Prédiction de séries temporelles de rayonnement solaire global et de production d'énergie photovoltaïque à partir de réseaux de neurones artificiels

Cyril Voyant

▶ **To cite this version:**

Cyril Voyant. Prédiction de séries temporelles de rayonnement solaire global et de production d'énergie photovoltaïque à partir de réseaux de neurones artificiels. Energie électrique. Université Pascal Paoli, 2011. Français. NNT : . tel-00635298

HAL Id: tel-00635298

https://theses.hal.science/tel-00635298v1

Submitted on 25 Oct 2011

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

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

UNIVERSITE DE CORSE-PASCAL PAOLI
ECOLE DOCTORALE ENVIRONNEMENT ET SOCIETE
UMR CNRS 6134 (SPE)

Thèse pour l'obtention du grade de

DOCTEUR EN PHYSIQUE

Mention : Energétique

Soutenue publiquement par

CYRIL VOYANT

Le 16 Novembre 2011

Prédiction de séries temporelles de rayonnement solaire global et de production d'énergie photovoltaïque à partir de réseaux de neurones artificiels

Directeurs :

M. Marc Muselli, Professeur, Université de Corse
M. Christophe Paoli, Dr, Université de Corse

Rapporteurs :

Mme Irina Martin Graour, Professeure, Université de Marseille
M. Philippe Lauret, Professeur, Université de la Réunion

Jury :

Mme Irina Martin Graour, Professeure, Université de Marseille
M. Dominique Lambert, Dr-HDR, Université de Toulouse
M. Philippe Lauret, Professeur, Université de la Réunion
M. Marc Muselli, Professeur, Université de Corse
M. Christophe Paoli, Dr, Université de Corse
Mme Ilana Peretti, Professeure, Faculté de médecine, Université Paris VII
M. Philippe Poggi, Professeur, Université de Corse

UNIVERSITE DE CORSE-PASCAL PAOLI
ECOLE DOCTORALE ENVIRONNEMENT ET SOCIETE
UMR CNRS 6134 (SPE)

Thèse pour l'obtention du grade de

DOCTEUR EN PHYSIQUE

Mention : Energétique

Soutenue publiquement par

CYRIL VOYANT

Le 16 Novembre 2011

Prédiction de séries temporelles de rayonnement solaire global et de production d'énergie photovoltaïque à partir de réseaux de neurones artificiels

Directeurs :

M. Marc Muselli, Professeur, Université de Corse
M. Christophe Paoli, Dr, Université de Corse

Rapporteurs :

Mme Irina Martin Graour, Professeure, Université de Marseille
M. Philippe Lauret, Professeur, Université de la Réunion

Jury :

Mme Irina Martin Graour, Professeure, Université de Marseille
M. Dominique Lambert, Dr-HDR, Université de Toulouse
M. Philippe Lauret, Professeur, Université de la Réunion
M. Marc Muselli, Professeur, Université de Corse
M. Christophe Paoli, Dr, Université de Corse
Mme Ilana Peretti, Professeure, Faculté de médecine, Université Paris VII
M. Philippe Poggi, Professeur, Université de Corse

Remerciements

Avant d'attaquer la phase de remerciements, je voudrais dire à quel point il m'a été difficile d'entreprendre une thèse en étant salarié, de reprendre mes études à trente ans et surtout à quel point je suis content d'avoir pu la mener à son terme. Sans une rencontre avec le professeur Antoine Folacci en 2007, cette aventure n'aurait certainement jamais vu le jour. Merci de m'avoir orienté ainsi et merci pour les anecdotes scientifiques passionnantes (Cf Grothendieck).

Je voudrais tout d'abord remercier mes directeurs de thèses. J'ai beaucoup appris à leur contact : ils m'ont permis de progresser dans de nombreux domaines tant scientifiques que rédactionnels ou organisationnels. Donc, un grand merci à Marc Muselli, Christophe Paoli et Marie Laure Nivet (même si tu ne l'es pas officiellement, je te considère bien évidement comme ma directrice….). La disponibilité dont vous avez fait preuve me touche beaucoup, et me laisse penser que j'ai eu de la chance de vous rencontrer.

Je souhaite aussi faire part de ma gratitude aux membres du jury. Merci à Mme Irina Martin Graour et M. Philippe Lauret d'avoir accepté d'être rapporteurs de cette thèse. Je tiens aussi à remercier Mme Ilana Peretti, M. Philippe Poggi et M. Dominique Lambert, d'avoir consenti à être examinateurs dans mon jury de thèse.

Je tiens aussi à remercier tous les membres de l'équipe ENR du laboratoire Sciences pour l'Environnement de l'UMR CNRS 6134 de l'Université de Corse, car même si je n'étais basé ni sur l'Université ni sur le Centre Scientifique G. Peri de Vignola, tout le monde s'est toujours rendu disponible et m'a toujours facilité la tâche quand j'avais des questions ou des problèmes. C'est ainsi que je remercie plus particulièrement Jean-François Muzy, Rachel Baile, Christophe Darras ou encore Pierrick Haurant.

Les autres personnes sans qui je n'aurais pu mener à bien cette thèse sont l'ensemble du personnel du service de radiothérapie de l'hôpital de Castelluccio. En effet, sans leur aide, il m'aurait été très difficile de concilier mes activités de physicien médical le jour et de thésard

la nuit. Tout le monde a été très arrangeant et compréhensif, me permettant d'aborder avec sérénité les phases de réunion, d'expérimentation et de rédaction. Je tiens à remercier principalement Katia Biffi (ma collègue directe) et Céline Lantieri (une chef de service que beaucoup de monde rêverait d'avoir), ainsi que le restant de l'équipe de radiothérapie : Julie-Pierre, Audrey, Cécile, Françoise, Sylvie, Amélie, Emilie, Marina, Delphine et Jérôme.

Une partie des manipulations présentées dans ce manuscrit a été réalisée avec l'aide de stagiaires de l'Université de Corse et du master OSE de l'Ecole des Mines de Paris. Sans leur contribution, je n'aurais sans doute pas pu mener à terme l'ensemble des simulations, un grand merci pour votre contribution, Auline, Prisca, Mickael et Paul.

Outres toutes les personnes que je viens de présenter et qui m'ont aidées directement ou indirectement durant cette thèse, il y a trois autres personnes que je tiens tout particulièrement à remercier. Il s'agit tout d'abord, de mes deux enfants Solèn et Timéo. Je pense qu'il n'a pas toujours été facile pour eux de voir leur papa « trop souvent devant son ordinateur ». Il en est de même pour Aurore : pendant toutes les périodes où j'étais mobilisé par la thèse, je ne pouvais pas être très disponible à la maison, je me rends bien compte que je ne suis pas le seul à avoir fait des sacrifices pour que cette thèse soit menée à bien. Merci aussi pour l'aide apportée durant les phases de rédaction.

Résumé et mots clés

La Corse faisant partie des petits réseaux insulaires non-interconnectés, son approvisionnement énergétique est très particulier. En effet, comme toutes les îles, elle doit se suffire à elle-même. Une solution souvent adoptée pour pallier à cet isolement, consiste à recourir aux énergies renouvelables. Cependant, à cause de leur caractère intermittent, elles ne sont insérées que de manière limitée au sein des réseaux électriques. Il est nécessaire d'utiliser en parallèle d'autres moyens de production d'énergie, avec comme principale difficulté, la gestion optimale de la bascule entre ces deux types d'énergie. Cette étude s'inscrit dans le cadre de la prédiction de la ressource solaire et photovoltaïque dans le but de quantifier l'énergie disponible et de permettre une gestion optimale de la transition entre énergies intermittentes et conventionnelles. Tout au long de ces travaux, nous avons ainsi testé différentes techniques de prédiction sur quatre horizons susceptibles d'intéresser un gestionnaire de réseau : j+1, h+24, h+1 et m+5. A l'issue de toutes ces manipulations, nous pouvons conclure que suivant l'horizon considéré, la hiérarchisation des différents prédicteurs fluctue. On retiendra ainsi que, pour l'horizon j+1, il est intéressant d'utiliser une approche à base de réseaux de neurones en prenant soin de stationnariser les séries temporelles et d'utiliser des variables exogènes. Pour l'horizon h+1, une méthodologie hybride couplant la robustesse des modèles autorégressifs et la non-linéarité des modèles connexionnistes permet d'obtenir des résultats très satisfaisants. Pour le cas h+24, les réseaux de neurones à sorties multiples donnent de très bons résultats. Concernant l'horizon m+5, les conclusions sont moins catégoriques. Ainsi, même si les réseaux de neurones sont les plus performants, la simplicité et les résultats d'une approche basée sur la persistance, nous conduisent à préconiser principalement ce prédicteur. L'ensemble des méthodologies proposées et des résultats obtenus sont complémentaires avec les travaux de prédiction bibliographiques étudiés. Les méthodologies développées pourraient, à terme, être reprises comme éléments de prédiction dans des outils globaux de contrôle et de commande des systèmes énergétiques.

Mots clés : prédiction, rayonnement global, réseau de neurones, photovoltaïque, séries temporelles, stationnarité.

Abstract and keywords

As Corsica is a non-interconnected island, its energy supply is very special case. Indeed, as all islands, a large part of the electricity production must be generated locally. Often, renewable energies are considered as a good solution to overcome the isolation problem. However, because of their intermittent nature, they are included in a limited way in power systems. Thus, it's necessary to use in addition other energy productions, with main problem the management of the dispatch between these two energy types. This study is related to the solar and PV prediction in order to quantify available energy and to allow the optimal transition between intermittent and conventional energies sources. Throughout this work, we tested different techniques of prediction concerning four horizons interesting the power manager: d+1; h+24, h+1 and m+5. After all these manipulations, we can conclude that according the considered horizon, the prioritization of the different predictors varies. Note that for the d+1 horizon, it is interesting to use an approach based on neural network being careful to make stationary the time series, and to use exogenous variables. For the h+1 horizon, a hybrid methodology combining the robustness of the autoregressive models and the non-linearity of the connectionist models provides satisfactory results. For the h+24 case, neural networks with multiple outputs give very good results. About the m+5 horizon, our conclusions are different. Thus, even if neural networks are the most effective, the simplicity and the relatively good results shown by the persistence-based approach, lead us to recommend it. All the proposed methodologies and results are complementary to the prediction studies available in the literature. In conclusion, we can say that methodologies developed could eventually be included as prediction tools in the global command – control systems of energy sources.

Keywords : forecasting, global radiation, neural network, photovoltaic, time series, stationarity.

Liste des publications faites durant cette thèse

Liste des publications

- Solar radiation forecasting using ad-hoc time series preprocessing and neural networks. *Emerging Intelligent Computing technology and Applications 2009,* Springer (Berlin / Heidelberg). p 898-907. C Paoli, **C Voyant**, M Muselli, et ML Nivet ;

- Forecasting of preprocessed daily solar radiation time series using neural networks. *Solar Energy*. Decembre 2010;84(12), p 2146-2160. C. Paoli, **C Voyant,** M. Muselli, ML Nivet ;

- Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation. *Energy*, janvier 2011, p 348-359. **C Voyant,** M Muselli, C Paoli, ML Nivet ;

- Hybrid method dedicated to the hourly global horizontal radiation forecasting for mediterranean locations. *Energy conversion and management*, 2011, In press. **C Voyant,** M Muselli, C Paoli, ML Nivet ;

- Use of ALADIN forecast data, artificial neural network and ARMA model to predict hourly global radiation time series. *Energy*, 2011, In press. **C Voyant,** M Muselli, C Paoli, ML Nivet.

Liste des conférences

- Etude des séries chronologiques d'une source renouvelable d'énergie ; prédiction de la ressource d'irradiation solaire sur un plan horizontal. *Journée des Doctorants 2008*, université de Corse, Corte, France. **C Voyant**, M Muselli, C Paoli, ML Nivet, P Poggi ;

- Predictability of PV power grid performance on insular sites without weather stations: use of artificial neural networks. *24th European Photovoltaic Solar Energy Conference 2009*, Hambourg. **C Voyant**, M Muselli, C Paoli, ML Nivet, P Poggi.

Award dans la thématique des systèmes photovoltaïques (1583 candidats) ;

- Solar radiation forecasting using ad-hoc time series preprocessing and neural networks. *2009 International Conference on Intelligent Computing*; Ulsan, Korée. Christophe Paoli, **C Voyant**, M Muselli, ML Nivet ;

- Optimization of an Artificial Neural Network (ANN) Dedicated to the Daily Global Radiation and PV Plant Production Forecasting Using Exogenous Data. *25th European Photovoltaic Solar Energy Conference and Exhibition (25th EU PVSEC) 2010/ 5th World Conference on Photovoltaic Energy Conversion (WCPEC-5). Session 5BV.4.23*; Valence, Espagne. **C Voyant**, M Muselli, C Paoli, ML Nivet ;

- Use of exogenous data to improve an artificial neural networks dedicated to daily global radiation forecasting. *IEEE, international conference on environment and electrical engineering*. 2010, Prague, république tchèque. C Paoli, **C Voyant**, M Muselli, ML Nivet ;

- Use of artificial neural network and exogenous data to the daily global radiation forecasting. *Journée des Doctorants 2010*, université de Corse, Corte, France. **C Voyant**, M Muselli, C Paoli, ML Nivet ;

- Problématique de la prédiction du rayonnement global et de l'énergie PV : Approche connexionniste. *Les Journées des Doctorants, 2011*, Université de Corse, Corte, France. **C Voyant**, A Rodler, M Muselli, C Paoli, ML Nivet.

Liste des travaux de recherches relatifs à d'autres domaines scientifiques, faits durant cette thèse

- Dose comparison between three prostate planning : 3D conformational radiotherapy, coplanar arctherapy and non coplanar arcthérapy. *Cancer Radiothérapie*, septembre 2008, p 343-351. **C Voyant**, A Baadj, K Biffi, D Leschi, C Lantieri ;

- Application Mise en œuvre du control de qualité d'une gamma camera TEMP-TDM selon les recommandations de l'AFSSAPS. *ACOMEN 2010*; Ajaccio. **C Voyant**, K Biffi ;

- Mise en œuvre du contrôle de qualité en médecine nucléaire. *Médecine nucléaire*, mars 2010, p158-164. **C Voyant** ;

- Therapeutic potential of atmospheric neutrons. *Report of practical oncology and radiotherapy*, Janvier-Février 2011, p 21-31; volume 16 issue 1. **C Voyant**; R Roustit ; J Tatje; K Biffi; D Leschi; C L Marcovici ;

- Dosimetric uncertainties related to elasticity of bladder and rectal wall : adenocarcinoma of the prostate. *Cancer Radiothérapie*, 2011, Juillet 2011, p 270-278. **C Voyant**, K Biffi, J Briançon, D Leschi, C Lantieri.

Sommaire

CHAPITRE 2

CHAPITRE 3

CHAPITRE 4

Introduction

SOMMAIRE

La maîtrise du caractère aléatoire des sources d'énergies renouvelables telles que le rayonnement solaire au sol pourrait permettre aux gestionnaires de réseaux électriques de mieux les intégrer. Dans cette optique, nous nous sommes efforcés durant cette étude de contribuer à la recherche de méthodologies de modélisation du rayonnement solaire mais aussi de la puissance photovoltaïque. Ce type de prédiction est essentiel car il pourrait, à terme, permettre de mieux utiliser l'Energie Renouvelable (*ER*) solaire dont l'intermittence pénalise lourdement l'intégration dans les réseaux d'exploitation électrique existants. Nous nous sommes concentrés sur les séries temporelles et leur « prévision », ou plutôt leur « prédiction ». Il est important de distinguer ces deux termes : on emploiera prévision pour une estimation analytique d'une grandeur sur un temps continu et prédiction pour la connaître à un temps donné. Une phrase explicite bien cette différence : « tout le monde prévoit sa mort, mais personne ne sait prédire quand ». Pour effectuer cette prédiction à temps fixe (ou à horizon déterminé) il est nécessaire de se doter d'outils statistiques dédiés à ce type d'analyse. Cette introduction va permettre dans un premier temps de mieux appréhender le contexte politique et économique dans lequel s'inscrit cette étude. Ensuite nous allons aborder le cas très particulier de l'approvisionnement énergétique des îles en présentant plus particulièrement celui de la Corse. Enfin nous expliquerons en quoi la prédiction de production d'énergie électrique d'origine solaire est importante pour un gestionnaire de réseau électrique.

1. Problématique des énergies renouvelables

Le monde s'est industrialisé et les besoins en énergie se multiplient de façon exponentielle pour soutenir tant l'évolution économique que les besoins en termes de confort et de consommation des populations. Actuellement, nous arrivons à un moment critique de l'exploitation en énergie : nous réalisons la fragilité et l'incohérence de notre fonctionnement.

En effet, les ressources de la planète en sédiment fossile s'épuisent, le pétrole se raréfie et, outre les conséquences économiques, force est de constater que soit il faut trouver des alternatives aux sources énergétiques actuelles, soit trouver une alternative à notre mode de civilisation même. Sans énergie, c'est tout notre quotidien qui disparait.

De plus, l'exploitation des énergies fossiles pose un autre problème : les répercussions sur l'environnement sont massives. Si cela a longtemps été ignoré, la préservation de l'environnement devient un enjeu planétaire, avec, là encore des enjeux économiques non négligeables. Le milieu associatif environnemental est depuis longtemps sensibilisé à ce problème, rejoint par les politiques, et depuis peu par l'opinion publique. Les catastrophes écologiques et industrielles plongent les populations dans une prise de conscience violente des dangers que génère l'impact de l'humanité sur notre planète.

Nous allons voir dans cette section l'évolution des mentalités et des politiques depuis trente ans face à ces difficultés et à cette demande toujours croissante en énergie. Nous nous focaliserons ensuite sur les répercussions de ces différentes directives au niveau international, national et régional.

1.1. L'enjeu énergétique au niveau politique

Pour la première fois, l'environnement devient réellement un enjeu international en juin 1972, lors de la conférence des nations unies de Stockholm. Une série de principes pour une gestion politique écologiquement raisonnable est énoncée [Cognet et al. 2004]. Cela marque le début d'un dialogue entre pays industrialisés et pays en voie de développement sur le lien qui existe entre la croissance économique, la pollution et le bien-être des peuples [Multon 2003].

C'est à Genève, en 1979, qu'a lieu la première conférence mondiale sur le climat. Elle est le siège du lancement du programme de recherche climatologique mondial. Ce dernier est confié à l'Organisation Météorologique Mondiale (*OMM*), au Programme des Nations Unies pour l'Environnement (*PNUE*) et au Conseil International des Unions Scientifiques (*CIUS*).

Six ans plus tard, en 1985, il est établi lors de la convention de Vienne que la protection de la couche d'ozone devient une priorité. Les états parties s'engagent à protéger la couche d'ozone et à coopérer scientifiquement afin d'améliorer la compréhension des processus atmosphériques.

Ce n'est qu'en 1992, au sommet de la Terre de Rio de Janeiro, que les Nations Unies adoptent la Convention Cadre des Nations Unies sur les Changements Climatiques (*CCNUCC*). Cette conférence fait le point sur la dégradation de la situation concernant la

pollution et met à jour la corrélation entre le progrès économique à long terme et la nécessité d'une protection de l'environnement. L'objectif de la déclaration est de stabiliser les concentrations atmosphériques des Gaz à Effet de Serre (*GES*) en prenant en compte la responsabilité différenciée des pays industrialisés et des pays en voie de développement.

Il faut attendre cinq ans pour voir les premiers effets de cette conférence. En décembre 1997 à Kyoto (Japon), *l'ONU* ratifie le « protocole de Kyoto » qui prévoit une réduction moyenne de 5,2 % des émissions des *GES* dans les pays industrialisés d'ici 2012. Ce protocole n'est entré en vigueur qu'en février 2005 lorsque la Russie a enfin accepté de le ratifier[1].

En avril 2007, le *GIEC* (Groupe d'experts Intergouvernemental sur l'Evolution du Climat) rend une synthèse alarmante des conséquences du réchauffement climatique. Un chiffre consternant y est avancé : il semblerait que 20 à 30 % des espèces végétales et animales soient déjà éteintes ou menacées d'extinction.

En décembre 2008, à Poznań a lieu la 14^{ième} conférence des états partis à la *CCNUCC*. Il s'agit d'un forum de discussion politique pour attirer l'attention du public et des politiques sur les changements climatiques. L'un des principaux objectifs de cette conférence est de s'assurer de l'application du protocole de Kyoto en travaillant sur les points de résistances ou de difficultés existants.

En décembre 2009, lors de la conférence de Copenhague, les 192 pays représentés renégocient un accord international sur le climat remplaçant le protocole de Kyoto. Selon le secrétaire général de l'*ONU*, elle aboutit au « premier accord réellement mondial » visant à réduire de moitié les émissions de gaz à effet de serre d'ici 2050 par rapport à celles de 1990 [Ben Ahmed et al. 2003].

Nous venons de voir que pour parvenir à envisager un réel positionnement politique mondial face à la crise écologique, plus de 30 années ont été nécessaires. Cela témoigne des difficultés rencontrées par les politiques pour parvenir à articuler des enjeux divergents. En effet, tant au niveau des nations qu'individuellement, la mise en balance des problématiques environnementales et des enjeux politiques et économiques n'a pas toujours été facile et évidente.

[1] Notons que les États-Unis et l'Australie refusent toujours d'intégrer ce protocole.

1.2. Les énergies renouvelables dans le contexte international

Actuellement la principale source d'énergie au niveau mondial, provient des combustibles fossiles conventionnels (pétrole, gaz naturel et charbon) qui cumulent deux aspects négatifs importants : ils sont présents sur terre en quantité limitée, et émettent des *GES* lors de leur combustion. C'est pourquoi il est nécessaire de trouver des alternatives à leur exploitation car, comme nous l'avons vu, les besoins en électricité s'accroissent avec les évolutions technologiques, l'industrialisation et les exigences de qualité et de confort de vie. Les sources d'énergies renouvelables présentent l'avantage d'être disponibles en quantité illimitée et d'être non émettrices de *GES* excepté lors de la fabrication des procédés d'exploitation [Observ'ER 2009]. Leur utilisation est le moyen de conjuguer production d'énergie et respect de l'environnement. Ainsi à l'échelle mondiale l'électricité d'origine renouvelable provient de six sources [Ministère de l'Ecologie 2009a, *EDF* 2006, BRGM 2006 et Energie21 2008] réparties comme suit :

- 84,3 % d'**hydraulique** : l'énergie électrique est produite par la transformation de l'énergie cinétique de l'eau en énergie électrique par l'intermédiaire d'un rotor alternateur relié à un ensemble mécanique situé autour de la roue motrice : la turbine ;

- 7 % d'**éolien** : cette énergie est issue des déplacements de masse d'air à l'intérieur de l'atmosphère (hors offshore). C'est l'énergie mécanique de ces déplacements de masse d'air qui est à la base de l'énergie éolienne par l'intermédiaire de pales reliées à un rotor ;

- 6,3 % de **biomasse** : il s'agit d'énergie stockée sous forme organique grâce à la photosynthèse. Elle est exploitée après combustion ;

- 1,7 % de **géothermie** : il existe dans la croûte terrestre un gradient de température qui est en moyenne de 3°C par 100 mètres. La géothermie vise à étudier et exploiter ce phénomène d'augmentation de la température en fonction de la profondeur. L'exploitation de cette ressource consiste à extraire l'énergie géothermique contenue dans le sol pour l'utiliser sous forme de chauffage ou pour la transformer en électricité à l'aide de turbines ;

- 0,6 % de **solaire** : l'énergie photovoltaïque (ou énergie *PV*) est basée sur l'effet photoélectrique. Celui-ci permet de créer un courant électrique continu à partir d'un

rayonnement électromagnétique. Le soleil émettant ce type de rayonnement, cette ressource a donc l'avantage d'être inépuisable et utilisable en tout point d'un territoire (au moins durant les 4,6 Milliards d'années qui arrivent…) ;

- 0,01 % d'**énergies marines** : les sources d'énergie que l'on peut utiliser provenant des mers sont multiples. La liste exhaustive est : l'énergie marémotrice (issue du mouvement de l'eau créé par les marées), l'énergie des vagues (utilisant la puissance du mouvement des vagues), l'énergie thermique des mers (produite en exploitant la différence de température entre les eaux superficielles et les eaux profondes des océans), les éoliennes offshore, et l'énergie osmotique (diffusion libre des particules par l'arrivée d'eau douce dans l'eau salée de la mer).

