étude saisonnière des performances des prédicteurs, entamée lors de l’étude des composantes diffuse et directe, mérite d’être étendue et améliorée.

Le développement et l’utilisation de plusieurs modèles conduit à un constat, certains modèles sont meilleurs que d’autres pour des horizons différents. L’hybridation des modèles pour en tirer le meilleur de chacun est alors une perspective d’amélioration intéressante. Il serait alors utile de trouver une manière d’hybrider les différents modèles dont nous disposons en fonction de l’horizon de prévision pour tenter de maintenir une erreur de prévision la plus faible possible.

Bibliographie

AER, 2016. INTERMITTENCY AND THE COST OF INTEGRATING SOLAR IN THE GB POWER MARKET.

Aggarwal, S.K., Saini, L.M., 2014. Solar energy prediction using linear and non-linear regularization models: A study on AMS (American Meteorological Society) 2013–14 Solar Energy Prediction Contest. Energy 78, 247-256.

Almeida, M.P., Perpiñán, O., Narvarte, L., 2015. PV power forecast using a nonparametric PV model. Sol. Energy 115, 354-368.

Alobaidi, M.H., Marpu, P.R., Ouarda, T.B.M.J., Ghedira, H., 2014. Mapping of the Solar Irradiance in the UAE Using Advanced Artificial Neural Network Ensemble. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7, 3668-3680.

Alonso, A.M., Peña, D., Romo, J., 2002. Forecasting time series with sieve bootstrap. J. Stat. Plan. Inference 100, 1-11.

Anderson, D., Leach, M., 2004. Harvesting and redistributing renewable energy: on the role of gas and electricity grids to overcome intermittency through the generation and storage of hydrogen. Energy Policy 32, 1603-1614.

Bacher, P., Madsen, H., Nielsen, H.A., 2009. Online short-term solar power forecasting. Sol. Energy 83, 1772-1783.

Barlier, Y., 2000. The situation of electricity generation in Corsica. In: Proceedings of the dissemination of the advanced control technologies and SCADA systems for the isolated power networks with increased use of renewable energies. Ajaccio, p. 27-35.

Baromètre éolien 2018 | EurObserv'ER [WWW Document], 2018. URL https://www.eurobserv-er.org/barometre-eolien-2018/ (consulté le 11.5.18).

Baromètre Photovoltaïque 2018 | EurObserv'ER [WWW Document], 2018. URL https://www.eurobserv-er.org/barometre-photovoltaique-2018/ (consulté le 11.5.18).

Ben Taieb, S., Bontempi, G., Atiya, A.F., Sorjamaa, A., 2012. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. Expert Syst. Appl. 39, 7067-7083.

Bhatt, G., Khanna, M., Pani, B., Chaturvedi, A., 2016. Consumption pattern, behaviour and awareness towards e-waste among mobile users in New Delhi. Int. J. Environ. Policy Decis. Mak. 2, 28.

Bigdeli, N., Salehi Borujeni, M., Afshar, K., 2017. Time series analysis and short-term forecasting of solar irradiation, a new hybrid approach. Swarm Evol. Comput.

Bilan électrique SEI 2016, 2016.

Bilionis, I., Constantinescu, E.M., Anitescu, M., 2014. Data-driven model for solar irradiation based on satellite observations. Sol. Energy 110, 22-38.

Bird, L., Milligan, M., Lew, D., 2013. Integrating Variable Renewable Energy: Challenges and Solutions.

Bird, R.E., Huldstrom, R.L., 1980. Direct insolation models. Trans. ASME J. Sol. Energy Eng. 103, 182-192.

Black, M., Strbac, G., 2006. Value of storage in providing balancing services for electricity generation systems with high wind penetration. J. Power Sources 162, 949-953.

Borenstein, S., 2012. The Private and Public Economics of Renewable Electricity Generation. J. Econ. Perspect. 26, 67-92.

Bourbonnais, R., Terraza, M., 2008. Analyse des séries temporelles: Applications à l'économie et à la gestion.

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

Box, G.E.P., Jenkins, G.M., 1976. Time series analysis: Forecasting and control, 1re ed.

Brancucci Martínez-Anido, C., Florita, A., Hodge, B.M., 2014. The Impact of Improved Solar Forecasts on Bulk Power System Operations in ISO-NE Preprint The Impact of Improved Solar Forecasts on Bulk Power System Operations in ISO-NE.

Breiman, L., 1996. Bagging Predictors. Mach. Learn. 24, 123-140.

Breiman, L., 2001. Random Forest, Mach Learn.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees., Chapman & Hall.

Bühlmann, P., 2002. Bootstraps for Time Series. Stat. Sci. 17, 52-72.

Burrows, W.R., 1997. CART Regression Models for Predicting UV Radiation at the Ground in the Presence of Cloud and Other Environmental Factors. J. Appl. Meteorol. 36, 531-544.