Malheureusement, ces sources d'électricité d'origine renouvelable sont, pour la plupart, à production intermittente. En effet, pour illustrer ce phénomène, on peut remarquer que l'énergie solaire ne peut être exploitée la nuit et que son potentiel est plus faible en hiver qu'en été. Pour parer à cet inconvénient et afin de maîtriser la régulation des flux électriques sur le réseau, ces sources doivent être couplées avec des systèmes de stockages. Les piles à hydrogène, les batteries ou les volants d'inertie sont des exemples de couplage possibles. A cause de tous ces inconvénients et de la difficulté à s'en affranchir, la production d'électricité renouvelable n'atteignait guère plus de 19 % de la production d'électricité mondiale en 2009. Cette part reste supérieure à la production d'origine nucléaire (14 %) mais très largement inférieure à l'électricité produite à partir des combustibles fossiles (67 %).

Les difficultés dues à la montée du prix du pétrole, à sa raréfaction, à l'impact écologique des énergies fossiles et les récents problèmes liés aux réacteurs nucléaires, prédisent une croissance des filières d'énergies renouvelables. Cette prévision s'appuie sur la dynamique en cours, tant dans les pays industrialisés que dans les pays à croissance rapide. La mondialisation de l'exploitation des énergies renouvelables permet à ces filières d'être compétitives et de gagner en efficacité. Cependant, du fait de leur intermittence, elles doivent être utilisées en complémentarité avec d'autres systèmes de production. Même si cela peut paraître paradoxal, cela facilite leur intégration dans les réseaux déjà en place en limitant le facteur concurrentiel, qui serait plus important si elles pouvaient s'y substituer totalement.

Chaque nation est laissée libre de transposer les grandes directives mondiales que l'on vient de spécifier précédemment, en fonction de son propre cadre législatif. Nous allons voir dans le cas de la France comment cela s'organise.

1.3. Les énergies renouvelables dans le contexte national

La France s'inscrit dans la dynamique européenne et mondiale, en s'engageant à réduire ses émissions de CO_2 et à développer des énergies renouvelables [LEADER 1999]. Ainsi, en 2001 la France s'est fixée comme objectif, dans le cadre d'une directive européenne, d'obtenir 21 % de sa consommation d'électricité à partir d'énergies renouvelables à l'horizon 2010 [CRE 2009]. Cet objectif a été conforté par la table ronde du Grenelle de l'environnement qui a eu lieu les 24 et 25 octobre 2007. Il a été défini un certain nombre de propositions, de mesures et d'annonces, selon trois priorités que sont la lutte contre le réchauffement climatique, la protection de la biodiversité et la réduction des pollutions. Ces engagements sont formalisés dans la loi Grenelle I (publiée le 3 août 2009) qui est une loi d'orientation intitulée « loi de programmation relative à la mise en œuvre du Grenelle de l'environnement » [Ministère de l'Ecologie 2009b]. Elle a été complétée en Juillet 2010 par la loi dite Grenelle II portant engagement national pour l'environnement[2]. Elle décline le projet environnemental plus concrètement, par objectif, chantier et secteur.

Malgré l'objectif des 21 % et sa mise à jour lors de ces tables rondes du Grenelles qui recommandent une augmentation de la production d'énergies renouvelables de 20 Mtep (millions de tonnes équivalent pétrole), soit 23 % de la production d'énergie à l'horizon 2020 [Livre Blanc 1997], la France ne produit qu'environ 14 % de son énergie à partir de sources renouvelables dont 75% par le bois et l'hydraulique.

Malgré des taux de croissance annuels voisins de 50 %, l'éolien reste très peu développé : environ 11 % de la production électrique renouvelable, soit une production électrique en 2009 de 7891 GWh [Obser'ER 2010]. La France est parmi les mauvais élèves européens en matière de surface solaire installée par habitant : 511 MWc en 2010 [Obser'ER 2010]. Des dispositifs ont été mis en place afin de pallier à ce problème, ainsi des aides visent à améliorer la situation, comme les crédits d'impôt. Le principe du tarif d'achat (prix du kWh

[2] Le projet de loi Grenelle III fut abandonné en 2008 au profit de mesures déjà intégrées dans d'autres lois comme la loi de Finances 2009 ou la loi Grenelle II.

électrique renouvelable fixé à l'avance pour une durée déterminée) a été retenu pour soutenir les producteurs et investisseurs et encourager l'émergence de nouvelles technologies. Cependant, depuis septembre 2010 les tarifs de rachat ont baissés de 12% pour les projets professionnels. Le nombre de projets continue tout de même d'augmenter. La régulation de ces projets passe par l'expertise de différentes instances. Parmi les institutions garantes du respect de l'environnement, on peut citer le Comité de Liaison des Energies Renouvelables (*CLER*), l'Agence De l'Environnement et de la Maîtrise de l'Energie (*ADEME*) ou encore la Fédération Française des Energies Renouvelables (*FFER*).

1.4. Les énergies renouvelables dans le contexte régional

En France, le Schéma Régional du Climat, de l'Air et de l'Energie (*SRCAE*) est l'un des grands schémas régionaux créés par les lois Grenelle I et Grenelle II (Article 68). Il décline aussi aux échelles régionales une partie du contenu de la législation européenne sur le climat et l'énergie. Ce schéma doit intégrer, dans un seul et même cadre, divers documents de planification en lien avec l'énergie et le climat, notamment les schémas éoliens et les schémas de services collectifs de l'énergie [*EDF* 2009 ; *EDF*-SEI 2009a]. Ce projet fixe des objectifs de maîtrise de l'énergie à l'horizon 2020 et 2050 dans le but d'atténuer les effets du changement climatique et de s'y adapter. Le grand challenge est de diviser par quatre les émissions françaises de gaz à effet de serre entre 1990 et 2050. Ce schéma régional du climat, de l'air et de l'énergie représente aussi indirectement un schéma régional des énergies renouvelables.

Au niveau régional, le Plan Climat-Energie Territorial (*PCET*) peut être intégré au *SRCAE*. Il est élaboré « en tenant compte des aspects économiques et sociaux » et comporte notamment un volet sur l'évaluation régionale du potentiel énergétique, renouvelable et de récupération. Pour aider à la mise en œuvre de ce plan, il existe un « comité de suivi des énergies renouvelables » qui est créé, au sein du conseil supérieur de l'énergie, pour « évaluer la progression vers l'objectif de 23 % d'énergies renouvelables dans la consommation d'énergie finale en 2020 ». Les gestionnaires des réseaux publics doivent dès lors, élaborer un schéma régional de raccordement au réseau des énergies renouvelables. Cette étape est délicate et nécessitent de multiples investigations pour parer aux difficultés liées à

l'intermittence de ce type de ressource. De plus, les collectivités quant à elles, doivent veiller à la bonne coordination des différents plans de développement des réseaux d'énergie.

Sous certaines conditions, les départements et les régions peuvent, sur leurs territoires, aménager, exploiter, faire aménager et faire exploiter des installations de production d'électricité en bénéficiant de l'obligation d'achat de l'électricité. Outre ces institutions, on peut citer la Commission de Régulation de l'Energie (*CRE*) qui est une autorité administrative indépendante en charge du bon fonctionnement des marchés de l'énergie et de l'arbitrage des éventuels conflits entre utilisateur et exploitant. A ce titre, elle lance régulièrement des appels d'offre *PV*/stockage et éolien/stockage pour installer en région des systèmes de production renouvelable.

2. Les énergies renouvelables et les spécificités insulaires : le cas de la Corse

L'approvisionnement énergétique des îles est un cas très particulier. En effet, les réseaux électriques insulaires ont la particularité de devoir se suffire à eux-mêmes car ils ne sont pas, ou peu, connectés au réseau continental [*EDF*-SEI 2009a ; *EDF* 2010]. Ainsi chaque région doit produire la totalité (ou presque) de l'électricité qu'elle consomme sur place.

A ce titre, la Corse est un système insulaire non interconnecté, comme les DOM-TOM. La puissance électrique totale installée est d'environ 700 MW en 2011. Les tarifs sont identiques à ceux de la métropole mais les coûts de production sont plus élevés. Pour compenser ce décalage, une taxe est répartie sur l'ensemble des consommateurs français, taxe récoltée par la *CRE*. En tant que gestionnaire du système, *EDF* assure une mission de service public sur la sécurité, la qualité et les moyens nécessaires à l'approvisionnement ainsi que sur la politique de recherche et de développement et la protection de l'environnement.

Les zones non interconnectées ont une fragilité structurelle. En effet, chaque incident peut provoquer un déséquilibre et entrainer une forte perturbation sur le réseau. Ces creux de tension génèrent un risque de coupure, c'est pourquoi il est nécessaire d'avoir des réserves de puissance permanentes. Du fait d'une croissance de la consommation en énergie, la gestion des *ER* (30 MW en Corse [Observ'ER 2010]) dans ces territoires est un enjeu énergétique majeur. La Corse, bien que reliée à l'Italie (*SACOI* 50 MW) et à la Sardaigne (*SARCO* 80

MW) par câble électrique, doit donc produire la presque totalité (70%) de sa consommation électrique qui a atteint un pic de 477 MW au début de l'hiver 2010.

Le réseau électrique corse se caractérise, comme presque toute zone insulaire, par sa petite dimension, sa sensibilité aux variations de production et sa capacité restreinte en nouveaux moyens de production. Dans ce contexte, et malgré l'ouverture du marché à la concurrence, aucun fournisseur à part *EDF* ne s'est positionné sur le marché corse. *EDF* gère, aussi bien la production, la gestion du réseau que la fourniture au client. Au-delà de sa propre production, il rachète l'ensemble de la production indépendante par le biais des contrats d'obligation d'achat. La limite majeure de cette démarche réside dans le fait que le parc de production n'est pas forcément adapté aux besoins actuels ni a fortiori futurs, compte tenu des prévisions de croissance des consommations. La Corse connait une forte croissance économique à laquelle s'ajoute une forte croissance démographique et touristique [*EDF* 2010], entrainant ainsi une augmentation de la demande en électricité. La consommation électrique en Corse a connu une croissance de 3 % ces cinq dernières années sauf en 2009 où elle était de 2,4 %. Comparativement, la consommation électrique sur le continent a connu une baisse de 1,6 % en 2009, et une croissance de 1 % seulement sur les cinq dernières années. Pour faire face à ces besoins croissants en énergie, il est donc nécessaire d'investir dans de nouvelles capacités de production et de gestion du réseau, en intégrant le plan national de développement des énergies renouvelables qui a pour objectif de porter à 23 % la part des énergies renouvelables dans la consommation d'énergie à l'horizon 2020 (section 1-3 de l'introduction).

La Corse, comme les DOM, présente un très fort potentiel d'ensoleillement et éolien. Ces territoires insulaires se prêtent donc au développement de ces énergies. Le parc de production électrique de la Corse se compose majoritairement de centrales hydrauliques et thermiques, le reste de l'approvisionnement provenant essentiellement des interconnexions avec l'Italie et la Sardaigne. Les moyens thermiques sont constitués de centrales diesel et de turbines à combustion, utilisées respectivement en base (186 MW) et en pointe (+115 MW). Le parc hydraulique est lui composé de centrales totalisant 139 MW auxquels viendront s'ajouter les 55 MW du barrage du Rizzanese, et de nombreuses centrales de petite hydraulique totalisant 22 MW. Quant au parc éolien, il totalise une capacité installée de 18 MW. Le parc photovoltaïque de Rapale (opérationnel depuis juillet 2010) représente une puissance installée de 7,7 MW. Avant sa mise en place la puissance *PV* sur l'île était

inférieure à 300 kW. L'unique centrale biogaz de la Corse se situe sur le site de Tallone et est opérationnelle depuis 2009. Elle représente une capacité installée de 1,7 MW. Au total, la Corse se place au 21e rang des régions ayant la plus grande puissance installée. La répartition de cette puissance est donnée par la Figure 0-1.

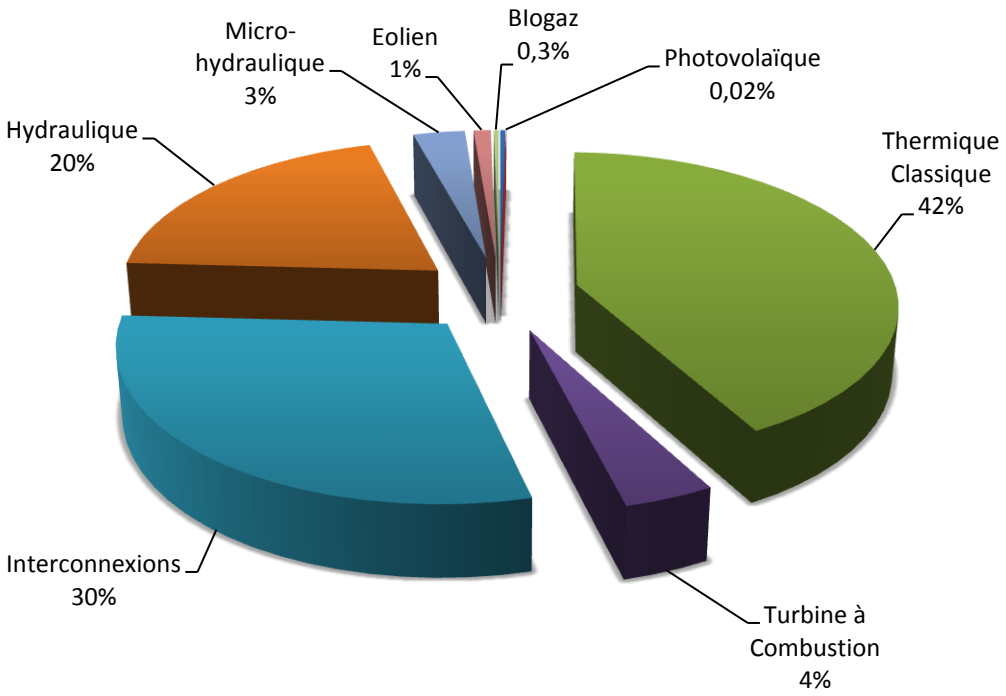


Figure 0-1 : répartition de l'énergie produite par sources début 2010

Notons qu'un projet de raccordement de la Corse à l'Algérie par le biais du gazoduc *GALSI* (Gazoduc Algérie-Sardaigne-Italie) est à l'étude et permettrait de remplacer la production thermique diesel par une production thermique au gaz moins polluante. En Corse, le tarif de rachat de l'électricité *PV* est fixé à 40 Cts/kWh pour les 20 prochaines années[3]. C'est sans doute pour cette raison que 74 projets de champs *PV* ont été déposés auprès de l'Agence de Développement Economique de la Corse (*ADEC*) depuis Juin 2009. Cette dernière a pour missions d'aider à la création et au développement d'entreprises et d'emplois en Corse. Le potentiel de tous ces projets réunis se monte à une puissance installée théorique de 257 MW. Ce chiffre est en parfaite contradiction avec la part maximale des 90 MW correspondant à 30 % *d'ER* (hors hydraulique) à injecter dans le réseau électrique (arrêté du 23 avril 2008 fixant la limite de la puissance *ER* produite). Dans cette optique, *EDF* a la

[3] Le tarif est de 30 Cts/kWh sur le continent.

possibilité de déconnecter des producteurs privés afin de préserver la stabilité du réseau [*EDF* 2009]. En effet, *EDF* ne garantit le rachat de l'électricité sans déconnexion que pour ces 90 MW [Haurant et al. 2010]. Ce plafond est fixé pour préserver la sûreté des systèmes électriques insulaires. Compte tenu des variations météorologiques, le solaire, tout comme l'éolien, est une énergie dont il est difficile de garantir la puissance à l'avance. Une variation subite de la météo peut provoquer une chute de la puissance disponible, ce qui peut conduire à un écroulement du réseau. Lorsque la puissance *ER* atteinte dépasse le seuil des 30 % et que cela va durer, le chargé de conduite[4] d'*EDF* [*EDF*-SEI 2009b] émettra, via le système de téléconduite[5], les demandes de déconnexion, selon une règle simple : le dernier arrivé en file d'attente sera le premier à être déconnecté.

Ainsi, au sein des réseaux électriques insulaires, il semble primordial pour le gestionnaire électrique de maîtriser la prévision des ressources *ER* et de s'équiper de moyens performants de stockage de l'énergie. La maîtrise de la prévision pourrait certainement permettre d'augmenter le seuil des 90 MW.

3. Pourquoi prédire la production d'énergie électrique d'origine solaire

Comme nous l'avons vu précédemment, l'intégration des *ER* au sein de réseaux électriques est une problématique importante. Dans le cadre insulaire, celle-ci l'est d'autant plus que les moyens de production sont limités et locaux. Une partie de la solution pour l'intégration des énergies renouvelables intermittentes au sein de réseau électrique réside donc dans la prédiction des différentes ressources.

3.1 Problématique de l'intermittence

A cause de leur caractère aléatoire et intermittent, les *ER* ne peuvent être insérées que de manière limitée au sein d'un réseau afin de préserver au mieux la sécurité de ce dernier et une qualité de fourniture. Pour cela, une solution est d'utiliser en parallèle d'autres moyens de

[4] Personne en charge de la bascule *ER* – énergie conventionnelle.
[5] Installation complexe qui assure l'interconnexion des lignes et la transformation des niveaux de tension.

production d'énergie (hors *ER*) avec comme principale difficulté la gestion optimale de la bascule. Dans un futur proche, le gestionnaire de réseau devra disposer d'outils de prédiction des productions éoliennes et photovoltaïques. A terme, la prédiction de la production électrique pourra être un outil décisionnel pour les producteurs et le gestionnaire de réseau, que ce soit dans le cas des systèmes insulaires ou dans le cas général des marchés de gros électriques (lorsque seuls des fournisseurs concurrents produisent pour des revendeurs d'électricité). Pour les petits réseaux insulaires, la prédiction de production intermittente sera probablement couplée avec des systèmes de stockage. De plus, une réduction de l'intermittence et de la variabilité de la production pourra être assurée par une multiplication de sources éloignées, procédé appelé foisonnement géographique. Ainsi, le stockage, le foisonnement et la prédiction des énergies intermittentes devraient permettre une meilleure intégration des *ER* dans les réseaux électriques. Leur prédiction permettra de prévoir les réserves énergétiques qui prendront le relais pour préserver la sécurité du système.

Le développement des *ER* en milieu insulaire a toute sa place, il s'agit de l'intégrer correctement dans un mix énergétique. Différents projets impliquant l'Université de Corse, et plus particulièrement le laboratoire *SPE* (UMR CNRS 6134) sont actuellement lancés sur cette thématique. On peut citer :

- le partenariat avec *EDF* sur la caractérisation de l'énergie photovoltaïque et éolienne à la fois sur le volet prévision et sur le volet stockage (signature prévue en octobre 2011) ;

- les projets de plateformes de stockage de l'énergie appelées *PAGLIA ORBA* [2011] (Plateforme Avancée de Gestion éLectrique en milieu Insulaire Associant stOckage et énergie Renouvelable oBjectif Autonomie) et *MYRTE* [2006] (Mission hYdrogène Renouvelable pour l'inTégration au réseau Electrique) ;

- la réponse à l'appel d'offre (en cours) *ADEME* sur les smart grids (l'informatique au service du réseau de distribution d'électricité) : http://www2.ademe.fr ;

- la réponse à l'appel d'offre (en cours) concernant l'éolien et son stockage : http://www.cre.fr.

3.2 Quels horizons de prédiction pour le gestionnaire de réseau ?

Du point de vue du gestionnaire de réseau (*EDF*), les besoins en termes de prédiction de la ressource solaire et des ressources intermittentes en général sont très variés. On peut les distinguer suivant l'horizon envisagé : la ressource qui sera disponible les jours suivants (j+1, j+2 et j+3), le lendemain par pas horaire (h+24), dans 1 heure (h+1), et dans 5 min (m+5). Ces différents horizons permettent d'appréhender les différents aspects de la prédiction : le moyen terme, le court terme et le très court terme. La prédiction j+1 est importante pour le gestionnaire car elle a une application industrielle immédiate. En effet, dans le cadre de l'interconnexion *SACOI* et *SARCO*, il faut être capable d'organiser la gestion des stocks fossiles. L'achat de l'interconnexion se fait à j+1 voire j+2. Concernant l'horizon h+1, il correspond notamment au délai d'allumage des dispositifs thermiques du gestionnaire de réseau (après commande au chargé de conduite *EDF*). En fait, la mise en œuvre des moyens de production thermique *EDF* dure environ 30 min, il faut donc être capable de prédire les baisses d'énergies intermittentes au moins 1 heure à l'avance. De plus, cette échelle correspond au pas de temps des bases de prévisions Météo-France, qui sont déjà utilisées pour caractériser certains paramètres météorologiques en lien direct avec la production d'énergie. Pour ce qui est de la prédiction h+24, elle revêt un intérêt mêlant les deux horizons que l'on vient de présenter. La connaissance 24 heures à l'avance par pas horaire de l'*ER* permet une meilleure gestion des stocks fossiles, et une anticipation sur la journée du lendemain, en connaissant les moments critiques où le gestionnaire de réseau devra être vigilent. Enfin, l'horizon quelques minutes (entre m+1 et m+5) concerne les moyens de production hydraulique et ceux liés à l'interconnexion qui sont très rapidement exploitables. En effet, il suffit de quelques minutes pour que l'électricité soit disponible. Le schéma ci-dessous récapitule le lien qui relie horizon de prédiction et moyens de production.

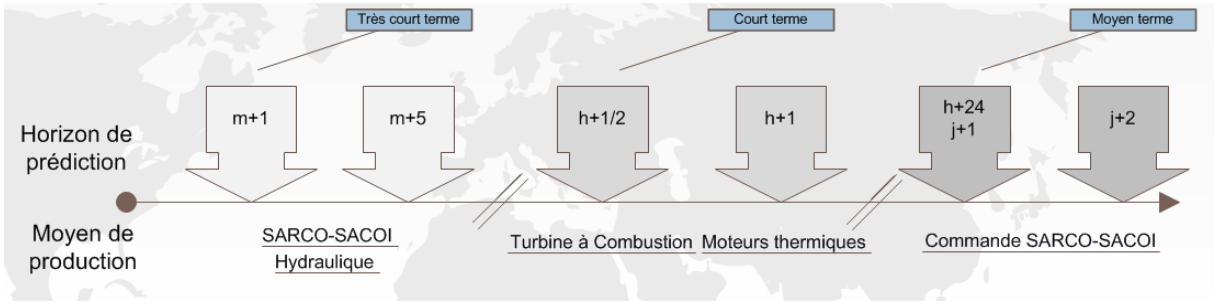


Figure 0-2 : horizons de prédiction et moyens de production associés

Afin de gérer au mieux l'équilibre entre production d'électricité et consommation, la prévision à courte et très courte échéance (quelques heures à quelques minutes) de l'énergie produite par les fermes photovoltaïques comme la centrale *MYRTE,* est cruciale [Darras 2010]. Cette thématique intéresse de nombreux acteurs parmi lesquels de nombreux producteurs et/ou exploitants mais aussi les collectivités locales et donc indirectement les utilisateurs.

4. Organisation du manuscrit

Cette étude s'inscrit dans le cadre de la caractérisation de la ressource solaire et *PV* dans le but de quantifier l'énergie disponible. Nous avons vu dans la section précédente que cette démarche est essentielle mais qu'elle revêt différentes modalités d'approche comme notamment l'horizon considéré. Nous avons ainsi essayé de tester différentes techniques de prédiction sur quatre horizons différents susceptibles d'intéresser un gestionnaire de réseau : j+1, h+24, h+1 et m+5.

Tout d'abord, nous décrirons dans le premier chapitre le phénomène de rayonnement global à la surface de la Terre, en exposant les faits théoriques qui permettent de comprendre les différents éléments qui le composent. Nous verrons que comme la périodicité (annuelle et journalière) est régie par une composante déterministe, il est possible de fonder des modèles dits « ciel clair » (en l'absence de nuage) du rayonnement global. L'utilisation de cette ressource et sa conversion *PV* seront aussi expliquées, permettant d'introduire les grandeurs de puissance pré et post-onduleur (P_{dc} et P_{ac}) qui seront utilisées par la suite.

Dans le second chapitre, seront introduits la théorie et le formalisme des séries temporelles. Cela servira de base à la prédiction de l'énergie solaire ou *PV* dans ce manuscrit. Nous verrons notamment des prédicteurs de natures différentes comme les réseaux de neurones, les autorégressions à moyenne mobile, les k plus proches voisins, les inférences Bayesiennes ou encore les chaînes de Markov. Nous verrons également comment les comparer sur la base de tests statistiques pertinents.

Ensuite, nous présenterons, dans le chapitre 3, la méthodologie de prédiction adoptée dans ce manuscrit. Nous avons expérimenté différentes approches et avons essayé de coupler les aspects positifs de chacune. On exposera le principe de la stationnarisation et comment

nous l'avons utilisée, l'hybridation des réseaux de neurones avec des modèles d'autorégressions, l'utilisation de mesures et de prévisions météorologiques en compléments des réseaux de neurones.

Nous énoncerons dans le quatrième chapitre les résultats de prévision obtenus pour les quatre horizons étudiés. Nous commencerons par l'analyse de l'horizon j+1 avec la comparaison des différents prédicteurs de références, puis nous étudierons l'horizon h+1 en tenant compte des résultats précédents. Ensuite, nous dégagerons les résultats issus des deux derniers horizons que sont h+24 et m+5. Si dans les deux premiers cas les techniques de prédictions étaient relativement équivalentes, nous verrons que ces deux horizons de prédictions nécessitent une adaptation de la méthodologie.

Enfin, nous conclurons cette étude en reprenant tous les résultats issus des manipulations et en essayant de dégager une synthèse des différents prédicteurs ou méthodes de prédiction en fonction de l'horizon considéré. L'objectif est de trouver une approche générique pour l'ensemble des horizons étudiés. Cependant, comme nous le verrons, cet objectif n'est peut-être pas encore réaliste et dépend de l'ordre de grandeur des erreurs de prédictions souhaité. Tout au long de cette thèse, nous nous efforcerons de hiérarchiser les différents prédicteurs pour chaque horizon étudié. Cette dernière partie servira de récapitulatif pour la comparaison des prédicteurs. Il nous a semblé important de pouvoir proposer des règles de prédiction que d'autres équipes de recherche pourraient exploiter. A partir de ces conclusions, nous verrons quelles perspectives sont envisageables et intéressantes pour la poursuite des travaux.

5. Bibliographie

Ben Ahmed H, Multon B, Robin G, et Gergaud O. Consommation d'énergie et ressources énergétiques, Techniques de l'Ingénieur, traité D3900, 2003.

BRGM, rapport de la compagnie Française de Géothermie. Usine géothermique de Bouillante, 2006.

Cognet G, Hesto P, Houzelot JL, Rombaut C et al. L'énergie au XXIème siècle, une réflexion prospective, Ministère Délégué Recherche et Nouvelles Technologies, 2004.

CRE. Cahier des charges de l'appel d'offres portant sur des installations au sol de production d'électricité à partir de l'énergie solaire. Technical report, Ministère de l'Ecologie, de l'Energie, du Développement durable et de l'Aménagement du territoire, 2009.

Darras Christophe. Modélisation de systèmes hybrides Photovoltaïque/hydrogène. Thèse de l'université de Corse, 2010.

EDF. Bilan 2009 des activités d'*EDF* en Corse. Technical report, 2010.

EDF-SEI. Corse : Bilan Prévisionnel Pluriannuel Investissements en Production. Technical report, Direction SEI, 2009.

EDF-SEI. Corse : Seuil de 30 % d'insertion des ENR. CCP SEI du 19 Mai 2009.

EDF. Direction Participations & Activités Nouvelles. La biomasse : production d'électricité, 2006.

ENERGIE 21 Centre France. L'éolien, énergie du futur, 2008.

Haurant P, Muselli M, Oberti P, Pillot B et al. Multicriteria Decision Aiding for Selection of Photovoltaic Plants on Farming Fields in Corsica. Dans: EU PVSEC Proceedings. Valencia, Spain, p 5267-5270, 2010.

LEADER/AEIDL. Observatoire européen « Sources d'énergies renouvelables, source de développement durable », 1999.

Livre Blanc de la Commission Européenne, Communication de la Commission, « Énergie pour l'avenir : Les sources d'énergies renouvelables », Livre Blanc établissant une stratégie et un plan d'action communautaires, 1997.

Ministère de l'Ecologie, de l'Energie, du Développement durable et de l'Aménagement du territoire, « Les grands barrages hydroélectriques », Publication DRIRE Limousin, 2009.

Ministère de l'Ecologie, de l'Energie, du Développement durable et de la Mer, « Le Livre Bleu des engagements du Grenelle de la Mer », 2009.

MYRTE. Avant-Projet Solaire du projet *MYRTE*, Université de Corse, document interne, diffusion restreinte, 2006.

Multon B. Production d'énergies électrique par sources renouvelables. Techniques de l'Ingénieur, traité D4005, 2003.

PAGLIA ORBA. Plateau Technique Solaire de Vignola : Micro-réseuau et stockage de l'énergie électrique Paglia Orba, Université de Corse, document interne, diffusion restreinte, 2011.

Observ'ER. La production d'électricité d'origine renouvelable dans le monde. Collection chiffres et statistiques, Onzième inventaire, Edition 2009.

Observ'ER. Le baromètre 2010 des ENR. Collection chiffres et statistiques, Edition 2010.

Chapitre 1

Le rayonnement solaire et la production d'énergie

SOMMAIRE

1

1. Introduction

La prédiction de l'énergie surfacique solaire au sol (aussi appelée rayonnement global), peut se faire sans connaissance experte sur le signal étudié ni sur ses composantes intrinsèques. En effet, certaines théories peuvent *a priori* être appliquées de manière systématique sur la prévision d'indice comme le CAC40, la concentration d'ozone dans l'atmosphère, ou encore le nombre de vente automobile [Lam et al. 2008]. Avant de détailler ce type d'approche dans le chapitre 2, nous présentons ici différentes théories permettant de comprendre et de modéliser le rayonnement solaire. Nous définirons plus particulièrement celles qui ont finalement été retenues durant cette étude. Ensuite nous décrirons de façon précise les différentes étapes qui permettent la production d'électricité à partir du rayonnement solaire. Enfin nous présenterons les données expérimentales, essentiellement météorologiques, qui seront utilisées par la suite dans le cadre de nos expérimentations.

2. Le rayonnement solaire

Il existe de nombreuses théories permettant de comprendre et de modéliser le rayonnement solaire au sol [Kasten 1996 ; Liou 1980 ; Badescu 2008]. Les concepts nécessaires à la compréhension de ce phénomène sont liés à la nature même de ses composantes. En effet, comme nous le verrons, il en existe une que l'on qualifie de composante déterministe, et l'autre qui revêt un caractère aléatoire et qui porte le nom de composante stochastique.

2.1. Composante déterministe et stochastique

Avant d'être une source d'énergie renouvelable, le Soleil est avant tout l'élément essentiel rendant la vie organique possible sur Terre (chaleur, photosynthèse, et création des énergies fossiles). Il est situé au centre de notre système solaire et est composé principalement d'hydrogène (75 %) et d'hélium (25 %). Les proportions varient au fur et à mesure que le Soleil convertit l'hydrogène en hélium par réaction de fusion nucléaire. L'énergie dégagée au sein du Soleil (386 milliards de milliards de MW) est produite par ces réactions de fusion nucléaire. Chaque seconde au cœur du Soleil, environ 700 millions de tonnes d'hydrogène sont converties en 695 millions de tonnes d'hélium et 5 millions de tonnes d'énergie sous forme de rayons gamma. Les conditions de température et de pression au centre du Soleil sont extrêmes : il y règne une température de 15 millions de °C et la pression est de 250 milliards d'atmosphère. Les gaz y sont compressés à une densité équivalente à 150 fois celle de l'eau. Pendant qu'elle voyage vers la surface du Soleil, l'énergie est continuellement absorbée et réémise à des températures de plus en plus basses. De ce fait, lorsqu'elle en atteint la surface, elle est principalement constituée de lumière visible (99,9 % des photons ont une énergie comprise entre 0,2 et 0,8 μm), assimilable à celle d'un corps noir d'une température d'environ 6000 K. Cette énergie émise est appelée rayonnement solaire [Aguiar et al. 1988 ; Badescu 2008]. La puissance émise par le Soleil sous forme de rayonnement électromagnétique est évaluée à $1{,}8.10^{14}$ kW. L'angle solide sous lequel le Soleil est vu de la Terre, est d'environ 6.10^{-5} sr. Ainsi, notre planète n'intercepte environ que deux milliardièmes du rayonnement émis par le soleil. En dehors de l'énergie géothermique, toutes les formes d'énergie disponibles dans le système Terre-atmosphère résultent de la transformation du rayonnement solaire. Au sommet de l'atmosphère, la répartition de cette puissance n'est pas homogène du fait que la Terre est en orbite autour du Soleil, et de l'inclinaison du plan équatorial de la Terre par rapport au plan orbital selon la latitude et les saisons [Bird et Hulstrum 1981 a et b ; Ineichen et al 1990 et Kaplanis 2006]. Avant d'atteindre le sol, le rayonnement incident est fortement influencé par l'atmosphère terrestre (Figure 1-1), à travers des interactions de type absorption et réflexion [Ulgen et Hepbasli 2009].

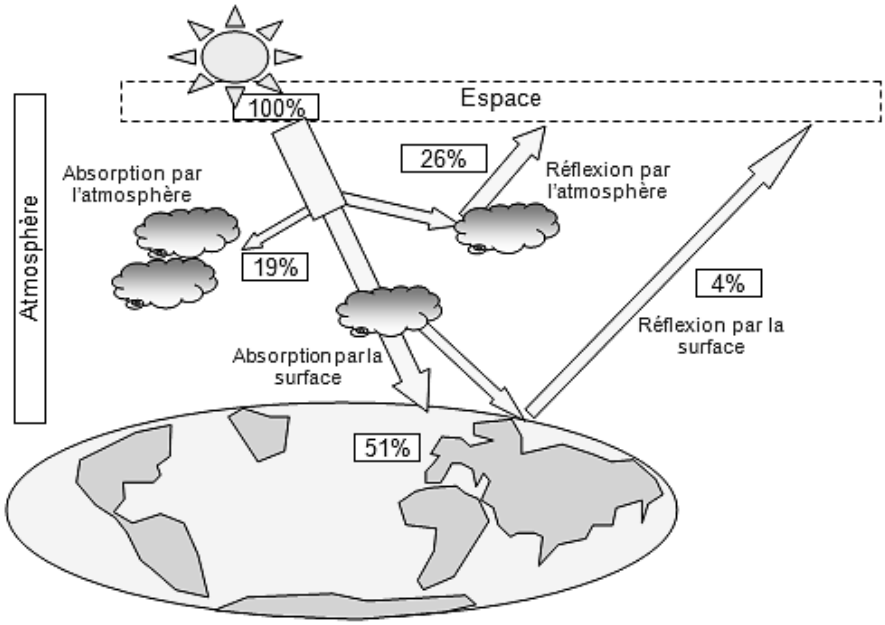


Figure 1-1 : interaction entre le rayonnement solaire et l'atmosphère terrestre

La valeur de l'éclairement en un point du sol et à un instant donné dépend non seulement de la position du Soleil, mais aussi de l'état optique de l'atmosphère [Batlles et al. 2000 ; Rehman 1999]. Ce phénomène permet de scinder en deux parties distinctes le phénomène de rayonnement global au sol [Oumbe 2009 ; Perez et al. 1990] : une composante que l'on nommera « déterministe » et qui ne dépend que de la distance entre le point de mesure et le centre du Soleil, et une composante que l'on nommera « stochastique » qui dépend de l'état de la couverture nuageuse, du taux d'ozone, du taux d'humidité, etc. [Hay et Davies 1978 ; Ineichen 2006 ; Bird et Hulstrum 1981 a et b ; Mellit et al. 2008]. La première composante peut se déterminer aisément avec des règles de mécanique céleste et de physique énergétique [Iqbal 1983 et Jin et al. 2005]. Concernant la seconde, elle est beaucoup plus difficile à appréhender [Klucher 1979 ; Marty et Philipona 2000] du fait purement probabiliste des différents éléments qui la composent : occurrence de nuage, de concentration d'ozone, etc. (Figure 1-1). Le modèle de prévision que l'on propose d'étudier dans cette thèse ne concerne que la seconde composante puisque nous avons considéré la première comme déterminée avec un modèle de connaissance dédié. Au final, en combinant outils statistiques et modèle de connaissance, l'étude sur la prévision de la série temporelle de rayonnement global peut se transposer en une étude de prédiction de la nébulosité comme le montre la Figure 1-2.

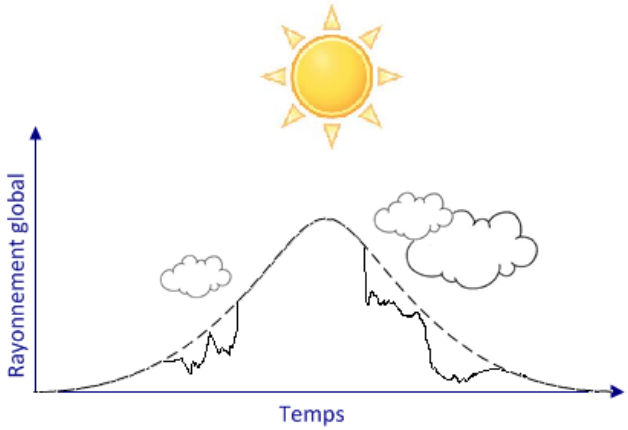


Figure 1-2 : impact de la couverture nuageuse sur la série temporelle de rayonnement global

Avant d'énoncer les modèles sophistiqués de modélisation du rayonnement global, il convient de formuler les définitions et éléments essentiels à la bonne compréhension de ses fondements par le biais de méthodes de calcul du bilan du rayonnement solaire en un point de la surface terrestre.

2.2. Eléments de mécanique céleste et de physique énergétique solaire

Nous allons commencer par énoncer des définitions essentielles à la quantification du rayonnement hors atmosphère nécessaires à la compréhension et à l'explication des rayonnements directs et diffus à la surface de la terre.

a. Distance Terre-Soleil

La Terre subit deux types de déplacement aux répercutions totalement différentes. Le premier est lié au mouvement de révolution de la Terre autour du Soleil de périodicité annuelle, qui permet de comprendre le phénomène des saisons [Badescu 2008 ; Liu et Jordan 1961]. Le second mouvement est lié à la rotation de la Terre sur elle-même autour de l'axe des pôles d'une périodicité de 24 heures. Cette dernière est à l'origine de la succession des jours et des nuits. Il est bien sûr possible de déterminer de façon précise la position du Soleil en

fonction du temps et de la position de l'observateur sur la terre. La Terre tourne autour du Soleil selon une trajectoire elliptique d'excentricité 0,0167 [Oumbe 2009], comme le montre la Figure 1-3.

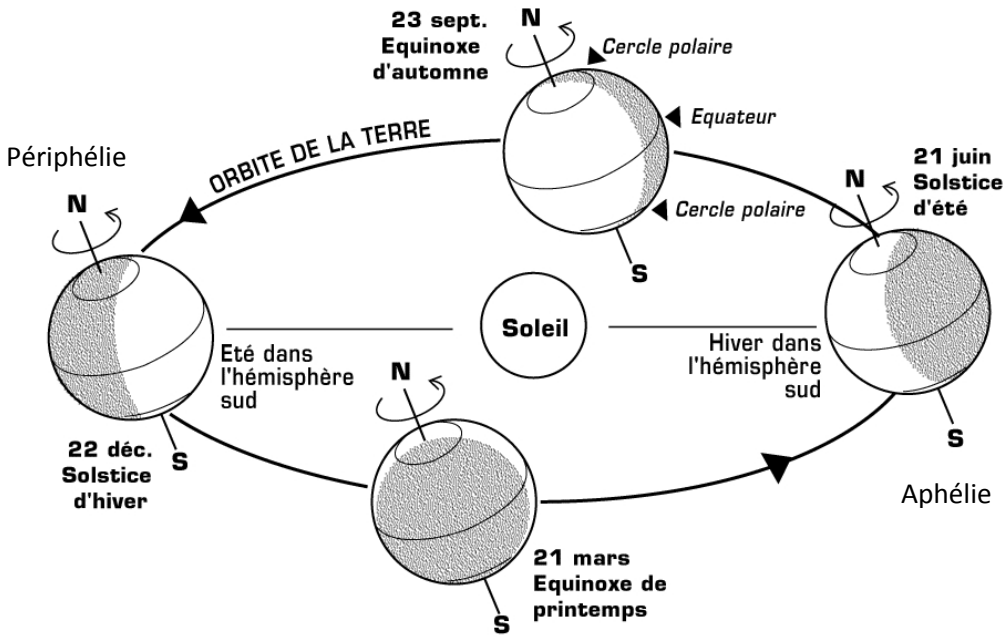


Figure 1-3 : mouvement de révolution de la Terre autour du Soleil

La distance Terre-Soleil varie donc au cours de l'année. En moyenne, la distance Terre-Soleil sert de base à l'« unité astronomique », soit 1 ua correspondant à 150.10^6 km (r_0). Elle atteint son maximum au solstice d'été (Aphélie ; 1,017 ua soit 152.10^6 km) et son minimum au solstice d'hiver (Périhélie ; 0,983 ua soit 147.10^6 km). Elle dépend du numéro j du jour de l'année qui varie de 1 à 365 (ou 366 pour les années bissextiles). La distance Terre-Soleil r est donnée par l'Equation 1-1 [Rigollier et al. 2000].

$$\left(\frac{r}{r_0}\right)^2 = 1 + \epsilon \qquad (1\text{-}1)$$

ϵ désigne la correction relative suivant le jour considéré, et peut être approximé en utilisant la notion d'angle journalier ω_j défini par [Randimbivololona 2010] :

$$\omega_j = j.2\pi/365{,}2 \qquad (1\text{-}2)$$

On remarquera qu'il est possible de définir un angle horaire ω_h sur le même principe, en remplaçant 365,2 par 24 (qui correspond au nombre d'heures dans une journée). En utilisant l'Equation 1-2, on peut redéfinir le rapport $\left(\frac{r}{r_0}\right)$ par [Oumbe 2009] :

$$\left(\frac{r}{r_0}\right)^2 = 1 + 0{,}3344.\,cos(\omega_j - 0{,}049) \qquad (1\text{-}3)$$

La déclinaison solaire notée δ représente l'angle entre la direction Terre-Soleil et le plan de l'équateur au moment du midi solaire. C'est une grandeur qui intervient souvent dans les différents calculs liés au rayonnement global. Elle varie de façon sinusoïdale au cours de l'année entre -23,45 (solstice d'hiver) et 23,45 (solstice d'été) et s'annule aux équinoxes. Elle peut être déterminée pour le jour *j* considéré de la façon suivante [Badescu 2008] :

$$\delta = 23{,}45.\,sin(\tfrac{360(284+j)}{365}) \qquad (1\text{-}4)$$

b. Systèmes de coordonnées horizontales et horaires

Les systèmes de coordonnées horizontales et horaires permettent de déterminer la position des astres tels qu'on les voit depuis la Terre. Ils sont basés sur un concept astronomique appelé sphère céleste. Cette dernière est représentée comme une surface fictive dont tous les points sont situés à une même distance (indéterminée), d'une origine basée sur le centre de la Terre, et sur laquelle les astres sont considérés comme projetés. Les astres se voient alors attribuer deux coordonnées angulaires uniques. A partir de la définition de la sphère céleste, plusieurs systèmes de coordonnées peuvent être définis dont notamment les systèmes de coordonnées horizontales et horaires. Concernant le premier système (Figure 1-4a), il s'agit d'un système de coordonnées locales. Son origine est le centre de la Terre O, et ses plans de référence sont l'horizontal du lieu d'observation (H) et le demi-méridien sud (m). La position d'un objet sur la voute céleste est définie par deux coordonnées : l'azimut *a* et la hauteur *h*, ou son complémentaire la distance zénithale *z* = 90°-*h*. L'azimut *a* est l'angle en degrés entre le demi-cercle vertical passant par l'astre M et le demi-méridien sud. Il est compté de 0 à 360° positivement vers l'ouest. La hauteur *h* est l'angle en degré qui sépare l'astre *M* du plan horizontal (H). Il est mesuré à partir du plan horizontal selon la verticale de l'observateur et prend ses valeurs entre - 90 et + 90°. Le zénith Z est l'intersection de la verticale du lieu d'observation et du demi-méridien sud. Le nord et le sud sont définis par

l'intersection du plan horizontal et le méridien du lieu d'observation. La Figure 1-4a illustre l'ensemble du système de coordonnées horizontales.