Butler, P., Miller, J.L., Taylor, P.A., 2002. SANDIA REPORT Energy Storage Opportunities Analysis Phase II Final Report A Study for the DOE Energy Storage Systems Program.

Cano, D., Monget, J.-M., Albuisson, M., Guillard, H., Regas, N., Wald, L., method, al A., 1986. A method for the determination of the global solar radiation from meteorological satellites data. Elsevier.

Cao, J.C., Lin, X., 2008. Application of the diagonal recurrent wavelet neural network to solar irradiation forecast assisted with fuzzy technique. Eng. Appl. Artif. Intell. 21, 1255-1263.

Cao, J.C., Lin, X., s. d. Study of hourly and daily solar irradiation forecast using diagonal recurrent wavelet neural networks. Energy Convers. Manag. 49, 1396-1406.

Chaabene, M., Ben Ammar, M., 2008. Neuro-fuzzy dynamic model with Kalman filter to forecast irradiance and temperature for solar energy systems. Renew. Energy 33, 1435-1443.

Chakraborty, P., Marwah, M., Arlitt, M.F., Ramakrishnan, N., 2012. Fine-Grained Photovoltaic Output Prediction Using a Bayesian Ensemble. In: AAAI.

Chaouachi, A., Kamel, R.M., Nagasaka, K., 2010. Neural Network Ensemble-based Solar Power Generation Short-Term Forecasting. Int. J. Electr. Comput. Energ. Electron. Commun. Eng. 14, 69-75.

Chaouachi, A., Kamel, R.M., Nagasaka, K., 2010. Neural Network Ensemble-Based Solar Power Generation Short-Term Forecasting. JACIII 14, 69–75.

Chaturvedi, D.K., 2016. Solar Power Forecasting: A Review, International Journal of Computer Applications.

Chauvin, R., Nou, J., Thil, S., Grieu, S., 2014. Intra-Day DNI Forecasting Under Clear Sky Conditions Using ANFIS. IFAC Proc. Vol. 47, 10361-10366.

Chen, B., Gel, Y.R., Balakrishna, N., Abraham, B., 2011. Computationally efficient bootstrap prediction intervals for returns and volatilities in ARCH and GARCH processes. J. Forecast. 30, 51-71.

Chen, C., Duan, S., Cai, T., Liu, B., 2011. Online 24-h solar power forecasting based on weather type classification using artificial neural network. Sol. Energy 85, 2856-2870.

Cheng, H.-Y., 2016. Hybrid solar irradiance now-casting by fusing Kalman filter and regressor. Renew. Energy 91, 434-441.

Chu, Y., Li, M., Pedro, H.T.C., Coimbra, C.F.M., 2015a. Real-time prediction intervals for intra-hour DNI forecasts. Renew. Energy 83, 234-244.

Chu, Y., Pedro, H.T.C., Coimbra, C.F.M., 2013. Hybrid intra-hour DNI forecasts with sky image processing enhanced by stochastic learning. Sol. Energy 98, Part C, 592-603.

Chu, Y., Pedro, H.T.C., Li, M., Coimbra, C.F.M., 2015b. Real-time forecasting of solar irradiance ramps with smart image processing. Sol. Energy 114, 91-104.

COST, 2012. Weather Intelligence for Renewable Energies (WIRE).

COST action ES1002 Weather intelligence for Renewable Energies (WIRE), 2012.

Cybenkot, G., 1989. Mathematics of Control, Signals, and Systems Approximation by Superpositions of a Sigmoidal Function*, Math. Control Signals Systems.

D'Haultfoeuille, X., Givord, P., 2014. La régression quantile en pratique. Econ. Stat. 471, 85-111.

da Silva Fonseca Junior, J.G., Oozeki, T., Ohtake, H., Shimose, K., Takashima, T., Ogimoto, K., 2014. Regional forecasts and smoothing effect of photovoltaic power generation in Japan: An approach with principal component analysis. Renew. Energy 68, 403-413.

David, M., Ramahatana, F., Trombe, P.J., Lauret, P., 2016. Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models. Sol. Energy 133, 55-72.

De'ath, G., 2007. Boosted Trees for Ecological Modeling and Prediction. Ecology 88, 243-251.

De Felice, M., Petitta, M., Ruti, P.M., 2015. Short-term predictability of photovoltaic production over Italy. Renew. Energy 80, 197-204.

DeMeo, E.A., Jordan, G.A., Kalich, C., King, J., Milligan, M.R., Murley, C., Oakleaf, B., Schuerger, M.J., 2007. Accommodating wind's natural behavior. IEEE Power Energy Mag. 5, 59-67.

Demirtas, M., Yesilbudak, M., Sagiroglu, S., Colak, I., 2012. Prediction of solar radiation using meteorological data. 2012 Int. Conf. Renew. Energy Res. Appl. 1-4.