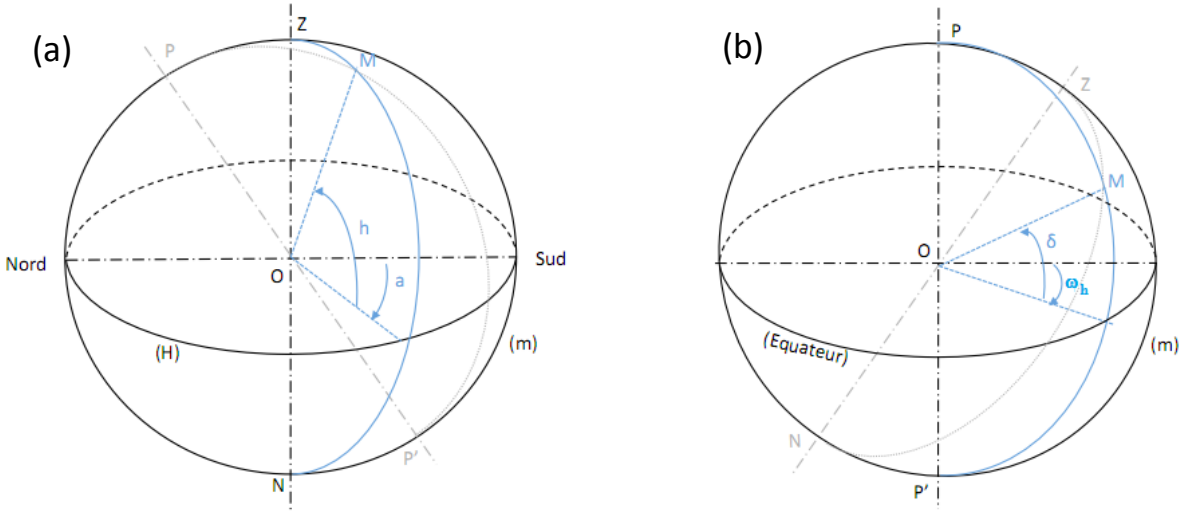


Figure 1-4 : système de coordonnées horizontales (a) et horaires (b)

Le système de coordonnées horaires a également pour origine le centre de la Terre O. Il est déterminé par l'équateur terrestre, l'axe des pôles de la Terre (PP'), et le zénith de l'observateur (Z). Les coordonnées de l'astre M sont définies par la déclinaison δ et l'angle horaire ω_h. Comme nous l'avons vu précédemment, la déclinaison δ peut se définir par l'angle entre la droite OM (reliant le centre de la Terre et l'astre) et l'équateur. Elle prend des valeurs comprises entre - 90 et + 90° à partir du plan équatorial. Le cercle horaire qui contient le zénith de l'observateur est appelé méridien. L'angle horaire ω_h est l'angle entre le méridien et le cercle horaire qui contient l'astre M. Il est compté positivement vers l'ouest à partir du méridien. La Figure 1-4b illustre l'ensemble du système de coordonnées horaires. Une relation liant ces deux systèmes de coordonnées est très utilisée dès lors que l'on étudie les effets radiatifs solaires dans l'atmosphère (Equation 1-5).

$$sin(h) = cos(a) = sin(\varphi)sin(\delta) + cos(\varphi)cos(\delta)cos(\omega_h) \qquad (1\text{-}5)$$

c. Equation du temps, temps solaire vrai et temps local

Dans la pratique, le calcul des coordonnées angulaires est relatif au Temps Solaire Vrai (*TSV*). Souvent, les séries temporelles de rayonnement solaire, sont mesurées en utilisant le

Temps Local (*TL*) qui correspond à ce que donne une horloge ou une montre habituellement. Pour passer d'une base de temps à une autre, il est nécessaire de faire appel à l'Equation du Temps (*ET*). Cette dernière est liée à la chronométrie du soleil. Comme le Soleil fait un tour complet autour de l'observateur terrestre en 24 heures, l'heure réelle est simplement issue de l'angle que fait le Soleil avec une direction fixe (le sud, si l'observation se fait dans l'hémisphère nord). Si l'on considère le Soleil comme étant une horloge, on peut se poser la question de sa précision chronométrique [Vardi 2009]. L'erreur chronométrique peut être observée directement en notant la position du Soleil dans le ciel à la même heure civile pour chaque jour de l'année, ou alors en notant l'heure locale quotidienne où le Soleil est orienté plein sud par exemple. La Figure 1-5 donne l'erreur chronométrique du Soleil pour tous les jours de l'année 2010, commençant par le 1er janvier.

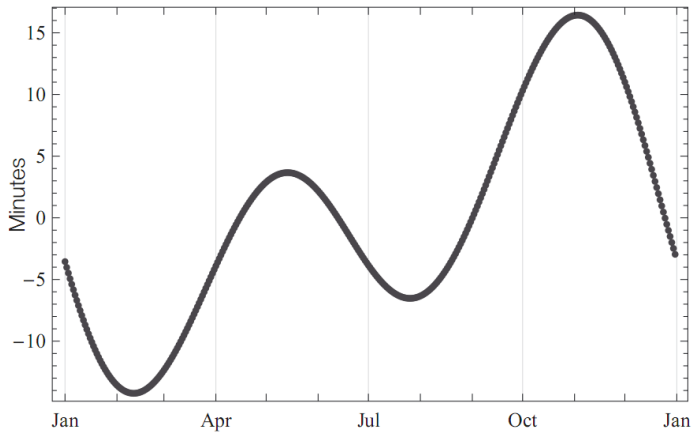


Figure 1-5 : erreur chronométrique du soleil

On s'aperçoit que l'erreur est assez conséquente (-14 minutes en février, et +16 minutes en novembre), par contre l'équation du temps ne varie pratiquement pas d'une année sur l'autre [Vardi 2009]. Habituellement, l'équation du temps est définie par :

$$ET = TSM\text{-}TSV \qquad (1\text{-}6)$$

Le Temps Solaire Moyen (*TSM*) est défini par un Soleil fictif qui fait une révolution autour de l'observateur terrestre en exactement 24 heures. Ce dernier est différent pour chaque longitude. Il est donc primordial de connaître la longitude pour chaque calcul lié à l'équation du temps. Si l'on utilise le temps *UTC* (ou *GMT*), c'est-à-dire le temps solaire

moyen de la ville de Greenwich, la formule pour le *TSM* d'un lieu est très simple comme on peut le voir sur l'Equation 1-7. En fait la longitude (λ) est exprimée en quantité horaire en utilisant la correspondance 1°⇔ 4 minutes correspondant au fait que 360° équivaut à 24 heures.

$$TSM=UTC+4.\lambda \qquad (1\text{-}7)$$

Les horloges des détecteurs alimentant les bases de données utilisées généralement (et en particulier durant cette thèse) n'affichent pas le temps *UTC*, mais plutôt le temps civil local (*TL*). Ces deux grandeurs ne différent que par la prise en compte du fuseau horaire (f_u correspondant à +1 heure en France) et du changement entre heure d'été et d'hiver (T_{E_H} correspondant à +1 heure en été) au moyen de la formule de l'Equation 1-8.

$$UTC = TL - f_u - T_{E\_H} \qquad (1\text{-}8)$$

En combinant toutes ces équations, on obtient la conversion entre *TL* et *TSV* comme le montre l'Equation 1-9[6].

$$TL = TSV + ET - 4.\lambda + f_u + T_{E\_H} \qquad (1\text{-}9)$$

Tout ce que nous venons de détailler permet de dégager des éléments essentiels concernant le rayonnement global. En effet, il est désormais possible, sous condition d'introduire la constante solaire *Esc*, de quantifier le rayonnement à l'entrée de l'atmosphère habituellement nommé « extra-terrestre » ou « hors atmosphère » que l'on notera H_0 dans la suite.

d. Le rayonnement extra-terrestre et son entrée dans l'atmosphère

Esc exprime la quantité d'énergie solaire que recevrait une surface de 1 m^2 située à une distance de 1 ua et exposée perpendiculairement aux rayons du Soleil, en l'absence d'atmosphère. Pour la Terre, c'est donc la densité de flux énergétique au sommet de l'atmosphère. La valeur admise d'*Esc* a varié au cours des années avec l'accroissement de la précision des instruments de mesure. Aujourd'hui, la valeur 1367 $W.m^{-2}$ est adoptée avec une

[6] Notons, qu'en simplifiant cette équation (avec $ET \rightarrow 0$ et $4.\lambda \rightarrow 0$), il est possible de retomber sur la formule triviale du *TL* donné par $TL = TSV + \begin{cases} 1 \text{ heure en hiver} \\ 2 \text{ heures en été} \end{cases}$.

incertitude de 0,15 % due à l'activité solaire [Oumbe 2009]. Le rayonnement hors atmosphère peut se déduire assez simplement des précédentes équations. Il ne dépend que de la constante solaire *Esc*, de la hauteur solaire *h* et de la distance Terre-Soleil *r* (Equation 1-10).

$$H_0 = \left(\frac{r_0}{r}\right)^2 . Esc. sin(h) \qquad (1\text{-}10)$$

Ce rayonnement hors atmosphère possède une seule composante du fait de la faible diffusion des photons durant leur parcours entre la surface du Soleil et le début de l'atmosphère. Cette règle n'est plus respectée dès lors que la limite de l'atmosphère est franchie. Une fois dans l'atmosphère, le rayonnement solaire se disperse et se divise alors en deux composantes bien distinctes : la composante directe et la composante diffuse. La première concerne les photons n'ayant pas interagis avec les particules de l'atmosphère et la seconde, ceux ayant interagis au moins une fois. Ce phénomène dispersif vient du fait que l'atmosphère peut être considérée comme une couche concentrique contenant un ensemble hétérogène de particules en suspension [Molineaux et al. 1998 ; Mondol et al. 2008 ; Notton et al. 2006 a]. A l'exception de quelques constituants provenant des activités humaines ou naturelles locales, la composition globale de l'atmosphère est presque la même en tout point du globe (~80 % de diazote et ~20 % de dioxygène). L'intensité de l'absorption du rayonnement par les gaz ne dépend pas seulement de la quantité des molécules dans l'atmosphère, mais aussi de leur absorption spécifique liée à leurs natures intrinsèques. Les profils verticaux de l'atmosphère montrent que 99 % de sa masse est contenue dans les 70 premiers kilomètres au-dessus du sol. Les principaux constituants de l'atmosphère sont d'origine nuageuse. Les nuages sont des amas de vapeur d'eau en suspension dans l'atmosphère qui se condensent autour de particules microscopiques (1 μm à 100 μm de diamètre) pouvant être de différentes natures (cristaux de sable ou de sel, suies volcaniques, particules de poussières, etc.) et de provenances diverses. Les nuages recouvrent en permanence près de 70 % de la surface terrestre et jouent un rôle majeur dans le bilan radiatif terrestre du système Terre-atmosphère. Les aérosols y sont aussi très présents : ce sont de fines particules liquides ou solides en suspension, de dimensions comprises entre quelques fractions de nanomètres et une centaine de nanomètres. Ils sont d'origines naturelles (cendres volcaniques, poussières désertiques, embruns marins) et humaines (fumées industrielles, particules issues de la combustion des matières fossiles, de la déforestation ou de feux agricoles, condensation de composés gazeux), ce qui induit une grande diversité de leurs

propriétés dans le temps, l'espace et en fonction des conditions météorologiques. L'essentiel des interactions entre l'atmosphère et le rayonnement solaire est principalement influencé par les éléments que l'on vient de présenter, à savoir : les nuages, les aérosols et les gaz atmosphériques (Tableau 1-1). Notons que concernant les ultraviolets, la principale source d'interaction, et donc d'absorption, est l'ozone (O_3). En moyenne, environ un tiers du rayonnement est réfléchi vers l'espace par les nuages et le sol ou rétrodiffusé par les constituants de l'atmosphère (26% + 4% = 30% d'après la Figure 1-1 située dans la section 2-1 de ce chapitre). Approximativement la moitié du rayonnement est absorbée par la surface terrestre.

Constituant	***Diffusion***	***Absorption***
Ozone	~0	Très forte pour $\lambda_0<0,3\mu m$
Gaz (autre que l'ozone)	Forte et croit en λ_0^{-4}	Faible
Vapeur d'eau	~0	Forte pour $\lambda_0>0,65\mu m$
Aérosols	Croit en λ_0^{-a} avec $0<a<4$	Faible
Nuages	Forte et dépend de λ_0	Faible

Tableau 1-1: résumé de la contribution des différents constituants de l'atmosphère à l'atténuation du rayonnement de longueur d'onde λ_0 avec « a » un paramètre d'ajustement dépendant de la nature des aérosols présents [Santos et al. 2003]

Il existe de nombreux modèles capables de modéliser le parcours des photons issus du Soleil au travers de l'atmosphère. Dans la section suivante, nous allons détailler les modèles qui ont été utilisés durant cette étude.

3. Les modèles de rayonnement utilisés

Dans la section précédente, nous avons posé un certain nombre de définitions nous permettant de formuler le concept de rayonnement solaire hors atmosphère [Kaplanis 2006]. Cette grandeur est facilement quantifiable car l'on considère qu'elle ne dépend que de la distance séparant le point de mesure du Soleil. Nous allons voir dans la présente section, que la traversée de l'atmosphère va tendre à compliquer la compréhension du phénomène, de par les multiples interactions susceptibles de se produire. Nous commencerons par exposer les fondements du modèle ciel clair (e.g. sans couverture nuageuse) appelé *SOLIS*, puis du

modèle d'inclinaison du rayonnement global (*CLIMED2* et Klutcher) choisis durant cette étude. Comme nous allons le voir, ces choix ont été déterminés par l'analyse et l'étude de documents bibliographiques.

3.1. Le modèle ciel clair

Avant de vouloir modéliser le ciel « normal » (molécule en suspension avec nébulosité), il paraît utile de réussir à modéliser le ciel en négligeant la couverture nuageuse. Ce concept se nomme la modélisation « ciel clair » [Aguiar et al. 1988]. Une bonne estimation du rayonnement par ciel clair permet de pouvoir accéder à des estimations correctes du rayonnement solaire en l'absence de nuage (notamment en été). Le concept de ciel clair tient compte de l'effet des aérosols et des gaz atmosphériques, mais s'affranchit de la contrainte nuageuse. Si l'on se réfère à ce qui a été énoncé dans le paragraphe précédent, la modélisation par ciel clair permet de quantifier la composante déterministe du rayonnement solaire. Le développement de ces modèles a commencé vers la fin des années 1970 avec le modèle empirique de Bird et Hulstrom [1981a et 1981b] qui s'attache à exprimer la transmittance du ciel clair pour différents processus d'atténuation dans l'atmosphère. Il s'agissait d'une modélisation simplifiée du rayonnement solaire direct et diffus sur les surfaces horizontales. Les années qui suivirent, de nombreux modèles vinrent compléter et améliorer cette première étude. On peut, par exemple, citer le modèle de Kasten [Kasten 1996], celui de Molineaux [Molineaux et al. 1998], ou encore le modèle *SOLIS* de Mueller [Mueller et al. 2004]. Durant cette étude c'est principalement ce dernier modèle qui a été retenu pour décrire la composante déterministe du rayonnement. En effet, ce modèle est issu du projet européen Héliosat-3 de l'Université d'Oldenburg et donne d'excellents résultats lorsqu'il est comparé à des mesures réalisées au sol en Europe [Ineichen et al. 1990 ; Ineichen 2006]. Le modèle simplifié *SOLIS* est une approximation des équations de transfert radiatif (Radiative Transfer Model ; *RTM*) basées sur des relations de type Lambert-Beer, (le lecteur intéressé peut se référer à [Ineichen 2008]). Ces relations permettent en outre de quantifier H_{bn} (Equation 1-11) qui est le rayonnement direct monochromatique (une seule longueur d'onde considérée) sur une surface normale au niveau du sol (avec *M* le facteur optique de la masse de l'air et τ la profondeur optique).

$$H_{bn} = H_0 . exp(-M.\tau) \qquad (1\text{-}11)$$

Dans le cas polychromatique (cas réel d'un mélange de photons de longueurs d'onde différentes) cette équation n'est plus valide et doit être remplacée par la relation proposée dans l'Equation 1-12 où *b* est une constante d'ajustement [Ineichen 2008] et *h* la hauteur solaire.

$$H_{bn} = H_0 . exp(-\frac{\tau}{sin^b(h)}) \qquad (1\text{-}12)$$

A cause des différentes diffusions dans l'atmosphère, il serait logique de penser que ce type de relation n'est pas applicable concernant le rayonnement global horizontal. Cependant, Mueller [Mueller et al. 2004] a montré que ce modèle d'atténuation peut aussi être utilisé dans le cas diffus et même global horizontal (Equation 1-13) à condition de modifier quelque peu l'expression du rayonnement hors atmosphère (H_0').

$$\begin{cases} H_{gh} = H_0' . exp\left(-\frac{\tau_g}{sin^g(h)}\right) . \sin(h) \\ H_{dh} = H_0' . exp(-\frac{\tau_d}{sin^d(h)}) \end{cases} \qquad (1\text{-}13)$$

Les différents paramètres nécessaires à l'emploi de l'Equation 1-13 et relatifs à l'expression du rayonnement extra-terrestre sont donnés par l'Equation 1-14 où *w* est la colonne de vapeur d'eau, *p* et p_o les pressions du lieu considéré et au niveau de la mer, et *aod* la profondeur optique concernant les aérosols et les radiations de 700 nm de longueur d'onde.

$$H_0' = H_0 . (0{,}12 . w^{0{,}56} . aod^2 + 0{,}97 . w^{0{,}032} . aod + 1{,}08 . w^{0{,}0051} . ln(\frac{p}{p_0}) \qquad (1\text{-}14)$$

Des relations similaires sont disponibles pour le calcul des différents paramètres d'optimisation tels que : *b*, *g*, *d*, τ_b, etc. [Mueller et al. 2004 ; Ineichen 2008].

3.2. Le rayonnement global sur une surface inclinée

Ce que nous avons vu dans le paragraphe précédent concerne le rayonnement global arrivant sur une surface plane horizontale posée sur le sol. Souvent les systèmes réels faisant appel à l'énergie solaire ne sont pas posés sur le sol mais sont inclinés par rapport à ce dernier [Klucher 1979]. Parallèlement à cette inclinaison, il peut y avoir aussi une orientation azimutale (écart entre le sud et un vecteur normal au plan considéré), une valeur positive signifiant une orientation vers l'ouest [Notton et al. 2006 a; Randimbivololona 2010]. Pour

modéliser l'inclinaison β du plan de détection par rapport à l'horizontale et l'angle azimutal *a*, il suffit de considérer le rayonnement global H_{gh} comme la somme d'une composante directe H_{bh} et d'une composante diffuse H_{dh}. On peut alors écrire [Noorian et al. 2008 ; Badescu 2008 ; De Miguel et al. 2001] :

$$H_{g\beta} = H_{bh}.R_b + R_d.H_{dh} + R \qquad (1\text{-}15)$$

Les trois facteurs *R* sont des facteurs de prise en compte de l'inclinaison, R_d concerne le rayonnement diffusé, *R* le diffusé issu du sol et R_b, le rayonnement direct (voir Figure 1-6).

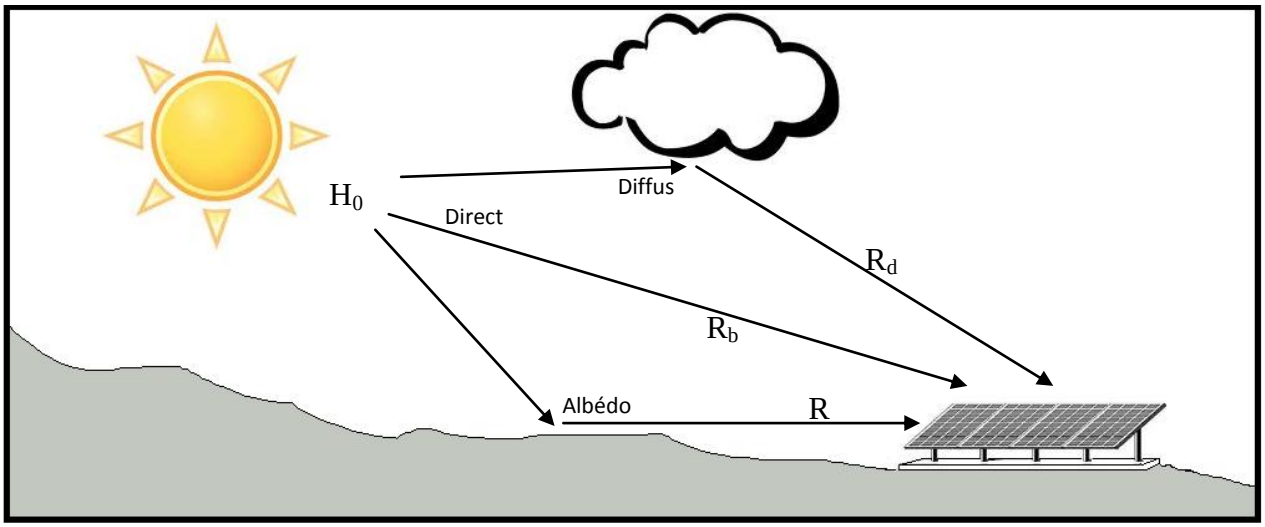


Figure 1-6 : rayonnement global sur une surface inclinée composé du rayonnement direct (facteur R_b), du rayonnement diffusé (facteur R_d) et du diffusé du sol (facteur *R*)

R_b est un facteur purement géométrique pouvant s'écrire à l'aide du facteur θ représentant l'angle d'incidence des rayons solaires sur la surface inclinée [Iqbal 1983 ; Badescu 2008].

$$R_b = \frac{cos(\theta)}{sin(h)} \qquad (1\text{-}16)$$

$$= \frac{(sin\,\varphi\,cos\,\beta - cos\,\varphi\,sin\,\beta\,cos\,\gamma)\,sin\,\delta + (cos\,\varphi\,cos\,\beta + sin\,\varphi\,sin\,\beta\,cos\,\gamma)\,cos\,\delta\,cos\,\omega_h + cos\,\delta\,sin\beta\,sin\gamma\,sin\,\omega_h}{sin\,\varphi\,sin\,\delta + cos\,\varphi\,cos\,\delta\,cos\,\omega_h}$$

Les deux autres facteurs intervenant lors de l'inclinaison du détecteur et relatifs au rayonnement diffus (du sol et de l'atmosphère), sont estimables grâce à de nombreux modèles, nous en présenterons quelques-uns ci-dessous.

Pour déterminer *R*, il est possible d'utiliser la relation faisant intervenir l'albédo (rapport de l'énergie solaire réfléchie par une surface à l'énergie solaire incidente ; ρ) [Notton et al. 2006 a et b] :

$$R = \frac{1}{2}.\rho.H_{gh}.(1 - cos(\beta)) \qquad (1\text{-}17)$$

Pour déterminer R_d, de nombreuses études ont été menées et ont permis de décrire une multitude de modèles. On peut citer le modèle isotropique de Liu et Jordan [1961], ou encore les modèles anisotropiques tels le modèle de Hay [Hay et Davies 1978], ou bien encore le modèle d'Iqbal [Iqbal 1983]. Une étude menée il y a quelques années au sein de notre laboratoire a permis de tester ces différents modèles sur les données solaires d'Ajaccio [Notton et al. 2004]. Il en résulte que le modèle offrant le plus de garantie est celui de Klucher [1979]. Son formalisme est relativement simple et fait intervenir un facteur *F*, défini par :

$$F = 1 - (\frac{H_{dh}}{H_{gh}})^2 \qquad (1\text{-}18)$$

On peut alors écrire que :

$$R_d = \frac{1}{2}.(1 + cos\left(\frac{\beta}{2}\right)).(1 + F.sin^3\left(\frac{\beta}{2}\right)).(1 + F.cos^2(\theta).sin^3(a)) \qquad (1\text{-}19)$$

Une des conditions nécessaires pour pouvoir appliquer les formules de R_b et de R_d est de connaître au préalable la proportion de rayonnement diffusé et de rayonnement direct [Rigollier et al. 2000]. Si cette proportion n'est pas connue il faut alors utiliser un modèle qui permet de la calculer. Là encore, de nombreux modèles sont dédiés à ce type d'étude (Cf. à Iqbal, Orgill, Erbs, etc. référencés dans [Notton et al. 2004]). Cependant, nous ne détaillerons que celui qui est utilisé régulièrement au sein de notre laboratoire, à savoir, le modèle *CLIMED2* qui est réputé comme étant le meilleur pour la zone méditerranéenne [Notton et al. 2004 ; Noorian et al. 2008]. Ce modèle propose de déduire la fraction de diffusé *f* du rayonnement global horaire à partir de l'indice de clarté horaire k_t [Kasten 1996 ; Notton et al. 2004 ; Reindl et al. 1990].

$$k_t = \frac{H_{gh}}{H_0} \qquad et \qquad f = \frac{H_{dh}}{H_{gh}} \tag{1-20}$$

Ainsi on peut appliquer :

$$\left.\begin{array}{l} f = 0{,}995 - 0{,}081k_t \quad pour\ \ k_t \leq 0{,}21 \\ f = 0{,}724 + 2{,}738k_t - 8{,}32{k_t}^2 + 4{,}967{k_t}^3 \quad pour\ \ 0{,}21 < k_t\ < 0{,}76 \\ f = 0{,}18 \quad pour\ \ k_t\ > 0{,}76 \end{array}\right\} \tag{1-21}$$

Le mode d'inclinaison que nous venons de détailler (*CLIMED2* et Klucher) est valable et a été testé sur Ajaccio [Notton et al. 2004]. Cependant, il est possible qu'il ne soit pas adapté à tous les sites. On peut alors s'affranchir de la localisation en calculant le rayonnement incliné à partir du rayonnement horizontal par l'Equation 1-22.

$$H_{gh}(\beta) = H_{gh}(\beta = 0)\left[\frac{H_{gh}(\beta)}{H_{gh}(\beta=0)}\right]^{ciel\ clair} \tag{1-22}$$

Il suffit donc d'utiliser un logiciel de type PVsyst™ [7] [Mermoud 2006] pour obtenir la partie entre crochet et ainsi incliner des données de rayonnement horizontal.