Diagne, H.M., David, M., Lauret, P., Boland, J., Schmutz, N., 2013. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. Renew. Sustain. Energy Rev. 27, 65-76.

Diagne, H.M., Lauret, P., David, M., 2012. Solar irradiation forecasting: state-of-the-art and proposition for future developments for small-scale insular grids. In: WREF 2012 - World Renewable Energy Forum. Denver, United States.

Dong, Z., Yang, D., Reindl, T., Walsh, W.M., 2015. A novel hybrid approach based on self-organizing maps, support vector regression and particle swarm optimization to forecast solar irradiance. Energy 82, 570-577.

Efron, B., 1979. Bootstrap Methods: Another Look at the Jackknife. Ann. Stat. 7, 1-26.

Elliston, B., MacGill, I., 2010. The potential role of forecasting for integrating solar generation into the Australian National Electricity Market. Sol. 2010 Proc. Annu. Conf. Aust. Sol. energy Soc.

Espinar, B., Aznarte, J.L., Girard, R., Moussa, A.M., Kariniotakis, G., 2010. Photovoltaic Forecasting: A state of art. Proc. 5th Eur. PV-Hybrid Mini-Grid Conf.

Espinar, B., Blanc, P., Wald, L., Hoyer-Klick, C., Shroedter Homscheidt, M., Wanderer, T. (Éd.), 2012. Controlling the quality of measurements of meterological variables and solar radiation. From sub-hourly to monthly average time periods. Eur. Geosci. Union 2012.

Fernández, Á., Gala, Y., Dorronsoro, J.R., 2014. Machine Learning Prediction of Large Area Photovoltaic Energy Production. In: Woon, W.L., Aung, Z., Madnick, S. (Éd.), Data Analytics for Renewable Energy Integration, Lecture Notes in Computer Science. Springer International Publishing, p. 38-53.

Ferrari, S., Lazzaroni, M., Piuri, V., Salman, A., Cristaldi, L., Rossi, M., Poli, T., 2012. Illuminance Prediction through Extreme Learning Machines. 2012 IEEE Work. Environ. Energy Struct. Monit. Syst. 97-103.

Franses, P.H., Paap, R., 1995. Seasonality and Stochastic Trends in German Consumption and Income, 1960.1-1987.4. Empir. Econ. 20, 109-132.

Freund, Y., Schapire, R.E., 1999. A Short Introduction to Boosting 14, 771-780.

Gastón, M., Pagola, Í., Fernández-Peruchena, C.M., Ramírez, L., Mallor, F., 2010. A new Adaptive methodology of Global-to-Direct irradiance based on clustering and kernel machines techniques. p. 11693.

Geiger, M., Diabaté, L., Ménard, L., Wald, L., 2002a. A web service for controlling the quality of measurements of global solar irradiation. Sol. Energy 73, 475-480.

Geiger, M., Diabaté, L., Ménard, L., Wald, L., 2002b. A web service for controlling the quality of measurements of global solar irradiation. Sol. Energy 73, 475-480.

Ghofrani, M., Ghayekhloo, M., Azimi, R., 2016. A novel soft computing framework for solar radiation forecasting. Appl. Soft Comput. 48, 207-216.

Gladyshev, E.G., 1963. Periodically and Almost-Periodically Correlated Random Processes with a Continuous Time Parameter. Theory Probab. Its Appl. 8, 173-177.

Gowrisankaran, G., Reynolds, S.S., Samano, M., Clark, R., Collard-Wexler, A., Cullen, J., Davis, L., Fowlie, M., Gilling-Ham, K., Handel, B., Hogan, B., Joskow, P., Keith, D., Lemoine, D., Pakes, A., Schmidt-Dengler, P., Teirilä, J., Wolfram, C., Wooders, J., Xiao, M., Zoettl, G., 2015. Intermittency and the Value of Renewable Energy *.

Grantham, A., Gel, Y.R., Boland, J., 2016. Nonparametric short-term probabilistic forecasting for solar radiation. Sol. Energy 133, 465-475.

Gross, R., Heptonstall, D., Anderson, D., Green, T., Leach, M., Skea, J., 2006. The costs and impacts of intermittency: an assessment of the evidence on the costs and impacts of intermittent generation on the British electricity network.

Gueymard, C., 1989. A two-band model for the calculation of clear sky solar irradiance, illuminance and photosynthetically active radiation at the earth surface. Sol. Energy 43, 252-265.

Gueymard, C., 2004. High performance model for clear sky irradiance and illuminance. In: ASES Conference.

H. Mori, N.K., 2001. Optimal regression tree based rule discovery for short-term load forecasting 421-426 vol.2.

Hammer, A., Heinemann, D., Lorenz, E., Lückehe, B., 1999. Short-term forecasting of solar radiation: a statistical approach using satellite data. Sol. Energy 67, 139-150.

Harvey, A.C., 1990. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press.

Hastie, T., Tibshirani, R., 1986. Generalized additive models. Stat. Sci. 1, 297-318.

Heinemann, D., Lorenz, E., Girodo, M., 2006a. Forecasting of solar radiation. Solar energy resource management for electricity generation from local level to global scale. NovaScience Publ.

Heinemann, D., Lorenz, E., Girodo, M., Dunlop, E.D., Wald, M., Suri, M., 2006b. Solar Energy Resource Management for Electricity Generation from Local Level to Global Scale. NovaScience Publ.

Hirth, L., 2014. The Economics of Wind and Solar Variability : how the Variability of Wind and Solar Power affects their Marginal Value, Optimal Deployment, and Integration Costs.

Hirth, L., Ueckerdt, F., Edenhofer, O., 2013. Integration Costs and the Value of Wind Power. SSRN Electron. J.

Hirth, L., Ueckerdt, F., Edenhofer, O., 2015. Integration costs revisited – An economic framework for wind and solar variability. Renew. Energy 74, 925-939.

Hirth, L., Ziegenhagen, I., 2015. Balancing power and variable renewables: Three links. Renew. Sustain. Energy Rev. 50, 1035-1051.

Hokoi, S., Matsumoto, M., Kagawa, M., 1990. Stochastic models of solar radiation and outdoor temperature. ASHRAE Trans. 96.

Holttinen, H., Meibom, P., Orths, A., Lange, B., O'Malley, M., Tande, J.O., Estanqueiro, A., Gomez, E., Söder, L., Strbac, G., Smith, J.C., van Hulle, F., 2011. Impacts of large amounts of wind power on design and operation of power systems, results of IEA collaboration. Wind Energy 14, 179-192.

Horin, C., Cohen, G.E., Apt, J., 2014. The cost of solar power variability. Carnegie Mellon Work. Pap. CEIC-11-04.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2, 359-366.

Hossain, M.R., Oo, A.M.T., Ali, A.B.M.S., 2012. Hybrid prediction method of solar power using different computational intelligence algorithms. In: Universities Power Engineering Conference (AUPEC), 2012 22nd Australasian. p. 1-6.

Huang, J., Troccoli, A., Coppin, P., 2014. An analytical comparison of four approaches to modelling the daily variability of solar irradiance using meteorological records. Renew. Energy 72, 195-202.

Humberto Marin, D., 2011. Intégration des éoliennes dans les réseaux électriques insulaires.

Ineichen, P., 2006. Comparison of eight clear sky broadband models against 16 independent data banks. Sol. Energy 80, 468-478.

Ineichen, P., 2008. A broadband simplified version of the Solis clear sky model. Sol. Energy 82, 758-762.

Iqbal, M., 1983. An Introduction To Solar Radiation - 1st Edition.

Jazwinski, A.H., 1970. Stochastic processes and filtering theory.

Joskow, P.L., 2008. Capacity payments in imperfect electricity markets: Need and design. Util. Policy 16, 159-170.

Joskow, P.L., 2011. Comparing the Costs of Intermittent and Dispatchable Electricity Generating Technologies. Am. Econ. Rev. 101, 238-241.

Kasten, F., 1980. A simple parameterization of the pyrheliometric formula for determining the Linke turbidity factor. Meteorol. Rundschau 33, 124-127.

Kasten, F., 1984. Parametrisierung der Globaslstrahlung durch Bedekungsgrad und Trübungsfaktor. Ann. der Meteorol. Neue Folge 20, 49-50.

Kasten, F., 1996. The linke turbidity factor based on improved values of the integral Rayleigh optical thickness. Sol. Energy 56, 239-244.

Katzenstein, W., Apt, J., 2012. The cost of wind power variability. Energy Policy 51, 233-243.

Kaur, A., Nonnenmacher, L., Pedro, H.T.C., Coimbra, C.F.M., 2016. Benefits of solar forecasting for energy imbalance markets. Renew. Energy 86, 819-830.

Kavadias, K.A., Zafirakis, D., Paliatsos, A.G., 2017. Application of Typical Meteorological Years for Sizing Building Integrated PV Systems under Zero Load Rejections. Energy Procedia 105, 881-887.

Kinsey Hill, G., 2008. BPA Calculates Administrative Costs of Wind Power. The Oregonian.

Koeppel, G., Korpås, M., 2008. Improving the network infeed accuracy of non-dispatchable generators with energy storage devices. Electr. Power Syst. Res. 78, 2024-2036.

Kraas, B., Schroedter-Homscheidt, M., Madlener, R., 2013. Economic merits of a state-of-the-art concentrating solar power forecasting system for participation in the Spanish electricity market. Sol. Energy 93, 244-255.

Krömer, P., Musílek, P., Pelikán, E., Krč, P., Juruš, P., Eben, K., 2014. Support Vector Regression of multiple predictive models of downward short-wave radiation. In: 2014 International Joint Conference on Neural Networks (IJCNN). p. 651-657.