En annexe A, quelques exemples simples de manipulations relatives à la validation des modèles que l'on vient de présenter (*SOLIS*, *CLIMED2* et Klucher) sont disponibles.

4. Du rayonnement à la production photovoltaïque

Le rayonnement global est induit par un apport d'énergie à la surface terrestre. Cependant, outre l'aspect calorifique que ce dernier représente spontanément, il est possible à partir de cellules spécialement créées, de convertir cette énergie surfacique en énergie électrique. Ce processus passe par l'utilisation d'un module photovoltaïque intégré dans un système photovoltaïque complet. Il est associé à un système de stockage (batterie, H_2, volant d'inertie, etc.) afin de stocker l'énergie produite. Dans cette section, nous allons décrire de façon détaillée les différentes étapes qui permettent la production d'électricité à partir du rayonnement global.

[7] Logiciel commercial permettant de simuler du rayonnement global par ciel clair, http://www.pvsyst.com.

4.1. L'effet photoélectrique

L'effet photoélectrique a été initialement découvert par le physicien allemand Heinrich Rudolf Hertz (1886). Il réalisa expérimentalement qu'un matériau métallique exposé à la lumière pouvait émettre des électrons (Figure 1-7). Cette découverte fut appelée « effet photoélectrique ». Malheureusement, Hertz ne fut pas en mesure d'expliquer théoriquement le phénomène, car certaines caractéristiques de cet effet n'étaient pas formalisable avec la théorie classique de l'électromagnétisme de l'époque.

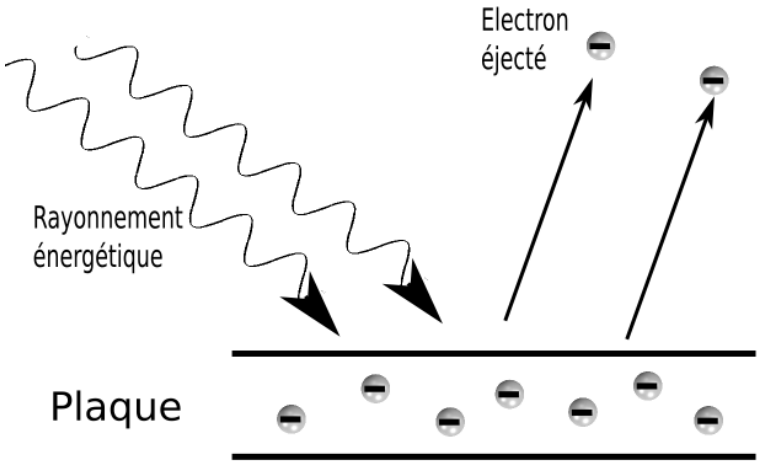


Figure 1-7 : principe de l'effet photoélectrique

Il faut attendre 1905 pour que le phénomène soit correctement expliqué avec la théorie quantique. Après « l'exploit théorique » réalisé par Max Planck en 1900 sur son interprétation du spectre du corps noir, Albert Einstein généralisa en 1905 le concept de perte d'énergie électromagnétique par quanta. Cette quantification de l'énergie permit de comprendre et d'appliquer le concept de photon comme particule transportant une quantité finie d'énergie. En appliquant l'hypothèse de l'existence du photon, Albert Einstein fut en mesure d'expliquer la nature quantique de l'effet photoélectrique et il fut récompensé du prix Nobel de physique en 1921 pour ses travaux. En résumé :

- lorsqu'une structure est exposée à la lumière, elle subit un bombardement de photons. Ceux-ci entrent en collision avec les électrons présents et peuvent être absorbés. S'il

y a absorption, l'énergie cinétique de l'électron augmente (apport égal à l'énergie du photon incident) ;

- pour que l'électron puisse être éjecté, il doit avoir suffisamment d'énergie pour quitter la structure. Cette énergie seuil porte le nom de « travail d'extraction » ;

- le photon doit, dès la collision, permettre à l'électron de changer d'état (état lié/état libre). Pour les électrons proches de la surface du matériau, il n'y a pas d'état de transition entre l'électron lié et l'électron libéré.

4.2. La conversion photovoltaïque

La conversion photovoltaïque peut être simplement définie comme la transformation de l'énergie des photons en énergie électrique grâce au processus d'absorption de la lumière par la matière [Petibon 2009]. Lorsqu'un photon interagit avec le matériau, il cède une partie de son énergie (ou toute) suite à la collision. Cette énergie permet d'arracher un électron à la matière (voir effet photoélectrique), créant un déséquilibre électrique au sein du milieu. Généralement, toute l'énergie des photons, n'arrivant pas à se transformer en électricité, est absorbée par le matériau sous forme thermique. Le matériau constituant les capteurs *PV* voit alors sa température interne augmenter proportionnellement à l'énergie solaire reçue. L'effet thermique est donc majoritaire sur la plupart des capteurs. Même si le phénomène électrique est secondaire devant le phénomène thermique, récupérer tout, ou une partie de l'énergie électrique est le premier objectif des capteurs photovoltaïques. Cette récupération est possible par l'utilisation de cellules solaires confectionnées en associant un matériau semi-conducteur [Mathieu et Fanet 2009] dopé N à un autre semi-conducteur dopé P comme le montre la Figure 1-8 :

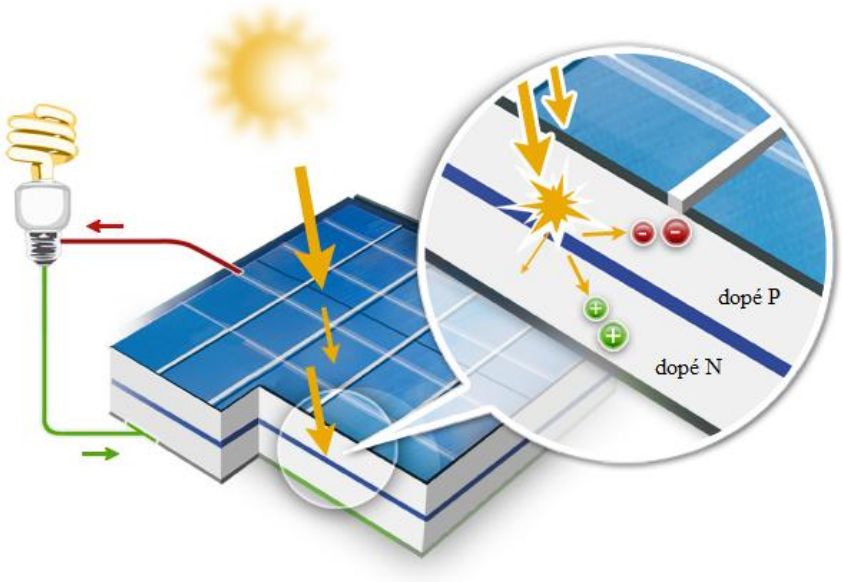


Figure 1-8 : principe de la cellule *PV*

L'énergie produite par l'absorption d'un photon dans un matériau se traduit du point de vue électrique par la création d'une paire électron-trou. Cette réaction entraîne une différence de répartition des charges générant une différence de potentiel électrique, c'est l'effet photovoltaïque. Le fait d'avoir associé deux types de matériaux pour créer une jonction permet de pouvoir récupérer les charges avant qu'elles ne se soient recombinées dans le matériau redevenu alors neutre. La présence de la jonction PN permet ainsi de maintenir une circulation de courant jusqu'à ses bornes. La plupart des cellules photovoltaïques utilise des semi-conducteurs pour récolter les paires électron-trou créées par la collision des photons dans le matériau. Chaque matériau possède son propre gap énergétique (bande d'énergie interdite entre la bande de valence et la bande de conduction [Mathieu et Fanet 2009]). Tout photon possédant une énergie inférieure à ce gap et arrivant à la surface du matériau n'aura pas assez d'énergie pour arracher un électron au matériau même s'il entre en collision avec l'un d'entre eux. Le courant produit par un capteur *PV* est donc beaucoup plus faible que la quantité de photons arrivant sur le matériau car plusieurs conditions doivent être réunies pour que l'énergie d'un photon puisse réellement se traduire en courant. Si le gap du matériau est grand, peu de photons auront assez d'énergie pour créer du courant, mais aux bornes de la cellule, la tension en circuit ouvert sera grande et facilitera d'autant plus l'exploitation de l'énergie électrique. A l'inverse, un matériau avec un faible gap absorbe plus de photons mais présente une tension plus faible à ses bornes. Le gap du silicium qui est aujourd'hui le

matériau le plus utilisé pour constituer des cellules dans les capteurs *PV* terrestres, est de 1,12 eV et le rendement maximum théorique pour une simple jonction *Si* (Silicium) est d'environ 29 %. Typiquement, la tension maximale d'une cellule (PN) est d'environ 0,5 à 0,8 V. Elle peut être directement mesurée à ses bornes sans charge (circuit ouvert). Cette tension est nommée tension de circuit ouvert (V_{co}). Lorsque les bornes d'une cellule sont court-circuitées, on peut mesurer le courant maximal produit par la cellule *PV* et on le nomme communément courant de court-circuit (I_{cc}). Ces valeurs peuvent changer fortement en fonction du matériau utilisé, de la température et de l'ensoleillement. La Figure 1-9 représente les caractéristiques typiques mesurables I_{cell}=f(V_{cell}) d'un module *PV* (mono-Si) soumis à un flux lumineux constant de 1000 W/m² et à une température de 25°C. L'ensemble de ces conditions expérimentales se nomme Standard Test Condition (*STC*[8]).

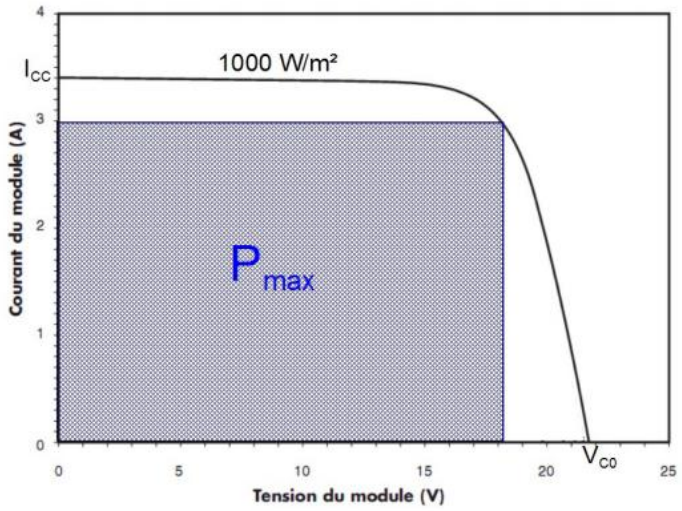


Figure 1-9 : courbe I-V et puissance maximale (ou puissance crête)

Cette courbe est la caractéristique électrique la plus importante d'un module photovoltaïque. La courbe I-V décrit la dépendance entre le courant photoélectrique (I_{mod}) généré par la lumière et la tension d'une cellule (V_{mod}). Afin de produire une puissance électrique, un module photovoltaïque doit donc créer une tension et un courant, le produit des deux étant la puissance (maximale pour I_{mod} = 3 A et V_{mod} = 18 V sur la Figure 1-9). De nombreux modèles électriques sont disponibles dans la littérature pour modéliser les courbes I-V des modules *PV*. On peut citer en particulier le modèle simple diode [Rosell et Ibanez

[8] Tests effectués en laboratoire, sous une irradiation de 1000 W/m2, une température de 25°, la lumière ayant le spectre attendu pour une pression atmosphérique de 1,5 atm.

2006] et le modèle double diode [Singh et al. 2008]. En pratique, pour déterminer la puissance électrique moyenne générée par un module (P_{DC}) durant un temps Δt, on aura tendance à favoriser la relation linéaire liant la puissance crête (P_C) théorique optimale du module (pour une irradiation de 1 kW/m² dans les conditions *STC*) et la puissance électrique produite comme le montre l'Equation 1-23 (unités requises entre les crochets).

$$P_{DC}[kW] = \frac{H_{gh}}{\Delta t}\left[\frac{kW}{m^2}\right].P_C[kWc]/1\left[\frac{kW}{m^2}\right] \qquad (1\text{-}23)$$

Notons que si le temps Δt est trop court (inférieur à un jour), cette formule risque de conduire à un taux d'erreur non négligeable, et il est alors préférable d'utiliser les modèles simples ou doubles diodes. En pratique, il est aussi possible d'introduire le concept de rendement η_{PV} (puissance électrique / puissance radiative) pour ainsi établir l'Equation 1-24 (*S* est la surface exposée des modules).

$$P_{DC}[kW] = \frac{H_{gh}}{\Delta t}\left[\frac{kW}{m^2}\right].S[m^2].\eta_{PV}[\%] \qquad (1\text{-}24)$$

Dans notre contexte, il est important de noter que la puissance *PV* augmente avec l'illumination solaire et qu'elle diminue quand la température augmente (perte de 0,4 % par degré supérieur à 25 °C). Il existe plusieurs techniques de modules solaires photovoltaïques :

- les modules solaires monocristallins : ils possèdent un très bon rendement ($\eta_{PV} \approx 15\%$), et sont essentiellement utilisés lorsque les espaces sont restreints. Leur coût, plus élevé que celui d'une autre installation de même puissance, entrave le développement de cette technique ;

- les modules solaires polycristallins : actuellement ils offrent le meilleur rapport qualité/prix et sont les plus utilisés. Ils ont un bon rendement ($\eta_{PV} \approx 12\%$) et une durée de vie supérieure à 35 ans ;

- les modules solaires amorphes : ils sont très intéressants car ils peuvent être souples et ont une meilleure production par faible lumière. Le silicium amorphe possède un rendement divisé par deux par rapport à celui du cristallin ($\eta_{PV} \approx 7\%$), ce qui nécessite plus de surface pour la même puissance installée, mais leur prix au m² étant plus faible, ils restent très compétitifs.

4.3. Les systèmes photovoltaïques

Dans des conditions d'ensoleillement standards (1000W/m² ; 25°C), la puissance maximale délivrée par une cellule en silicium de 150 cm² est d'environ 2,3 Wc sous une tension de 0,5 V. Une cellule photovoltaïque élémentaire constitue donc un générateur électrique de faible puissance, insuffisant en tant que tel pour la plupart des applications domestiques ou industrielles [Mondol et al. 2008]. Les générateurs photovoltaïques sont, de ce fait, réalisés par association, en série et/ou en parallèle, d'un grand nombre de cellules élémentaires. Outre cette juxtaposition de cellules et d'après ce que nous venons de voir dans le paragraphe précédent, le courant généré par les cellules photovoltaïques produit uniquement du courant continu à partir du rayonnement solaire [Wald 2007].

L'utilisation de ce courant continu diffère d'une installation à l'autre, selon le but de celle-ci. On distingue principalement deux types d'utilisation selon si l'installation photovoltaïque est connectée ou non à un réseau de distribution d'électricité.

Les installations non connectées peuvent directement consommer l'électricité produite. C'est le cas des sites non raccordés au réseau électrique (en montagne, sur des îles ou des voiliers, un satellite, etc.) qui peuvent en outre être équipés de batteries, d'accumulateurs pour disposer d'électricité au cours de périodes sans lumière (la nuit notamment).

Des installations photovoltaïques peuvent également être connectées à un réseau de distribution. Sur les grands réseaux de distribution (Amérique du Nord, Europe, Japon, etc.) des installations photovoltaïques produisent de l'électricité et l'injectent dans le réseau. Pour ce faire, ces installations transforment le courant continu en courant alternatif aux caractéristiques du réseau (fréquence de 50 Hz en Europe ou 60 Hz en Amérique du Nord par exemple). Elles n'ont pas besoin d'installation de stockage (batteries) car l'électricité est consommée à l'instant où elle est produite par les consommateurs les plus proches sur le réseau. Ce raccordement nécessite l'utilisation d'un onduleur, c'est-à-dire un dispositif qui a pour fonction de transformer un courant électrique continu en courant alternatif, identique à celui circulant sur le réseau. Il existe des onduleurs issus de différente technologie, cependant, leurs principes et objectifs sont les mêmes : faire passer l'énergie du *DC* (courant continu) à *l'AC* (courant alternatif). Cette transition entraîne une modification des performances globales du système. Il convient d'introduire le facteur *PR* (« performance ratio ») qui dépend du choix

des composants comme l'onduleur, la température, l'énergie perdue, etc. et peut modifier considérablement la production d'électricité au cours du temps. Le *PR* est calculé par le rapport entre le rendement mesuré et celui théorique. Son expression analytique est donnée par l'équation 1-25.

$$P_{AC}=PR.\eta_{PV}.\frac{H_{gh}}{\Delta t}.S \qquad (1\text{-}25)$$

L'onduleur est un appareil technologique sophistiqué qui fonctionne grâce à un microprocesseur et qui est soumis à des normes réglementaires. Il est protégé par un petit boîtier métallique que l'on place au plus près des modules photovoltaïques, afin d'éviter toute déperdition d'énergie. La durée de vie moyenne d'un onduleur est de 10 ans. En transformant la nature du courant, l'onduleur est le composant du module photovoltaïque qui permet d'utiliser l'électricité produite. Pour un site dit isolé (non raccordé au réseau), l'onduleur sert à faire passer l'électricité photovoltaïque du module de production aux appareils à alimenter. Dans le cas d'un site connecté au réseau, l'onduleur fait passer l'électricité photovoltaïque du site de production au réseau. En résumé, un onduleur n'ondule pas, il convertit. La figure 1-10 détaille le principe d'utilisation d'une source renouvelable d'énergie de type solaire raccordée au réseau électrique.

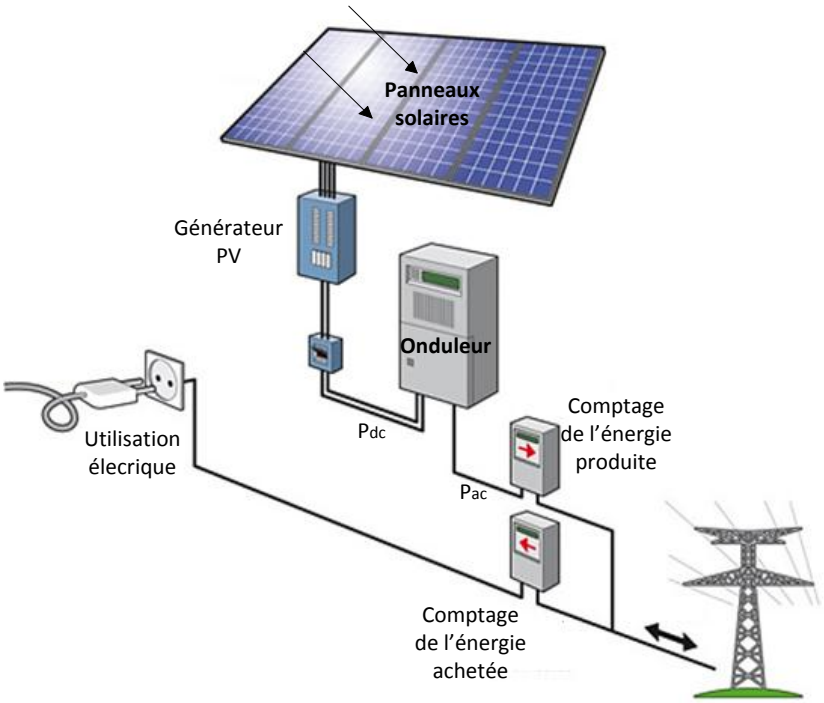


Figure 1-10 : principe du raccordement d'un module l'énergie *PV* au réseau électrique

Dans les différentes sections de ce paragraphe que nous venons de voir, nous avons illustré les différentes grandeurs liées à l'utilisation de modules *PV*. Les différentes manipulations qui seront traitées ultérieurement nécessitent de posséder un certain nombre de données expérimentales, afin de paramétrer et tester les différents modèles de prévision objet de cette thèse. Ainsi la section suivant détaille les bases de données, essentiellement météorologiques, utilisées dans cette étude.

5. Les données expérimentales

Pour valider notre approche, nous avons souhaité effectuer des prédictions de rayonnement global sur plusieurs sites géographiques méditerranéens. Nous avons axé notre étude sur le nord du bassin méditerranéen avec cinq sites d'étude en testant des données en rapport direct ou indirect avec le rayonnement global. Nous avons essayé de construire un outil de prédiction adaptable à différentes localités et à différents horizons. Les données que nous allons présenter peuvent se classer en trois catégories, les variables endogènes, exogènes et les prévisions. Se dit endogènes, les variables qui sont de mêmes natures que la prédiction (rayonnement global ou puissance *PV*) et exogènes celles de nature différentes (température, pression, etc.). La troisième catégorie fait partie des variables exogènes, mais représentera les paramètres qui ne sont pas mesurés mais estimés à partir de modèles physiques.

5.1. Données de mesures, rayonnement et variables météorologiques

Les cinq sites étudiés durant cette thèse sont Montpellier (43°4'N / 3°5'E, 2 m alt), Nice (43°4'N / 7°1'E, 2 m alt), Marseille (43°2'N / 5°2'E, 5 m alt), Bastia (42°3'N / 9°3'E , 10 m alt) et Ajaccio (41°5'N / 8°5'E, 4 m alt) (Figure 1-11). Ces villes présentent deux caractéristiques communes, à savoir une proximité immédiate de la mer Méditerranée et un relief montagneux proche. L'étude sur le rayonnement global que l'on a menée ne considère que les heures comprises entre 8h00 et 16h00 (*TSV*). En dehors de cette plage horaire, le rayonnement est trop faible pour une utilisation énergétique. De plus, les simulations réalisées en dehors de la tranche 8h00-16h00 peuvent être biaisées par les différents effets de terrain ou

de constructions masquant le rayonnement solaire. Ce phénomène appelé « masque solaire », peut pénaliser grandement les mesures faites en début ou en fin de journée.

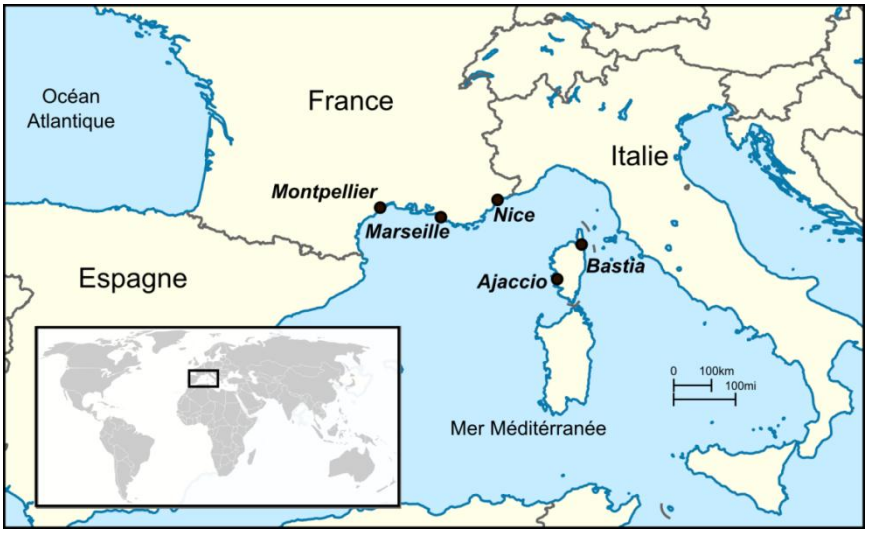


Figure 1-11 : localisation des 5 villes étudiées

Sur le Tableau 1-2 sont répertoriées les caractéristiques des données de rayonnement global horizontal dont nous disposons pour ces cinq villes. Ces données ont été gracieusement mises à notre disposition par Météo-France dans le cadre de la résolution 40 [OMM 1995] sur l'échange international des données et produits météorologiques. Météo-France est l'organisme français de météorologie (établissement public administratif) chargé de la prévision et de l'étude des phénomènes météorologiques.

Ville	Base	Rayonnement global		Variables exogènes	
Ajaccio	Horaire	01/01/1971 01/01/1998	31/12/1989 31/12/2007	01/01/1998	31/12/2007
Bastia	Horaire	01/01/1998	31/12/2007	01/01/1998	31/12/2007
Marseille	Horaire	01/01/1998	31/12/2007	01/01/1998	31/12/2007
Montpellier	Horaire	01/01/1998	31/12/2007	01/01/1998	31/12/2007
Nice	Horaire	01/01/1998	31/12/2007	01/01/1998	31/12/2007

Tableau 1-2: caractéristiques des données de rayonnement global horizontal et des variables exogènes disponibles (pression, nébulosité, température, vitesse du vent, direction du vent, vitesse de rafale, précipitation, humidité et insolation)

Les méthodologies d'acquisition et d'étalonnage de l'instrumentation sont celles qui étaient en vigueur au moment des relevés utilisés. De nos jours, les détecteurs sont des pyranomètres à thermocouple CM 11 de Kipp & Zonen possédant une sensibilité de 6 μV/W.m² et un temps de réponse de 24 s. Le taux de données manquantes et les modes de corrections sont disponibles dans l'annexe B.

Outre ces valeurs de rayonnement horizontal nous avons aussi pu tester notre approche sur du rayonnement incliné. Nous avions à disposition les données issues du mur photovoltaïque du laboratoire qui possède pour trois azimuts différents : 45° Sud-Est, 0° Sud et 45° Sud-Ouest. Les valeurs du rayonnement global sont inclinées à 80° par rapport à l'horizontal (Tableau 1-3). Les détecteurs utilisés ne sont pas des pyranomètres, mais des dispositifs à base de photodiode de type INGENIEURBURO SI-12TC (calibré par PTB Braunscheig permettant une mesure compris entre 0 et 1200W/m²).

Ville	Base	Période couverte	
Ajaccio/Vignola	5min SE	01/03/2009	31/08/2010
	5min S	01/03/2009	31/08/2010
	5min SO	01/03/2009	31/08/2010

Tableau 1-3: caractéristique des données de rayonnement global incliné à 80°

Certaines manipulations ont nécessité l'utilisation d'autres types de données (de type exogènes) comme la pression (*P*, Pa; moyenne et gradient journalier entre la pression moyenne des jours j et j-1, le tout mesuré par baromètre numérique durant 1 heure), la nébulosité totale (*N*, Octas), la température ambiante (*T*, °C; maximum, minimum, moyenne et de nuit à 3:00AM mesurée sur une demi-heure), la vitesse du vent (*Ws*, m/s; moyenne à 10m, mesurée pendant les 10 dernières minutes de chaque demi-heure), la direction du vent (*Wd*, deg à 10 mètres sur une demi-heure), la vitesse de rafale (*PKW*, m/s; vitesse maximale à 10 mètres, mesurée sur 30 minutes), les précipitations (*RP*, mm, 5 mesures cumulatives de 6 minutes durant une demi-heure), l'humidité (*RH*, %, mesure instantanée à la fin de chaque demi-heure) et l'insolation (*Su*, h, calculée avec la série de radiation globale et un seuil de 120 W/m²). L'utilisation de ces données conduit à une analyse dite multivariée, à opposer à l'analyse univariée ou seule les paramètres endogènes sont utilisés. Ces données exogènes, dont les périodes et caractéristiques de mesures sont résumées dans le Tableau 1-2, sont issues

des chaînes de mesure de Météo-France et sont disponibles sur les cinq sites précédemment introduits.

5.2. Données de prévisions Météo-France

Hormis les mesures Météo-France que l'on vient de citer, nous avons aussi utilisé pour certaines manipulations, des prévisions météorologiques. Pour mener à bien nos études, Météo-France a accepté de nous donner un libre accès à certaines de leur prévision. Le modèle retenu est basé sur la méthodologie *ALADIN* (Aire Limitée Adaptation dynamique Développement InterNational / consortium de pays européens) utilisée par le service national météorologique Météo-France (*CNRM*) [Pailleux et al. 2000 ; Pailleux et al. 2005 ; Bouttier 2010 ; Yessad 2010]. *ALADIN* fait partie du groupe des modèles baptisés « Numerical Weather Prediction » (*NWP*) [Radnoti et al. 1995]. Ce modèle est basé sur une méthodologie d'aire limitée bi-spectrale. Les données limites sont issues de sorties du modèle *ARPEGE* (Action de Recherche Petite Echelle Grande Echelle, [Radnoti et al. 1995]) qui est un modèle global lui aussi développé en partie par Météo-France (consortium européen). Le code *ALADIN* évolue en moyenne tous les six mois et suit les mises à jour apportées au modèle *ARPEGE*. Il existe quatre réseaux de prédiction associés à *ALADIN*, organisés suivant la production à 00, 06, 12 et 18 *UTC*. La prédiction est donnée à un horizon 54 heures. La résolution horizontale du modèle est approximativement 9,5 km, pour 60 niveaux verticaux. Les valeurs de sortie du modèle *ALADIN* sont calculées pour tous les points du maillage. Le modèle *ALADIN* génère plus de 20 sorties par pas horaire et il est à noter que le rayonnement fait depuis quelques années, partie des sorties générées. Cependant celui-ci étant très récent, nous avons choisi de ne pas l'utiliser par manque de recul sur son exploitation. Il n'était par ailleurs pas calculé durant les périodes d'étude allant de 1971 à 2010. Pour chaque localisation considérée, nous avons dû choisir entre le réseau le plus opportun (entre 00, 06, 12 et 18 *UTC*), la maille la plus proche du site considéré, et l'horizon de prédiction (1-54 heures). Parmi les 23 sorties possibles d'*ALADIN*, nous avons choisi les variables météorologiques les plus intéressantes pour la prédiction du rayonnement global. Nous avons sélectionné (chapitre 3, section 3-2) la pression (*P*, Pa), la nébulosité totale (*N*, Octas), les précipitations (*RP*, mm) et la température ambiante (*T*, °C). Concernant le maillage, nous avons utilisé le critère de proximité et ainsi choisi les points de la grille de calcul *ALADIN* les

plus proches des cinq stations retenues (écart d'environ cinq km). Le réseau sélectionné est celui de 06AM avec 12 heures d'horizon de prédiction par pas de une heure, ce qui nous permet de n'avoir besoin que d'une seule extraction *ALADIN* par jour pour récupérer les variables sur la plage horaire 8AM-04PM (en *TSV*).

Ville	Base	Période couverte	
Ajaccio	horaire	01/01/2002	31/12/2008
Bastia	horaire	01/01/2002	31/12/2008
Marseille	horaire	01/01/2002	31/12/2008
Montpellier	horaire	01/01/2002	31/12/2008
Nice	horaire	01/01/2002	31/12/2008

Tableau 1-4: séries temporelles de prévisions *ALADIN* (*P*, *N*, *RP* et *T*) disponibles pour le réseau horaire des 6AM (+12 heures)

5.3. Données de mesures électriques issues du laboratoire

Outre les données météorologiques que nous venons de lister, certaines expérimentations ont nécessité la validation de notre approche de prédiction sur le cas réel de systèmes photovoltaïques (Figure 1-12). Comme nous l'avons vu dans le cas du rayonnement global incliné, le laboratoire basé à Vignola possède trois façades inclinées de 80° (45° Sud-Est, 0° Sud et 45° Sud-Ouest) dotées de deux technologies de module *PV* : silicium monocristallin (*mono-Si*) et silicium amorphe *(a-Si*). Chaque champ *mono-Si* est connecté à un onduleur SUNNY BOY SMA (1,85 kW) et chaque champ *a-Si* à un autre onduleur SUNNY BOY SMA (0,7 kW).

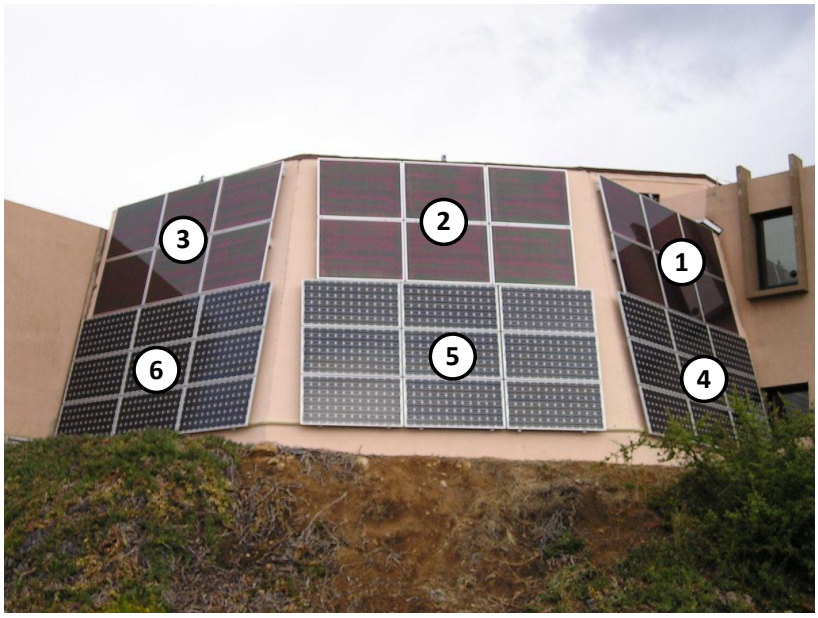


Figure 1-12 : mur expérimental du laboratoire mis en place en 2008

La surface totale est de 63m² et la puissance nominale est de 6,5 kWc (1,8kWc pour *a-Si* de type Mitsubishi heavy industries MA100T2 et 4,7 kWc pour *mono-Si* de type Suntech STP175S-24/Ac). Pour chaque technologie, nous avons une série temporelle de puissance *DC* (pré-onduleur), et une de puissance *AC* (post-onduleur).

Ville	N°	Base	Période couverte	
Ajaccio/Vignola	1	5min / SE / a-Si	15/01/2009	31/08/2010
	2	5min / S / a-Si	15/01/2009	31/08/2010
	3	5min / SO / a-Si	15/01/2009	31/08/2010
	4	5min / SE / mono-Si	15/01/2009	31/08/2010
	5	5min / S / mono-Si	15/01/2009	31/08/2010
	6	5min / SO / mono-Si	15/01/2009	31/08/2010

Tableau 1-5: caractéristiques des données de puissance électrique disponibles (puissances *AC* et *DC*)

Les différentes données que l'on vient de présenter dans cette section, ne sont pas exploitables directement. Il est fréquent d'avoir à nettoyer les séries brutes ou à recalibrer (*a posteriori*) ces valeurs. Les méthodologies de nettoyage et de calibration utilisées sont détaillées en annexe B et C de ce rapport.

6. Conclusion

Ce chapitre clôture les connaissances préalables qu'il a fallu acquérir pour pouvoir établir et utiliser des modèles prédictifs de rayonnement global. En effet, nous verrons plus loin dans ce manuscrit, que la prédiction renvoie fréquemment vers une analyse de la nature intrinsèque du rayonnement global [Marty et Philipona 2000 ; Mellit et al. 2008]. Nous avons vu à travers ce chapitre que le rayonnement global est composé de deux composantes élémentaires : l'une déterministe et l'autre stochastique. C'est cette caractéristique qui a permis de formuler des modèles empiriques de rayonnement de type « ciel clair ». Ces derniers utilisent un formalisme mathématique dédié, basés sur : la distance Terre Soleil, la déclinaison, les systèmes de coordonnées horizontales et horaires, l'équation du temps et le rayonnement hors atmosphère. Les modèles de rayonnement qui ont été détaillés dans ce chapitre, concernent l'estimation par ciel clair *SOLIS*, la fraction de rayonnement diffus issu du modèle *CLIMED2* et l'inclinaison de ce dernier par la méthode de Kutcher. Ce ne sont pas forcément les plus utilisés, ou ceux donnant en général les meilleurs résultats. Cependant, les données dont nous disposions et les caractéristiques des sites d'étude sont spécifiques et c'est ce qui a déterminé le choix de ces modèles. De plus, nous possédions déjà une connaissance et une expérience au sein du laboratoire de certains de ces modèles [Notton et al 2004]. D'autre part, nous avons aussi opté pour ceux qui nous paraissaient le plus adaptés d'après la littérature. Comme souvent en prédiction, il est difficile de définir un modèle universel qui s'appliquerait à tous les cas de figure. Il faut donc composer avec tous les outils disponibles, pour finalement choisir ceux qui seront les plus aptes à répondre aux problèmes posés. Dans le prochain chapitre, nous détaillerons les différents outils mathématiques utilisés pour la prédiction à horizon des séries temporelles de rayonnement global et de puissance *PV*.

7. Bibliographie

Aguiar RJ, Collares-Pereira M, Conde JP. Simple procedure for generating sequences of daily radiation values using a library of Markov transition matrices. Solar Energy n°40(3), p 269-279, 1988.

Badescu V. Modeling solar radiation at the earth's surface: recent advances. Springer; 2008.

Batlles FJ, Rubio MA, Tovar J, Olmo FJ et al. Empirical modeling of hourly direct irradiance by means of hourly global irradiance. Energy n°25(7), p 675-688, 2000.

Bird RE, Hulstrom RL. Evaluation, and Improvement of Direct Irradiance Models. J. Sol. Energy Eng n°103(3), p 182-192, 1981 a.

Bird RE, Hulstrom RL. A Simplified Clear Sky Model for Direct and Diffuse Insolation on Horizontal Surfaces. Solar Energy Research Institute, 1981 b.

Bouttier François, The Météo-France *NWP* system: description, recent changes and plans, CNRM, *NWP* report, 2010.

De Miguel A, Bilbao J, Aguiar R, Kambezidis H et al. Diffuse solar irradiation model evaluation in the North Mediterranean belt area, Solar Energy n°70, p 143-153, 2001.

Hay JE, Davies JA. Calculation of the Solar Radiation Incident on an Inclined Surface," in Proceedings. First Canadian solar radiation workshop, Toronto. pp. 59-72, 1978.

Ineichen P, Guisan O, Perez R. Ground-reflected radiation and albedo. Solar Energy n°44(4), p 207-214, 1990.

Ineichen P. Comparison of eight clear sky broadband models against 16 independent data banks. Solar Energy n°80(4), p 468-478, 2006.

Ineichen P. A broadband simplified version of the *SOLIS* clear sky model. Solar Energy.n°82(8), p 758-762, 2008.

Iqbal M. An introduction to solar radiation. Academic Press Inc. 1983.

Jin Z, Yezheng W, Gang Y. General formula for estimation of monthly average daily global solar radiation in China. Energy Conversion and Management n°46(2), p 257-268, 2005.

Kaplanis S. New methodologies to estimate the hourly global solar radiation; Comparisons with existing models. Renewable Energy n°31(6), p 781-790, 2006.

Kasten F. The linke turbidity factor based on improved values of the integral Rayleigh optical thickness. Solar Energy n°56(3), p 239-244, 1996.

Klucher T.M., Evaluation of models to predict insolation on tilted surfaces, Solar Energy n°23, p 111–114, 1979.

Lam J, Wan K, Lau C, Yang L. Climatic influences on solar modelling in China. Renewable Energy. N°33(7), p 1591-1604, 2008.

Liou K. An introduction to atmospheric radiation. Academic Press; 1980.

Liu B, Jordan R. Daily insolation on surfaces tilted towards the equator. ASHRAE Journal n°3, p 53-59, 1961.

Marty C, Philipona R. Clear-sky index to separate clear-sky from cloudy-sky situations in climate research. Geophysical Research Letters n°27(17), p 2649–2652, 2000.

Mathieu H, Fanet H. Physique des semiconducteurs et des composants électroniques: cours et exercices corrigés. Dunod; 2009.

Mellit A, Kalogirou S, Shaari S, Salhi H et al. Methodology for predicting sequences of mean monthly clearness index and daily solar radiation data in remote areas: Application for sizing a stand-alone PV system. Renewable Energy n°33(7), p 1570-1590, 2008.

Mermoud, A.Technico-economical Optimization of Photovoltaic Pumping Systems Pedagogic and Simulation Tool Implementation in the PVsyst Software". Research report of the Institut of the Environnemental Sciences, University of Geneva. 2006

Molineaux B, Ineichen P, O Neill N. Equivalence of Pyrheliometric and Monochromatic Aerosol Optical Depths at a Single Key Wavelength. Appl. Opt. n° 37(30), p 7008-7018, 1998.

Mondol J, Yohanis Y, Norton B. Solar radiation modelling for the simulation of photovoltaic systems. Renewable Energy, n°33(5), p 1109-1120, 2008.

Mueller RW, Dagestad KF, Ineichen P, Schroedter-Homscheidt et al. Rethinking satellite-based solar irradiance modelling: The *SOLIS* clear-sky module. Remote Sensing of Environment, n°91(2), p 160-174, 2004.

Noorian A, Moradi I, Kamali G. Evaluation of 12 models to estimate hourly diffuse irradiation on inclined surfaces. Renewable Energy, n° 6;33(6), p 1406-1412, 2008.

Notton G, Cristofari C, Muselli M, Poggi P. Calculation on an hourly basis of solar diffuse irradiations from global data for horizontal surfaces in Ajaccio. Energy Conversion and Management, n°45, p 2849-2866, 2004.

Notton G, Cristofari C, Poggi P. Performance evaluation of various hourly slope irradiation models using Mediterranean experimental data of Ajaccio. Energy Conversion and Management, n° 1;47(2), p 147-173, 2006 a.

Notton G, Poggi P, Cristofari C. Predicting hourly solar irradiations on inclined surfaces based on the horizontal measurements: Performances of the association of well-known mathematical models. Energy Conversion and Management, n°8;47(13-14), p 1816-1829, 2006 b

OMM, Congrès de l'Organisation Météorologique Mondiale (OMM) qui a adopté la résolution 40 sur l'échange international des données et produits météorologiques. Organisation Météorologique Mondiale, S/A/PMS, ANNEXE, Résolution 40 (Cg-XII). 1995.

Oumbe Ndeffotsing Bertrand, thèse de doctorat. Exploitation des nouvelles capacities d'observation de la Terre pour évaluer le rayonnement solaire incident au sol. Ecole nationale supérieure des mines de Paris. 2009

Pailleux J, Geleyn JF, Legrand E: La prévision numérique du temps avec ARPEGE et ALADIN : Bilan et perspectives. La Météorologie, 8e série, n°30, p 32-60, 2000.

Pailleux J, Geleyn JF, Tardy M. Les progrès récents de la prévision numérique à Météo-France. La Météorologie, n°48, p 2-5, 2005.

Perez R, Ineichen P, Seals R, Michalsky J et al. Modeling daylight availability and irradiance components from direct and global irradiance. Solar Energy, n°44(5), p 271-289, 1990.

Petibon Stéphane. Thése soutenue à l'université de Toulouse (laboratoire d'Analyse et d'Architecture des Sytèmes en 2009.

Radnoti G, Ajjaji R, Bubnova R, Caian M et al. The spectral limited area model *ARPEGE*/*ALADIN*. PWPR Report Series n°7, WMO-TD n° 699, p. 111-117, 1995.

Randimbivololona Prisca, Mastère spécialisé OSE. Prédiction du rayonnement global par pas horaire 24 heures à l'avance. Ecole des Mines de Paris, 2010.

Rehman S. Empirical model development and comparison with existing correlations. Applied Energy, n° 1;64(1-4), p 369-378, 1999.

Reindl D, Beckman WA, Duffie JA. Evaluation of hourly tilted surface radiation models. Solar Energy, n°45(1), p 9-17, 1990.

Rigollier C, Bauer O, Wald L. On the clear sky model of the ESRA -- European Solar Radiation Atlas -- with respect to the heliosat method. Solar Energy, n°68(1), p 33-48, 2000.

Rosell JI et Ibanez M. modelling power output in photovoltaic modules for outdoor operating conditions. Energy conversion and management Vol 47, n°15-16, p 2424-2430, 2006.

Santos JM, Pinazo JM, Cañada J. Methodology for generating daily clearness index index values Kt starting from the monthly average daily value. Determining the daily sequence using stochastic models. Renewable Energy n°28(10), p 1523-1544, 2003.

Singh P, Singh SN, Lal M, Husain M. Temperature Dependence of I–V Characteristics and Performance Parameters of Silicon Solar Cells, Solar Energy Materials and Solar Cells, vol. 92, n°12, p 1611 - 1616, 2008.

Ulgen K, Hepbasli A. Diffuse solar radiation estimation models for Turkey's big cities. Energy Conversion and Management n°50(1), p 149-156, 2009.

Vardi I. La marche du soleil, bulletin SSC n°62, Décembre 2009.

Wald L. Solar radiation energy (fundamentals and theory). In Solar Energy Conversion and Photoenergy Systems. EOLSS; 2007.

Yessad K, « basics about ARPEDE/IFS, *ALADIN* and AROME in the 37 of the cycles of *ARPEGE*/IFS, Météo-France/CNRM/GMAP/ALGO, research report, 2010.

Chapitre 2

Les séries temporelles et leur prédiction : application au rayonnement solaire et à la production d'énergie

SOMMAIRE

2

1. Introduction

Plusieurs méthodes permettent d'appréhender la prédiction du rayonnement global au sol. On peut les classer en deux grandes familles.