Kurtz, B., Mejia, F., Kleissl, J., 2017. A virtual sky imager testbed for solar energy forecasting. Sol. Energy 158, 753-759.

Lahouar, A., Ben Hadj Slama, J., 2015. Day-ahead load forecast using random forest and expert input selection. Energy Convers. Manag. 103, 1040-1051.

Lara-Fanego, V., Ruiz-Arias, J.A., Pozo-Vázquez, D., Santos-Alamillos, F.J., Tovar-Pescador, J., 2012. Evaluation of the WRF model solar irradiance forecasts in Andalusia (southern Spain). Sol. Energy, Progress in Solar Energy 3 86, 2200-2217.

Lauret, P., David, M., Calogine, D., 2012. Nonlinear Models for Short-time Load Forecasting. Energy Procedia 14, 1404-1409.

Lauret, P., Voyant, C., Soubdhan, T., David, M., Poggi, P., 2015. A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. Sol. Energy 112, 446-457.

Law, E.W., Prasad, A.A., Kay, M., Taylor, R.A., 2014. Direct normal irradiance forecasting and its application to concentrated solar thermal output forecasting – A review. Sol. Energy 108, 287-307.

Lazzaroni, M., Ferrari, S., Piuri, V., Salman, A., Cristaldi, L., Faifer, M., 2015. Models for solar radiation prediction based on different measurement sites. Measurement 63, 346-363.

Lew, D., Brinkman, G., Ibanez, E., Hodge, B.M., Hummon, M., Florita, A., Heaney, M., 2013. The Western Wind and Solar Integration Study Phase 2. Golden, CO (United States).

Li, H., Berretta, S., Tan, Y., 2016. Dynamic Performance of the Standalone Wind Power Driven Heat Pump. Energy Procedia 103, 40-45.

Li, M., Chu, Y., Pedro, H.T.C., Coimbra, C.F.M., 2016. Quantitative evaluation of the impact of cloud transmittance and cloud velocity on the accuracy of short-term DNI forecasts. Renew. Energy 86, 1362-1371.

Logan, J., Kaplan, S.M., Kaplan, S., 2008. Wind Power in the United States: Technology, Economic, and Policy Issues.

Long, H., Zhang, Z., Su, Y., 2014. Analysis of daily solar power prediction with data-driven approaches. Appl. Energy 126, 29-37.

Lorenz, E., Heinemann, D., Hammer, A., 2004. Short-term forecasting of solar radiation based on satellite data. Proc. Eurosun (ISES Eur. Sol. Congr.

Lorenz, E., Künhert, J., Hammer, A., Scheidsteger, T., Heinemann, D., 2012. Short term forecasting of solar irradiance by combining satellite data and numerical weather predictions.

Lorenz, E., Remund, J., Müller, S.C., Traunmüller, W., Steinmaurer, G., Pozo, D., Ruiz-Arias, J.A., Fanego, V.L., Ramirez, L., Romeo, M.G., Kurz, C., Pomares, L.M., Guerrero, C.G., 2009. Benchmarking of different approaches to forecast solar irradiance. Proc. 24th Eur. Photovolt. Sol. Energy Conf. Exhib.

Lorenz, E., Scheidsteger, T., Hurka, J., Heinemann, D., Kurz, C., 2011. Regional PV power prediction for improved grid integration. Prog. Photovoltaics Res. Appl.

Lueken, C.A., 2012. Integrating Variable Renewables into the Electric Grid: An Evaluation of Challenges and Potential Solutions.

Luickx, P.J., Delarue, E.D., D'haeseleer, W.D., 2010. Impact of large amounts of wind power on the operation of an electricity generation system: Belgian case study. Renew. Sustain. Energy Rev. 14, 2019-2028.

Marquez, R., Coimbra, C.F.M., 2011. Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database. Sol. Energy 85, 746-756.

Masa-Bote, D., Castillo-Cagigal, M., Matallanas, E., Caamaño-Martín, E., Gutiérrez, A., Monasterio-Huelín, F., Jiménez-Leube, J., 2014. Improving photovoltaics grid integration through short time forecasting and self-consumption. Appl. Energy 125, 103-113.

Mathiesen, P., Kleissl, J., 2011. Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. Sol. Energy 85, 967-977.

McCandless, T.C., 2015. Artificial intelligence techniques for short-range solar irradiance prediction. THE PENNSYLVANIA STATE UNIVERSITY.

McClelland, J.L., Rumelhart, D.E., 1981. An Interactive Activation Model of Context Effescts in Letter Perception: Part1. An Account of Basics Findings. Psychol. Rev. 88.

McGovern, A., Ii, G., John, D., Eustaquio, L., Titericz, G., Lazorthes, B., Zhang, O., Louppe, G., Prettenhofer, P., Basara, J., Hamill, T.M., Margolin, D., 2015. Solar Energy Prediction: An International Contest to Initiate Interdisciplinary Research on Compelling Meteorological Problems. Bull. Am. Meteorol. Soc.