La première consiste à utiliser un ensemble de procédés dits experts du phénomène à modéliser [Badescu 2008 ; Mubiru 2008]. On emploie le terme de « modèle de connaissances » pour chacun d'eux (section 3-1 du chapitre 1, p 43). Si ces méthodes sont issues d'équations algébriques, on emploie le terme de modélisation statique [Mueller et al. 2004]. Si elles proviennent d'équations différentielles ou d'équations aux différences récurrentes le modèle est dit dynamique, respectivement à temps continu ou à temps discret [Bouttier 2010].

La seconde famille de méthode de prédiction ne considère que l'aspect statistique du phénomène, et utilise uniquement des outils d'analyse stochastiques non spécifiques au phénomène étudié. On emploie alors le terme de modèle « boîte noire » pour chacun d'eux. Ces modèles sont construits essentiellement sur la base de mesures du processus à modéliser [Mellit et al. 2009] ou de mesures de processus directement dépendants de celui-ci. Cette série de mesures se nomme « série temporelle » ou « série chronologique » ou encore « chronique » [Mellit et Kalogirou 2008]. Le modèle repose sur une formulation mathématique reliant les entrées (données sources) et les sorties (les prédictions). Si ces deux entités sont de mêmes natures (endogènes), on emploiera le terme d'analyse univariée ; si elles sont de natures différentes (exogènes), on distinguera l'approche en la qualifiant de multivariée (voir la section 5 du chapitre 1, p 55).

Nous verrons dans la suite de ce manuscrit qu'un couplage entre le procédé expert et l'analyse statistique est souvent nécessaire afin d'accroître la qualité de la prédiction et ainsi utiliser un modèle de type « boîte grise » (ou modélisation semi-physique) [Dreyfus et Idan 1998 ; Dreyfus 2004]. Ces modèles sont utilisables lorsque des connaissances exprimables sous forme d'équations sont disponibles, mais insuffisantes pour concevoir un modèle de connaissances satisfaisant. C'est le cas de la modélisation ciel clair du rayonnement qui ne tient pas compte de l'occurrence nuageuse. Ainsi, le modèle « boîte grise » considère à la fois les connaissances et les mesures. Une telle démarche peut concilier les avantages de l'intelligibilité d'un modèle de connaissance avec la souplesse d'un modèle « boîte noire » comportant des paramètres ajustables [Oussar 1998].

Dans ce chapitre, nous commencerons par expliquer ce que sont les séries temporelles, en insistant sur les définitions d'ordre général. Nous dégagerons ensuite les différentes méthodologies de prédiction (et plus particulièrement celles liées au rayonnement global) en détaillant les principaux modèles utilisables. Deux méthodes de prédiction (modèles autorégressifs et réseaux de neurones artificiels) seront plus particulièrement détaillées parce qu'elles constituent la base de la méthodologie retenue et exposée dans le chapitre suivant. Pour finir, nous présenterons les principaux paramètres d'erreur permettant de comparer ces différentes méthodologies de prédiction. Sur la base de ces différents estimateurs, une comparaison bibliographique des méthodes de prédiction relatives au rayonnement global ou à la puissance *PV* sera également proposée.

2. L'analyse des séries temporelles

Que ce soit dans le domaine de la météorologie ou celui des marchés financiers, il est aujourd'hui fréquent de parler de « prédiction ». Comme énoncé en début d'introduction générale, dans ce rapport nous emploierons le terme de « prédiction » pour le formalisme des séries temporelles et les estimations à horizons donnés, le terme de « prévision » sera dédié aux modèles analytiques comme ceux proposés par Météo-France. En fait, l'estimation de la valeur future d'une grandeur en météorologie (comme le rayonnement global) ou d'un produit financier peut, sous l'aspect de l'analyse des séries temporelles, se traiter de la même manière. Pour cela, il est usuel d'utiliser un modèle « boîte noire ». Généralement, afin de pouvoir estimer la valeur future d'une variable, il est indispensable de disposer d'informations

relatives à son évolution passée. Une série temporelle se définit intuitivement comme la suite des valeurs passées de la variable que l'on essaie de prédire [Bourbonnais 1998 ; Bourbonnais et Terraza 2008 ; Rumelhard et al. 1986 et Simon 2007]. Pour passer de l'évolution antérieure aux valeurs futures, il est utile d'analyser, puis d'optimiser la modélisation de cette évolution. Cette étape peut revêtir de multiples formes ; elle est baptisée de façon générique « apprentissage du processus ». Ces questions d'analyse de la série et de la construction d'un modèle sont souvent le cœur même du problème en prédiction. Une fois l'apprentissage terminé et la série temporelle estimée, le modèle doit idéalement être capable de représenter aussi fidèlement que possible la dynamique de l'évolution propre à chaque série. Lorsque le modèle représente correctement cette évolution passée, on peut raisonnablement considérer que la prédiction sera pertinente par rapport à la dynamique qui a été mesurée et qui se trouve dans la série. Par contre, un modèle qui ne correspond pas aux valeurs mesurées durant l'apprentissage, permettra tout de même de faire une prédiction, mais cette dernière n'a pas grand sens au regard de la série temporelle considérée. Comme nous allons le confirmer tout au long de ce chapitre, la construction d'un modèle relatif à une série et l'analyse de sa dynamique sont des étapes importantes à tout problème de prédiction.

2.1. Définition

Une série temporelle est une succession d'observations (ou de mesures) au cours du temps représentant un phénomène. Dans ce manuscrit, les observations seront de la puissance *PV*, du rayonnement global au sol ou d'autres variables météorologiques comme la pression, la température, ou la nébulosité. Par hypothèse, le pas de temps des observations est considéré constant, et la modalité d'acquisition des signaux doit être identique tout au long de la période d'observation. En théorie, ces deux hypothèses fortes sont nécessaires à une analyse de qualité. Toutefois, en pratique il est impossible de les respecter scrupuleusement, il convient simplement de s'assurer de la cohérence et de l'harmonie des mesures au cours du temps. Sans cette vérification, le modèle n'étant qu'une représentation mathématique d'un phénomène, il va tout de même essayer d'approximer le plus possible les différents signaux sans avertir ni laisser présager d'une éventuelle dérive des détecteurs, ou d'une modification de la méthodologie d'acquisition. Une démarche d'assurance de qualité des appareillages est donc essentielle. De plus, la mise en conformité des séries temporelles brutes (issues de

l'acquisition), est souvent nécessaire afin de prendre en compte les éventuelles modifications, mises à jour, pannes, maintenances, etc. Quand on utilise le formalisme des séries temporelles comme outil de prédiction, il faut bien avoir à l'esprit que l'on ne va pas prédire une grandeur météorologique absolue mais uniquement la mesure de celle-ci par un dispositif physique. Il convient donc de faire converger autant que faire se peut la grandeur absolue et la mesure.

Pour utiliser le formalisme des séries temporelles, il est nécessaire, au préalable, de poser certaines définitions. Ainsi, la valeur courante en *t* de la chronique *x* est notée x_t où *t*, le temps, est compris entre 1 et n, avec n le nombre total d'observations. On appelle *h* le nombre de points ou de valeurs à prédire de la série temporelle. La prédiction de la série temporelle de (n+1) à (n+h), connaissant l'historique de x_1 à x_n, porte le nom d'horizon de la prédiction (horizon 1,…, horizon *h*). Aussi pour un horizon 1 (cas le plus simple), le formalisme général de la prédiction sera représenté par l'Equation 2-1 où ϵ représente l'erreur entre la prédiction et la mesure, f_n le modèle à estimer et *t* le paramètre temporel qui prend les (n-p) valeurs suivantes : n, n-1,…, p+1, p. Où n est le nombre d'observations et p le nombre de paramètres du modèle (on suppose que n ≫ p).

$$x_{t+1} = f_n\left(x_t, x_{t-1}, \dots., x_{t-p+1}\right) + \epsilon(t+1) \qquad (2\text{-}1)$$

Il est toujours utile en première analyse de représenter l'évolution temporelle d'un phénomène (profil et allure de la chronique) à l'aide d'un graphique ayant en ordonnée la valeur du phénomène x_t et en abscisse le temps *t*. Ainsi sur la figure 2-1, on peut se rendre compte que le phénomène de rayonnement global (X_t) est un ensemble de signaux périodiques avec un bruit (lié à la couverture nuageuse) qui semble plus important durant les mois d'hiver que durant ceux d'été. Les données relatives aux nuits ont été retirées pour une plus grande lisibilité, seules les heures comprises entre 8h00 et16h00 sont considérées (neuf heures par jour). Dans ce cas précis, les deux périodes du signal sont T_1 = 9 heures et T_2 = 365 jours = 365x9 heures.

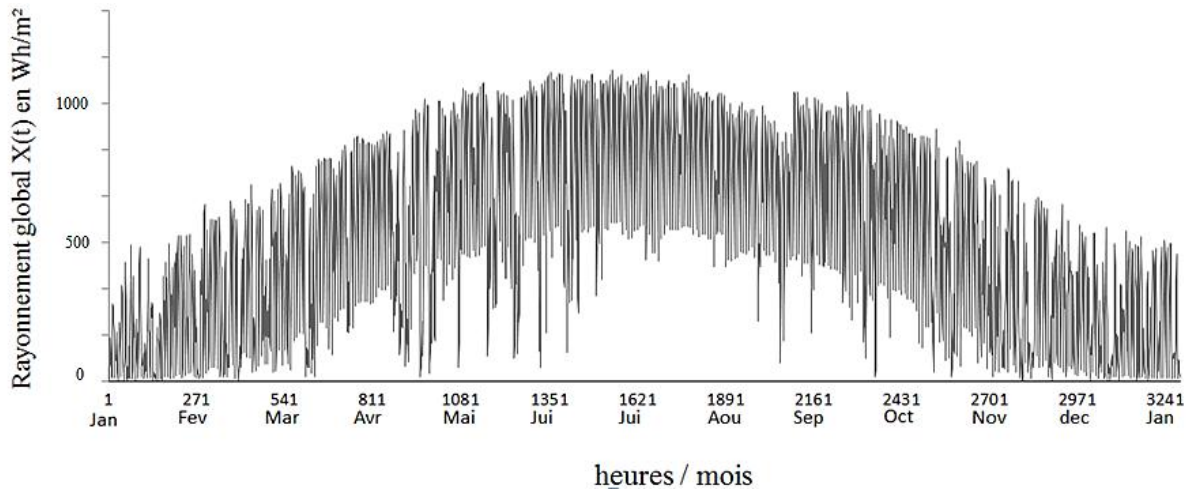


Figure 2-1 : représentation du rayonnement global horizontal horaire (X_t) pour le site d'Ajaccio entre janvier 1998 et janvier 1999

Pour aller plus loin et mieux comprendre les phénomènes entourant la chronique de rayonnement global, il faut utiliser certaines définitions liées à la périodicité, la tendance, la saisonnalité, le bruit résiduel, etc. Les sections suivantes vont permettre de mieux comprendre les fondements de l'analyse des séries temporelles.

2.2. Approche stochastique du signal

Les techniques traditionnelles d'analyse des chroniques procèdent souvent par décomposition et recomposition. L'approche de décomposition suppose que la structure de chaque chronique peut être scindée en éléments simples (modélisables), et donc plus facilement prévisibles, pour ensuite être reconstituée pour donner la prédiction de la chronique. Les premières études [Hamilton 1994 ; Bourbonnais et Terraza 2008] sur les chroniques ont amené à considérer de façon standard trois grandes composantes de séries temporelles :

- **la tendance**, notée E_t, qui décrit le mouvement sur le long terme (extra-annuel) ;

- **la composante saisonnière** Σ_t, qui est une composante cyclique relativement régulière de période intra-annuelle (notons qu'il peut y avoir plusieurs composantes saisonnières) ;

- **la composante résiduelle** notée R_t, qui est une composante par défaut, qui regroupe ce que les autres composantes n'ont pu intégrer.

Pour affiner le modèle, on choisit le schéma de décomposition qui convient le mieux à la nature du processus. Cela consiste à combiner les trois composantes élémentaires de série temporelle que l'on vient de définir, dans le schéma qui correspond le plus aux propriétés intrinsèques de la chronique. Habituellement, il est retenu trois grands groupes de décomposition applicables à la plupart des séries temporelles :

- **le schéma additif** qui nécessite l'orthogonalité des différentes composantes et qui s'écrit :

$$x_{t=} E_t + \Sigma_t + R_t \qquad (2\text{-}2)$$

- **le schéma multiplicatif simple**, utilisable quand il n'y a pas d'interaction entre le bruit résiduel et les deux autres composantes :

$$x_{t=} E_t . \Sigma_t + R_t \qquad (2\text{-}3)$$

- **le schéma multiplicatif complet** nécessaire dès qu'il y a interaction générale des trois composantes :

$$x_{t=} E_t . \Sigma_t . R_t \qquad (2\text{-}4)$$

Dans le cas du rayonnement global, noté X_t, on a une simplification liée à la constance apparente du cumul du rayonnement global annuel[9]. Cette simplification annule l'effet de la composante E_t. De plus, comme l'occurrence de nuages semble plus importante en hiver qu'en été, il serait logique de penser que le schéma multiplicatif complet de l'équation 2-4 matérialise le mieux le rayonnement global. Dans ce cas précis, X_t devient :

$$X_{t=} \Sigma_t . R_t \qquad (2\text{-}5)$$

9 En réalité il y a un phénomène périodique lié au cycle solaire (période de 11 ans). Cependant les écarts annuels sont très faibles, environ 0,2W/m² en fonction de l'année choisie [Fröhlich et Lean, 1998].

La Figure 2-1 suggère que l'allure ou la tendance du rayonnement global peut se représenter sous forme analytique de la façon décrite par l'Equation 2-6, où $\omega_1 (= \frac{2\pi}{T_1})$ et $\omega_2 (= \frac{2\pi}{T_2})$ sont les pulsations observables et *a(t)* et *b(t)* les amplitudes non constantes qui ne semblent pas avoir d'expression simple.

$$X_t = a(t) . cos(\omega_1 . t) + b(t) . cos(\omega_2 . t) \qquad (2\text{-}6)$$

En regroupant l'Equation 2-5 et l'Equation 2-6 on obtient l'Equation 2-7 qui est purement théorique et qui représente la décomposition du rayonnement global.

$$X_t = (a(t)\, cos(\omega_1 . t) + b(t)\, cos(\omega_2 . t)) . R_t \qquad (2\text{-}7)$$

Il existe différents types de saisonnalités, dans notre cas on peut parler de périodicité intrinsèque « rigide », ou déterministe car elle est bien marquée et répétitive. Sur la Figure 2-2, il est possible de constater cette périodicité intrinsèque au rayonnement global.

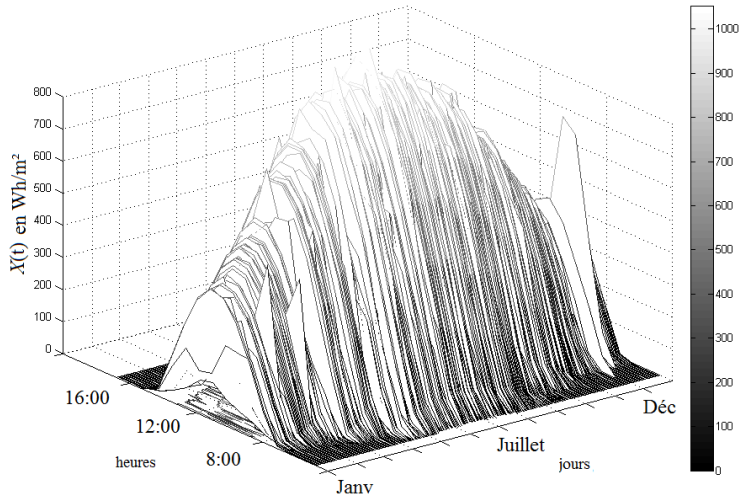


Figure 2-2 : représentation du rayonnement global *X(t)* sur une année à Ajaccio en 2000

Sur la Figure 2-3, l'utilisation de la transformée de Fourier permet de confirmer la rigidité de la double saisonnalité (annuelle et journalière), seules les heures comprises entre 8h00 et 16h00 sont considérées (9 heures /jour). Les deux pics sont relatifs aux périodes T_1 (9 heures) et T_2 (3285 heures = 9 x 365). Il est à noter que ce résultat peut être généralisé au rayonnement global incliné.

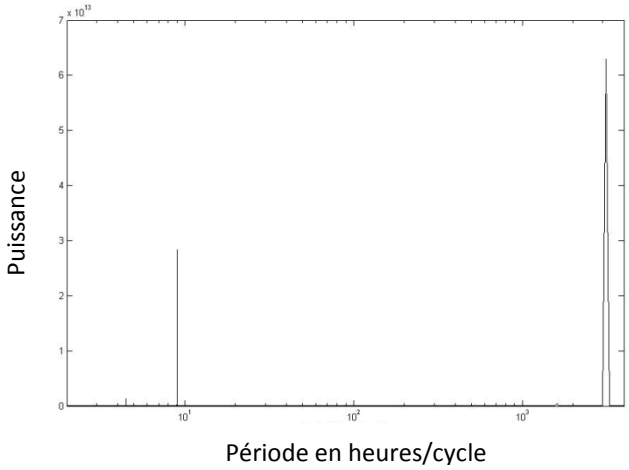


Figure 2-3: spectre du rayonnement global horizontal (seules les heures comprises entre 8h et 16h *TSV* sont considérées, soient 9 mesures par jour)

Il existe cependant un autre type de saisonnalité beaucoup plus difficile à appréhender, elle est qualifiée de « souple » et est aléatoire en amplitude et/ou en période. Nous verrons dans le chapitre 3 une méthode de prise en compte de cette saisonnalité, ce qui permettra d'envisager et d'élargir jusqu'à la notion même de stationnarité. Le paragraphe suivant va décrire les principales méthodes d'analyse des séries temporelles du domaine de la prédiction à horizon et notamment dans le cadre de la quantification du rayonnement global.

3. Les méthodes de prédiction à horizon

La prédiction à horizon des séries temporelles est un problème qui recouvre de nombreux domaines d'application. Les études menées dans la finance et l'économétrie ont permis de dégager de nombreux modèles plus ou moins sophistiqués. Ces derniers ont été repris dans le cadre d'autres thématiques, dont la prédiction du rayonnement global par modélisation des séries temporelles. Dans cette section, différentes approches de modélisation seront brièvement présentées. Le but étant de mieux situer le cadre général auquel appartiennent les modèles qui seront utilisés plus tard dans ce manuscrit.

3.1. Introduction

Il existe de nombreux modèles permettant de faire une prédiction de *ST*. Il est possible de les rassembler en quatre grands groupes [Brockwell et Davis 2006 ; Hamilton 1994] :

- Les **modèles de type « naïf »** qui sont primordiaux pour vérifier la pertinence des modèles complexes. On peut citer la persistance, la moyenne ou les k plus proches voisins ;

- **Les modèles à probabilités conditionnelles** rarement mentionnés dans la littérature en ce qui concerne le rayonnement global. On peut citer les chaînes de Markov et les prédictions basées sur les inférences Bayesiennes ;

- Les **modèles de référence,** de par le nombre d'études les ayant utilisés, qui sont issus de la grande famille des modèles autorégressifs à moyenne mobile saisonniers, les *SARIMA* ;

- Les **modèles de type connexionnistes (réseau de neurones)** et plus particulièrement le Perceptron MultiCouche (*PMC*) qui est un type de réseaux de neurones à fort potentiel prédictif et le plus souvent utilisé.

Il existe un très grand nombre de modèles, aussi la liste ci-dessus ne représente qu'une partie de l'ensemble des prédicteurs, et ne se veut absolument pas exhaustive. Les modèles énumérés ici sont ceux qui ont été étudiés tout au long de cette thèse et en rapport avec notre étude bibliographique. Un descriptif de chacun d'eux fait suite à ce paragraphe.

3.2. Persistance

La persistance est un prédicteur de type « naïf » très souvent choisi pour le rayonnement global et la puissance *PV*. C'est en fait la répétition d'une mesure de l'instant t à l'instant $t+h$ (dans le cas $h = 1$ on obtient l'Equation 2-8). Dans le cas de phénomènes T-périodiques, on peut aussi utiliser la valeur à l'instant $t+h-T$ pour prédire l'instant $t+h$, pour $h < T$ (Cf Equation 2-9 pour le cas $h = 1$).

$$\hat{x}_{t+1} = x_t \qquad (2\text{-}8)$$

$$\hat{x}_{t+1} = x_{t+1-T} \qquad (2\text{-}9)$$

On notera que dans le cas de phénomènes doublement périodiques, tels que le rayonnement global horaire, les expressions restent similaires pour $T=T_1$ ou $T=T_2$. Il est possible d'améliorer ce type de modèle en utilisant une méthodologie « hybride » groupant la persistance avec un modèle de connaissance. Dans le cas du rayonnement global *X*, si H_{gh} est une modélisation du phénomène (modèle ciel clair dans cet exemple défini au chapitre 1 section 2-1) dont l'expression analytique de la tendance est connue, on pourra corriger la persistance et ainsi obtenir l'Equation 2-10 en remplacement de l'Equation 2-8.

$$\hat{X}_{t+1} = X_t . \frac{H_{gh}(t+1)}{H_{gh}(t)} \qquad (2\text{-}10)$$

Ce type de prédicteur est parfois le seul réellement utilisable car il ne nécessite pas d'historique de la série temporelle, seule la valeur à l'instant t est nécessaire.

3.3. Moyenne

La moyenne est une manière de prédire relativement simplement. Cependant, elle est très coûteuse en terme d'historique. Si ces données sont insuffisantes, le bruit présent dans les mesures antérieures ne sera pas correctement pris en compte (il ne sera pas moyenné) et la prédiction se rapprochera de ce que l'on peut avoir avec de la persistance. Dans le cas de mesures *T*-périodiques avec un historique couvrant *N.T* mesures (concernant le rayonnement global journalier, *T* correspond à 365 jours et *N* au nombre d'années d'historique disponibles), il est possible de considérer que la prédiction à l'instant t+h correspond à l'Equation 2-11.

$$\hat{x}_{t+h} = \frac{1}{N} \sum_{i=1}^{N} x_{t+h-T.i} \qquad (2\text{-}11)$$

3.4. K plus proches voisins

La méthode de recherche de voisinage, appelée également méthode des k plus proches voisins (*k-NN*), est performante dans de nombreux domaines dont la prédiction de séries temporelles. Cette méthodologie issue de l'intelligence artificielle (*IA*) consiste à trouver dans l'historique (appelé données d'apprentissage), un groupe de données (formé de k données) le

plus semblable à celui dont on veut prédire les valeurs suivantes [Sharif et Burn 2006 ; Yakowitz 1987]. Il existe de nombreux algorithmes permettant de réaliser cet apprentissage supervisé. Aussi dans les différentes manipulations relatives à cette thèse et au rayonnement global, nous avons choisi une méthodologie basée sur une recherche d'extrema. Elle est décrite par les Equations 2-12 et 2-13 et représente la prédiction au temps *t+1* (horizon 1). Les arguments de cette prédiction sont relatifs à l'erreur d (distance ou écart quadratique) de similarité entre les mesures à l'instant présent et les mesures passées, pour N mesures et r configurations disponibles ($1 \leq r \leq N - k + 1$ et $k \ll N$).

$$r_0 = argmin_r(d(x_t, x_{t-r}) + d(x_{t-1}, x_{t-r-1}) + \cdots + d(x_{t-k}, x_{t-r-k}) \quad (2\text{-}12)$$

$$x_{t+1} = x_{t-r_0+1} \quad (2\text{-}13)$$

Suivant la valeur de k retenue, la prédiction peut totalement différer. L'optimisation de ce paramètre dépend de la série considérée.