Mellit, A., Pavan, A.M., 2010. A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy. Sol. Energy 84, 807-821.

Mihalakakou, G., Santamouris, M., Asimakopoulos, D.N., 2000. The total solar radiation time series simulation in Athens, using neural networks 66, 185-197.

Milligan, M., Donohoo, P., Lew, D., Ela, E., Kirby, B., Holttinen, H., Lannoye, E., Flynn, D., O'Malley, M., Miller, N., Eriksen, P.B., Gottig, A., Rawn, B., Gibescu, M., Lazaro, E.G., Robitaille, A., Kamwa, I., 2010. Operating Reserves and Wind Power Integration: An International Comparison; Preprint. Present. 9th Annu. Int. Work. Large-Scale Integr. Wind Power into Power Syst. as well as Transm. Networks Offshore Wind Power Plants, 18-19 Oct. 2010, Quebec, Canada.

Milligan, M., Ela, E., Hodge, B.-M., Kirby, B., Lew, D., Clark, C., DeCesaro, J., Lynn, K., 2011. Integration of Variable Generation, Cost-Causation, and Integration Costs. Electr. J. 24, 51-63.

Moreno-Munoz, A., Rosa, J.J.G. de la, Posadillo, R., Bellido, F., 2008. Very short term forecasting of solar radiation. In: 33rd IEEE Photovoltaic Specialists Conference, 2008. PVSC '08. p. 1-5.

Moreno, A., Gilabert, M.A., Martínez, B., 2011. Mapping daily global solar irradiation over Spain: A comparative study of selected approaches. Sol. Energy 85, 2072-2084.

Mori, H., Takahashi, A., 2012. A data mining method for selecting input variables for forecasting model of global solar radiation. In: Transmission and Distribution Conference and Exposition (T D), 2012 IEEE PES. p. 1-6.

Mueller, R.W., Dagestad, K.F., Ineichen, P., Schroedter-Homscheidt, M., Cros, S., Dumortier, D., Kuhlemann, R., Olseth, J.A., Piernavieja, G., Reise, C., Wald, L., Heinemann, D., 2004. Rethinking satellite-based solar irradiance modelling: The SOLIS clear-sky module. Remote Sens. Environ. 91, 160-174.

Multon, B., Thiaux, Y., Ben Ahmed, H., 2011. Consommation d'énergie, ressources énergétiques et place de l'électricité. Ref TIP301WEB - « Conversion l'énergie électrique ».

Muneer, T., Fairooz Bsc, F., 2002. Quality control of solar radiation and sunshine measurements-lessons learnt from processing worldwide databases. Build. Serv. Eng. Res. Technol 23, 151-166.

Muneer, T., Fairooz, F., 2002. Quality control of solar radiation and sunshine measurements – lessons learnt from processing worldwide databases. Build. Serv. Eng. Res. Technol 23, 151-166.

Noia, M., Ratto, C.F., Festa, R., 1993a. Solar irradiance estimation from geostationary satellite data: I. Statistical models. Sol. Energy 51, 449-456.

Noia, M., Ratto, C.F., Festa, R., 1993b. Solar irradiance estimation from geostationary satellite data: II. Physical models. Sol. Energy 51, 457-465.

Notton, G., 2015. Importance of islands in renewable energy production and storage: The situation of the French islands. Renew. Sustain. Energy Rev. 47, 260-269.

Notton, G., Nivet, M.-L., Voyant, C., Paoli, C., Darras, C., Motte, F., Fouilloy, A., 2018. Intermittent and stochastic character of renewable energy sources: Consequences, cost of intermittence and benefit of forecasting. Renew. Sustain. Energy Rev. 87.

Olaiya, F., Adeyemo, A.B., 2012. Application of Data Mining Techniques in Weather Prediction and Climate Change Studies. Int. J. Inf. Eng. Electron. Business(IJIEEB) 4, 51.

Pagano, M., 1976. Periodoc and multiple autoregressions.

Paoli, C., Voyant, C., Muselli, M., Nivet, M.L., 2010. Forecasting of preprocessed daily solar radiation time series using neural networks. Sol. Energy 84, 2146-2160.

Paulescu, M., Paulescu, E., Gravila, P., Badescu, V., 2013. Weather Modeling and Forecasting of PV Systems Operation, Green Energy and Technology. Springer London, London.

Pedro, H.T.C., Coimbra, C.F.M., 2015. Nearest-neighbor methodology for prediction of intra-hour global horizontal and direct normal irradiances. Renew. Energy 80, 770-782.

Pedro, H.T.C., Coimbra, C.F.M., David, M., Lauret, P., 2018. Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts. Renew. Energy 123, 191-203.

Pelland, S., Galanis, G., Kallos, G., 2011. Solar and photovoltaic forecasting through post-processing of the Global Environmental Multiscale numerical weather prediction model.

Perez, R., Ineichen, P., Moore, K., Kmiecik, M., Chain, C., George, R., Vignola, F., 2002. A new operational model for satellite-derived irradiances: description and validation. Sol. Energy 73, 307-317.

Perez, R., Kivalov, S., Schlemmer, J., Hemker Jr., K., Renné, D., Hoff, T.E., 2010a. Validation of short and medium term operational solar radiation forecasts in the US. Sol. Energy 84, 2161-2172.

Perez, R., Kivalov, S., Zelenka, A., Schlemmer, J., Hemker Jr., K., 2010b. Improving The Performance of Satellite-to-Irradiance Models using the Satellite's Infrared Sensors. Proc., ASES Annu. Conf.

Perez, R., Moore, K., Wilcox, S., Renné, D., Zelenka, A., 2007. Forecasting solar radiation – Preliminary evaluation of an approach based upon the national forecast database. Sol. Energy 81, 809-812.

Phillips, P.C.B., 1986. Understanding spurious regressions in econometrics. J. Econom. 33, 311-340.

Podestá, G.P., Núñez, L., Villanueva, C.A., Skansi, M.A., 2004. Estimating daily solar radiation in the Argentine Pampas. Agric. For. Meteorol. 123, 41-53.

Porter, K., Rogers, J., 2008. Survey of Variable Generation Forecasting in the West: August 2011 — June 2012.

Prokop, L., Misak, S., Snasel, V., Platos, J., Kroemer, P., 2013. Supervised learning of photovoltaic power plant output prediction models. Neural Netw. World 23, 321–338.

Radnoti, G., Ajjaji, R., Bubnova, R., Caian, M., Cordoneanu, E., Von Der Emde, K., Gril, J.D., Hoffman, J., Horanyi, A., Issara, S., Ivanovici, V., Janousek, M., Joly, A., Le Moigne, P., Malardel, S., 1995. The spectral limited area model ARPEGE/ALADIN. PWPR Rep. Ser. n°7 111-117.

Rana, M., Koprinska, I., Agelidis, V.G., 2015. 2D-interval forecasts for solar power production. Sol. Energy.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes in Machine Learning. In: Bousquet, O., Luxburg, U. von, Rätsch, G. (Éd.), Advanced Lectures on Machine Learning, Lecture Notes in Computer Science. Springer Berlin Heidelberg, p. 63-71.

Reikard, G., 2009. Predicting solar radiation at high resolutions: A comparison of time series forecasts. Sol. Energy 83, 342-349.

Remund, J., Perez, R., Lorenz, E., 2008. Comparison of solar rasiation forecasts for the USA.

REN21 Renewable Energy Policy Network for the 21th Century, 2017.

Rigollier, C., Bauer, O., Wald, L., 2000. On the clear sky model of the ESRA — European Solar Radiation Atlas — with respect to the heliosat method. Sol. Energy 68, 33-48.

Rigollier, C., Lefèvre, M., Wald, L., 2004. The method Heliosat-2 for deriving shortwave solar radiation from satellite images. Sol. Energy 77, 159-169.

Rudd, T.R., 2011. BENEFITS OF NEAR-TERM CLOUD LOCATION FORECASTING FOR LARGE SOLAR PV.

Saguan, M., 2007. The economic analyses of power market architectures.
Application to real time market design. Centre pour la communication scientifique directe.

Salcedo-Sanz, S., Casanova-Mateo, C., Pastor-Sánchez, A., Sánchez-Girón, M., 2014. Daily global solar radiation prediction based on a hybrid Coral Reefs Optimization – Extreme Learning Machine approach. Sol. Energy 105, 91-98.

Samanta, M., Srikanth, B.K., Yerrapragada, J.B., 2015. Short-Term Power Forecasting of Solar PV Systems Using Machine Learning Techniques.

Shannon, C., 1948. A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379-423,623-656.

Shi, J., Lee, W.J., Liu, Y., Yang, Y., Wang, P., 2012. Forecasting Power Output of Photovoltaic Systems Based on Weather Classification and Support Vector Machines. Appl. IEEE Trans. 48, 1064-1069.

Sjoerd Brouwer, A., Van Den Broek, M., Seebregts, A., Faaij, A., 2014. Impacts of large-scale intermittent renewable energy sources on electricity systems, and how these can be modeled. Renew. Sustain. Energy Rev. 33, 443-466.

Smith, J.C., Demeo, E., Oakleaf, B., Wolf, K., Schuerger, M., Zavadil, R., Ahlstrom, M., Nakafuji, W.D.Y., 2006. Grid Impacts of Wind Power Variability: Recent Assessments from a Variety of Utilities in the United States; Preprint.

Trapero, J.R., Kourentzes, N., Martin, A., 2015. Short-term solar irradiation forecasting based on Dynamic Harmonic Regression. Energy 84, 289-295.

Troncoso, A., Salcedo-Sanz, S., Casanova-Mateo, C., Riquelme, J.C., Prieto, L., 2015. Local models-based regression trees for very short-term wind speed prediction. Renew. Energy 81, 589-598.

Tso, G.K.F., Yau, K.K.W., 2007a. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy 32, 1761-1768.

Tso, G.K.F., Yau, K.K.W., 2007b. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy 32, 1761-1768.

Vapnik, V., 1986. The Nature of Statistical Learning Theory. Springer Science & Business Media.

Voyant, C., Darras, C., Muselli, M., Paoli, C., Nivet, M.L., Poggi, P., 2014. Bayesian rules and stochastic models for high accuracy prediction of solar radiation. Appl. Energy 114, 218-226.

Voyant, C., De Gooijer, J.G., Notton, G., 2018a. Periodic autoregressive forecasting of global solar irradiation without knowledge-based model implementation.

Voyant, C., Motte, F., Fouilloy, A., Notton, G., Paoli, C., Nivet, M.-L., 2017a. Forecasting method for global radiation time series without training phase: Comparison with other well-known prediction methodologies. Energy 120.

Voyant, C., Motte, F., Notton, G., Fouilloy, A., Nivet, M.-L., Duchaud, J.-L., 2018b. Prediction intervals for global solar irradiation forecasting using regression trees methods. Renew. Energy 126.

Voyant, C., Notton, G., Darras, C., Fouilloy, A., Motte, F., 2017b. Uncertainties in global radiation time series forecasting using machine learning: The multilayer perceptron case. Energy 125.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., Fouilloy, A., 2017c. Machine learning methods for solar radiation forecasting: A review. Renew. Energy 105, 569-582.

Voyant, C., Soubdhan, T., Lauret, P., David, M., Muselli, M., 2015. Statistical parameters as a means to a priori assess the accuracy of solar forecasting models. Energy 90, Part 1, 671-679.

Wang, S., Zhang, N., Zhao, Y., Zhan, J., 2011. Photovoltaic system power forecasting based on combined grey model and BP neural network. 2011 Int. Conf. Electr. Control Eng. ICECE 2011 - Proc. 4623-4626.

Wilcox, R.R., 2012. Introduction to robust estimation and hypothesis testing. Academic Press.

Wilks, D., 1995. Statistical Methods in the atmospheric sciences: an introduction" Academic Press San Diego.

Wiser, R., Barbose, G., Mills, A., Rosa, A., Lab, B., Porter, K., Fink, S., Tegen, S., Musial, W., Oteri, F., Heimiller, D., Roberts, B., Belyeu, K., Stimmel, R., 2008. 2008 WIND TECHNOLOGIES MARKET REPORT 2008 Wind Technologies Market Report i 2008 Wind Technologies Market Report Primary authors With contributions from.

Wolff, E.L.B., Kramer, O., 2015. Statistical Learning for Short-Term Photovoltaic Power Predictions (chapter). ResearchGate in-print.

Wu, J., Chan, C.K., Zhang, Y., Xiong, B.Y., Zhang, Q.H., 2014. Prediction of solar radiation with

genetic approach combing multi-model framework. Renew. Energy 66, 132-139.

Wu, Y.-K., Chen, C.-R., Abdul Rahman, H., 2014. A Novel Hybrid Model for Short-Term Forecasting in PV Power Generation. Int. J. Photoenergy 2014, e569249.

Xuan, Z., Xuehui, Z., Liequan, L., Zubing, F., Junwei, Y., Dongmei, P., 2019. Forecasting performance comparison of two hybrid machine learning models for cooling load of a large-scale commercial building. J. Build. Eng. 21, 64-73.

Yang, H.-T., Huang, C.-M., Huang, Y.-C., Pai, Y.-S., 2014. A Weather-Based Hybrid Method for 1-Day Ahead Hourly Forecasting of PV Power Output. IEEE Trans. Sustain. Energy 5, 917-926.

Yona, A., Saber, A.Y., Sekine, H., 2007. Application of Neural Network to One-Day-Ahead 24 hours Generating Power Forecasting for Photovoltaic System. 14thInternational Conf. Intell. Syst. Appl. to Power Syst.

Zamo, M., Mestre, O., Arbogast, P., Pannekoucke, O., 2014. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: Deterministic forecast of hourly production. Sol. Energy 105, 792-803.

Zhang, Y., Lundblad, A., Campana, P.E., 2016. Employing Battery Storage to Increase Photovoltaic Self-sufficiency in a Residential Building of Sweden. Energy Procedia 88, 455-461.

Zhu, T., Wei, H., Zhao, X., Zhang, C., Zhang, K., 2017. Clear-sky model for wavelet forecast of direct normal irradiance. Renew. Energy 104, 1-8.