3.5. Inférences Bayésiennes

Cette méthode s'appuie sur la probabilité qu'un événement futur apparaisse connaissant les événements antérieurs [Diday et al. 1982 ; Pole et al. 1994 ; Celeux et Nakache 1994, Lauret et al. 2008]. Pour estimer la probabilité que la série soit dans l'état y_k à l'horizon *h*, il est possible d'utiliser le théorème de Bayes comme décrit dans l'Equation 2-14 à partir de deux évènements *A* et *B* et de la relation reliant leurs probabilités conditionnelles.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (2\text{-}14)$$

L'expression de la méthodologie de prédiction est décrite dans l'Equation 2-15 et l'Equation 2-16, $\aleph$ représente l'ensemble des mesures initiales qui sont connues et y_k la valeur de la $k^{ième}$ classe (N mesures disponibles).

$$k_0 = argmax_k(P(x_{t+1} = y_k|\aleph)) = argmax_k(P(x_{t+1} = y_k).P(\aleph|x_{t+1} = y_k)) \quad (2\text{-}15)$$

$$x_{t+1} = y_{k_0} \quad (2\text{-}16)$$

Pour résoudre l'Equation 2-15 on peut se servir de la condition d'indépendance issue de l'Equation 2-17 ($1 \leq J \leq N$), permettant de transformer la probabilité $P(\aleph|x_{t+1} = y_k)$ en un produit d'éléments simples (en anglais méthode appelée « Naive Bayes »).

$$P(\aleph|x_{t+1} = y_k) = \prod_{j=1}^{J} P\left(x_{t+1-j}|x_{t+1} = y_k\right) \quad (2\text{-}17)$$

Grâce à cette simplification, on obtient une nouvelle formulation de la prédiction (Equations 2-18 et 2-19).

$$k_0 = argmax_k\left(P(x_{t+1} = y_k).\prod_{j=1}^{J} P\left(x_{t+1-j}|x_{t+1} = y_k\right)\right) \quad (2\text{-}18)$$

$$x_{t+1} = y_{k_0} \quad (2\text{-}19)$$

Pour recourir à ce type de prédicteur il est nécessaire d'établir préalablement la table des probabilités conditionnelles de la série afin de quantifier le dernier terme de l'Equation 2-18.

3.6. Chaîne de Markov

Dans le domaine des prévisions des séries temporelles, certains auteurs ont eu recourt aux processus de Markov [Logofet et Lesnaya 2000], en particulier aux chaînes de Markov [Muselli et al. 2001]. C'est un processus stochastique qui possède la propriété Markovienne signifiant que, étant donné l'état présent, les états futurs seront indépendants des états passés. En d'autres termes, la description de l'état actuel permet de définir les éléments qui pourraient influencer l'évolution du processus [Muselli et al. 2001]. Les états futurs seront modélisables grâce à un processus probabiliste dépendant uniquement des états présents. Pour initier l'utilisation de ces processus il est nécessaire de calculer d'abord la matrice de transition du phénomène. La probabilité de transition de l'état i à l'état j est définie par $p_{i,j}$ (Equation 2-20). La famille de ces probabilités est appelée la matrice de transition de la chaîne de Markov et est notée R_M.

$$p_{i,j} = P(x_t = j|x_{t-1} = i) \quad (2\text{-}20)$$

La formulation de la prédiction à l'instant t+h, est donnée par l'équation 2-21.

$$x_{t+h} = x_t.(R_M)^h \quad (2\text{-}21)$$

Le choix de la dimension de la matrice de transition est dépendant du nombre de classes que l'on considère. Pour une précision optimale, il faudra en prendre un nombre important. Néanmoins, cette condition est tributaire du nombre d'événements présents dans l'historique d'apprentissage. En effet, avec un nombre de classes élevé et peu d'historique, la matrice de transition comportera énormément de valeurs nulles. Le système d'Equations 2-22 et 2-23 où h = 3 et avec y_{k0} qui représente la valeur de la $k_0^{\text{ième}}$ classe et e_k qui représente les k vecteurs de base de la matrice de transition ($k \in [1,50]$), est un exemple lié à l'utilisation de ce type de prédicteur.

$$k_0 = argmax_k\left((x_t.R_M + x_{t-1}.(R_M)^2 + x_{t-2}.(R_M)^3).e_k\right) \quad (2\text{-}22)$$

$$x_{t+1} = y_{k_0} \quad (2\text{-}23)$$

3.7. Modèles autorégressifs à moyenne mobile saisonniers

La première étape avant de traiter des modèles complexes de type autorégressif à moyenne mobile saisonnière (*SARIMA)*, est d'expliciter ce que sont les modèles à moyenne mobile (*MA*) et les modèles autorégressifs (*AR*). En augmentant le degré de complexité, il vient ensuite les modèles autorégressifs à moyenne mobile (*ARMA*), puis en dernier lieu les modèles *SARIMA* [Bourbonnais 1998 ; Bourbonnais et Terraza 2008 ; De Gooijer et al. 2006 ; Hamilton 1994]. Tous ces modèles s'appliquent sur des processus univariés à temps discret. Excepté *SARIMA*, l'ensemble des modèles autorégressifs nécessite une hypothèse de stationnarité (détaillé dans le paragraphe suivant) [Hamilton 1994]. Des notions de statistiques étant nécessaires à la bonne utilisation de ces modèles, nous allons commencer par expliciter ce que signifie la stationnarité d'un processus, puis ce que sont les coefficients d'autocorrélation.

a. Stationnarité et autocorrélation

Si l'on se réfère à l'Equation 2-1 (dans ce chapitre, section 2-1, p 69), il convient d'estimer f_n pour connaître avec une erreur arbitraire, ce que sera la valeur future de la série temporelle. Souvent, et cela est d'autant plus vrai dans le cas *ARMA*, l'hypothèse de stationnarité est nécessaire à cette estimation [Leisch et al. 1998]. Cette condition est

synonyme de processus stable. Cela est lié au fait que certaines grandeurs comme la moyenne ou la variance changent suivant la période considérée, ou au contraire restent constantes. En fait, la série temporelle est appelée « faiblement stationnaire » si le premier moment est constant et si la covariance (*cov*) est indépendante du temps comme représentés sur l'Equation 2-24. Cette fonction d'autocovariance décrit la relation entre l'information contenue aux instants *t* et *t+h*, on la note $\gamma(h)$.

$$E[x_t] = \mu(t) = \mu \ \ and \ cov[x_t, x_{t+h}] = E[(x_t - \mu)(x_{t+h} - \mu)] = \gamma(h) \ \forall t, h \qquad (2\text{-}24)$$

Notons qu'un critère de stationnarité équivalent peut être fondé, en utilisant le coefficient de corrélation simple (*corr*). La covariance et la corrélation son liées par la relation décrite par l'Equation 2-25.

$$cov[x_t, x_{t+h}] = E[(x_t - \mu)(x_{t+h} - \mu)] = E[x_t . x_{t+h}] - \mu^2 = corr[x_t, x_{t+h}] - \mu^2 \ (2\text{-}25)$$

Ce critère de stationnarité est très souvent suffisant pour une utilisation de modèle de type *ARMA*, cependant un critère plus robuste existe. Il implique la stationnarité sur la globalité du processus. C'est-à-dire que non seulement la moyenne et la variance sont constantes mais tous les autres moments sont invariants pour tout changement de l'origine du temps. Dans ce cas, la série est dite « strictement stationnaire » et les deux premiers critères énoncés précédemment sont respectés. Cependant une condition supplémentaire est nécessaire. La densité de probabilité[10] (notée *F*) du processus stochastique x_t est invariant par translation dans le temps. Cette hypothèse est décrite par l'Equation 2-26.

$$F(x_{1,} \dots . x_{t,}) = F(x_{1+h,} \dots . x_{t+h}) \ \forall h \qquad (2\text{-}26)$$

Pour utiliser correctement les critères de stationnarité, il convient d'estimer le coefficient d'autocorrélation simple. Cette estimation notée ρ_h est représentée par l'Equation 2-27.

$$corr[x_t, x_{t+h}] \cong \rho_h = \frac{\sum_{t=h+1}^{N}(x_t - \bar{x}).(x_{t-h} - \bar{x})}{\sqrt{\sum_{t=h+1}^{N}(x_t - \bar{x})^2 . \sum_{t=h+1}^{N}(x_{t-h} - \bar{x})^2 .}} \qquad (2\text{-}27)$$

10 En théorie des probabilités ou en statistiques, une densité de probabilité est une fonction qui permet de représenter une loi de probabilité sous forme d'intégrales.

Dans l'analyse des séries temporelles, il est souvent utile d'analyser le coefficient de corrélation partiel noté ρ_{hh} (Equation 2-28). Ce dernier permet d'étudier la relation d'information exclusive entre x_t et x_{t+h}.

$$\rho_{hh} = corr[x_t, x_{t+h} | x_{t+1}, \dots, x_{t-1+h}] \quad (2\text{-}28)$$

L'Equation 2-29 montre le cheminement pour calculer ce coefficient à partir du rapport des déterminants de deux matrices. Ces dernières (P_h *et* P_h^*) sont construites spécifiquement pour le calcul des autocorrélations partielles. La matrice P_h est symétrique et sa formulation est décrite par l'Equation 2-30. La matrice P_h^* est équivalente à la matrice P_h où la dernière colonne est remplacée par le vecteur [$\rho_1 .. \rho_h$]).

$$\rho_{hh} = |P_h^*| / |P_h| \quad (2\text{-}29)$$

$$P_h = \begin{pmatrix} 1 & \rho_1 & \dots & \rho_{h-1} \\ \rho_1 & 1 & & \vdots \\ \vdots & & \ddots & \rho_1 \\ \rho_{h-1} & \dots & \rho_1 & 1 \end{pmatrix} \quad (2\text{-}30)$$

A partir de $h > 3$, il est plus facile d'utiliser la forme récurrente proposée par l'Equation 2-31.

$$\rho_{ii} = \begin{cases} \rho_1 \; pour \; i = 1 \\ \frac{\rho_i - \sum_{j=1}^{i-1} \rho_{i-1,j} \rho_{i-j}}{1 - \sum_{j=1}^{i-1} \rho_{i-1,j} \rho_j} \; pour \; i = 2 \dots h \end{cases} \quad (2\text{-}31)$$

La stationnarité est une hypothèse importante, dans le cas *ARMA* elle est même primordiale (hypothèse forte d'utilisation). De plus, les coefficients d'autocorrélations que l'on vient de définir, vont servir de base à l'élaboration des modèles *AR* que l'on se propose de détailler dans la suite.

b. Les modèles autorégressifs à moyenne mobile saisonniers

On appelle un processus *MA(q)* un processus linéaire x_t avec $t \in \mathbb{Z}$ dont les coefficients θ_i sont réels (**θ_0**= 1) et vérifient une relation du type de l'Equation 2-32, (ϵ_t est un bruit blanc de variance σ^2).

$$x_t = \sum_{i=0}^{q} \theta_i . \epsilon_{t-i}, \forall t \in \mathbb{Z} \qquad (2\text{-}32)$$

La prédiction d'une série temporelle est particulièrement simple quand elle peut être « bien approximée » par un modèle autorégressif paramétrique. La formulation de la prédiction pour x_t est simplement décrite par le modèle *AR(p)* de l'Equation 2-33.

$$x_t = \sum_{i=1}^{p} \varphi_i . x_{t-i} + \epsilon_t, \forall t \in \mathbb{Z} \qquad (2\text{-}33)$$

La notation réduite faisant intervenir l'opérateur retard L satisfaisant l'Equation 2-34 transforme l'Equation 2-32 en 2-35 et 2-33 en 2-36.

$$L^i . x_t = x_{t-i} \qquad (2\text{-}34)$$

$$x_t = \theta(L)\epsilon(t) \; ; \; \theta(L) = \sum_{i=0}^{q} \theta_i L^i \qquad (2\text{-}35)$$

$$\varphi(L)x_t = \epsilon(t) \; ; \; \varphi(L) = 1 - \sum_{i=1}^{p} \varphi_i L^i \qquad (2\text{-}36)$$

Le processus couplant les méthodologies *AR*(p) et *MA*(q) se nomme *ARMA*(p,q). C'est un processus stationnaire x_t vérifiant la relation réduite suivante :

$$\varphi(L)x_t = \theta(L)\epsilon_t \qquad (2\text{-}37)$$

On appelle *ARIMA*(p,d,q) un processus x_t pour lequel le processus différentié d'ordre *d*, noté $(1-L)^d x_{\ t}$ vérifie une relation de récurrence du type :

$$\varphi(L)\ (1-L)^d x_t = \ \theta(L)\epsilon_t \qquad (2\text{-}38)$$

Pour le cas saisonnier, il existe un type de modèle basé sur le même principe que précédemment, permettant de s'affranchir de la périodicité saisonnière *s*. Ces modèles s'appellent les *SARIMA*(p,d,q)(P,D,Q) et leur écriture est du type :

$$\varphi_p(L)\ \varphi_P(L)^s\ (1-L^s)^D (1-L)^d x_t = \ \theta_q(L)\theta_Q(L)^s\ \epsilon_t \qquad (2\text{-}39)$$

L'estimation des paramètres des modèles présentés ici, peut s'effectuer avec différentes méthodes. Nombre de ces méthodes sont sophistiquées et leur compréhension demande des connaissances importantes en algorithmique. Nous avons utilisé dans ce rapport, une méthode classique pour l'estimation de la partie *AR* : les équations de Yule Walker [Hamilton 1994]. Concernant la partie *MA*, nous avons utilisé un algorithme itératif basé sur le maximum de vraisemblance calculé par filtre de Kalman comme détaillé dans la toolbox Matlab™ E4

[Teirceiro et al. 2000] dédiée à l'analyse des séries temporelles [De Jong et Chu-chun-lin 1994]. Nous n'allons détailler ici, que le cas simple sans partie *MA*. Cela revient à se focaliser uniquement sur l'étude des équations de Yule Walker [Hamilton 1994] ou sur la méthodologie des Moindres Carrés Ordinaires (*MCO*).

Concernant la première méthode, si l'on reprend l'Equation 2-27 (paragraphe précédent) et la définition des autocovariances (ou après normalisation la définition des autocorrélations simples), on peut estimer les coefficients autorégressifs à partir de la forme de récurrence donnée par l'Equation 2-40 ou bien, à partir de sa forme matricielle représentée par l'Equation 2-41.

$$\gamma(h) = \sum_{k=1}^{p} \varphi_k \, \gamma(h-k) \; \forall \; h = 0, \dots, p \qquad \textit{(2-40)}$$

$$\begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{pmatrix} = \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & & \gamma(p-2) \\ \vdots & & \ddots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \cdots & \gamma(0) \end{bmatrix} . \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_p \end{pmatrix} \qquad \textit{(2-41)}$$

Dans ce manuscrit, l'estimation des paramètres *SARIMA* a nécessité l'emploi de deux logiciels dédiés : Scilab™ et sa toolbox grocer™ ainsi que Matlab™. Le premier, qui parait plus adapté dans le cas où il y a une partie *MA* (transposition de l'algorithme de Teirceiro réalisé initialement pour Matlab™) et le second dans le cas uniquement *AR* (algorithme de Yule Walker).

3.8. Les réseaux de neurones artificiels

Cette méthodologie tient une part importante dans les travaux développés durant cette thèse, aussi, nous avons décidé de présenter un peu plus en détail la genèse de cette approche. Les Réseaux de Neurones Artificiels (*RNA*), fabriqués à partir de structures cellulaires artificielles, constituent une approche permettant d'aborder sous des angles nouveaux les problèmes de perception, de mémorisation, d'apprentissage et de raisonnement, en d'autres termes d'Intelligence Artificielle (notée *IA*) [Crone 2005 ; Cortez et al. 2001 et 2002 ; De Gooijer et al. 2006 ; Faraway et Chatfield 1995]. Actuellement l'utilisation des réseaux de neurones est importante et la recherche continue à mobiliser les énergies. Les réseaux de neurones s'avèrent être des alternatives très prometteuses pour contourner certaines difficultés

ou limites rencontrées par les méthodes numériques classiques. Grâce à leur traitement parallèle de l'information et à leurs mécanismes inspirés des cellules nerveuses (neurones), ils infèrent des propriétés émergentes permettant de solutionner des problèmes complexes. Les premiers travaux sur les *RNA* ont été menés par Mc Culloch et Pitts en 1943 [Mc Culloch etPitts 1943]. Leur but était de comprendre les propriétés des systèmes nerveux à partir de composants élémentaires. Grâce à des modèles à base de neurones simplifiés (les neurones formels), ils montrent qu'il est possible de construire des systèmes capables de calculer des fonctions logiques. En 1949, Donald Hebb [Hebb 1949] amorce le problème de l'apprentissage : la loi de Hebb décrit la manière dont les cellules apprennent à modifier l'intensité des connexions (appelées poids) qui les relient. Dans les années 50, des chercheurs ont commencé à développer des modèles « software » (logiciel) mais aussi « hardware » (matériel) afin de simuler les fonctions de base du système nerveux. En 1956, une grande conférence était organisée à Darmouth sur le thème de l'intelligence artificielle et de l'apprentissage : elle sera le point de départ de l'âge d'or des *RNA* et de l'*IA*. Peu de temps après, Rosenblatt crée le premier Perceptron [Rosenblatt 1962], un réseau de neurones artificiels capable de classer des individus en catégories de façon robuste et capable de conserver ses performances face à des données bruitées et à des dégradations de sa structure (pertes de cellules, modifications d'une partie des connexions). De leur côté, Widrow et Hoff [1960] propose l'Adaline (Adaptative LINear Element) : le premier réseau de neurones destiné à résoudre un problème du monde réel, la suppression d'échos sur une ligne téléphonique. En 1965, Nilsson publie « Machine Learning » [1965] qui donne les fondements mathématiques de l'apprentissage automatique pour la reconnaissance des formes. Après des débuts prometteurs, les réseaux de neurones subissent la concurrence des techniques de calculs plus classiques et tombent en désuétude suite notamment aux critiques de Minsky et Paper [1969]. Ceux-ci mettent en évidence l'incapacité des réseaux à une couche à calculer l'opérateur logique XOR et les difficultés théoriques posées par l'apprentissage dans les réseaux multicouches. Les autres techniques d'intelligence artificielle vont conserver les faveurs des chercheurs pendant dix ans avant de montrer elles aussi leurs limites. Les réseaux de neurones vont réapparaitre dans les années 80 avec les travaux de Hopfield [1982 ; 1984] sur les réseaux entièrement interconnectés, ceux de Kohonen [1982] avec les cartes auto-organisatrices à apprentissage non supervisé et enfin ceux de McClelland et Rumelhart [1986a ; 1986b] qui introduisent l'algorithme d'apprentissage de rétropropagation de l'erreur en 1986.

Le prochain paragraphe vise à exposer succinctement les principales familles de réseaux de neurones, après avoir présenté la base de tout réseau, à savoir le neurone isolé, aussi appelé neurone formel.

a. Le neurone formel

Pour bien comprendre ce que sont les réseaux de neurones, nous allons commencer par expliciter l'élément fonctionnel nommé neurone formel. Ce dernier est une fonction algébrique paramétrée, à valeurs bornées, de variables réelles appelées entrées et nommées x_j (Figure 2-4). Celles-ci sont multipliées par un poids ω_{ij}^1 pour donner le produit $x_j\omega_{ij}^1$ qui est ensuite pris comme argument d'une fonction de transfert (ou d'activation) g pour former la sortie $y_i = g(\sum_{j=1}^{n} x_j\, \omega_{ij}^1 + b_i^1)$. Le poids qui n'est pas connecté aux entrées porte le nom de biais (constante d'ajustement indépendante des entrées ; b_i^1).

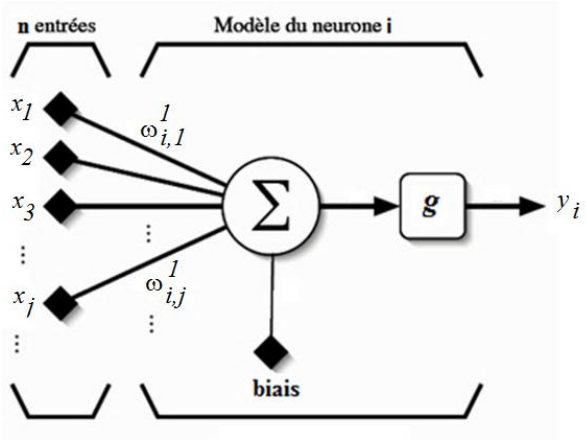


Figure 2-4 : principe du neurone formel

Les fonctions d'activation ont évolué au cours du temps. Dans les versions originelles, on utilisait une fonction d'activation de type « seuil », c'est-à-dire une fonction qui prenait uniquement deux valeurs, 0 ou 1, selon que la force des signaux arrivant au neurone dépassait un certain seuil dont la valeur est l'opposé de celle du biais (Figure 2-4). Malheureusement, ces fonctions (de type Heaviside ou échelon) ne sont pas dérivables sur $\mathbb{R}$, ce qui génère des difficultés lors de la détermination des poids et des biais lors de la phase d'apprentissage. Afin de pallier à ce problème de non-dérivabilité, plusieurs fonctions, dites sigmoïdes, dont la forme s'approche de la fonction échelon, peuvent être utilisées. Les fonctions classiquement

utilisées sont la fonction linéaire : $g(x) = x$, la tangente hyperbolique : $g(x) = \tanh(x) = e^{x} - e^{-x}/e^{x} + e^{-x}$ et la fonction sigmoïde standard : $g(x) = (\tanh(x) + 1)/2$. Ces fonctions sont bornées et proposent des valeurs de sortie comprises entre 0 et +1, ou entre -1 et +1. Il est aussi possible d'utiliser d'autres types de fonctions dérivables (gaussiennes, autres sigmoïdes, etc.). En fonction de l'intensité du signal qui arrive au niveau de la fonction d'activation, le neurone sera soit « allumé » soit « éteint ». Dans le premier cas, cela signifie que la sortie sera égale à 1, et dans le second que la sortie sera -1 (0 dans le cas de la fonction gaussienne ou de certaine sigmoïde). Notons que d'autres valeurs intermédiaires peuvent être atteintes ($\in [-1,1]$ ou $[0,1]$) en dehors des sorties booléennes classiques par l'intermédiaire de fonctions d'activation de type sigmoïde permettant la « fuzzification » des sorties (terme issu de. fuzzy logic ; logique floue).

b. Les familles de réseaux de neurones

Un neurone artificiel isolé n'a que peu d'intérêt et est rarement utilisé. Cependant, une fois interconnectés avec d'autres neurones, ils forment alors un réseau capable de résoudre des problèmes très complexes : classification, reconnaissance de forme, prédiction de séries temporelles, etc. C'est souvent une solution idéale lorsque l'on dispose de beaucoup de données et que l'on ne connait pas les règles qui régissent les phénomènes que l'on veut modéliser. Dans ce paragraphe, nous allons lister les différentes familles de réseaux qui agencent les neurones. Le réseau le plus utilisé dans la littérature est sans conteste le Perceptron MultiCouche (*PMC*) qui est organisé en plusieurs couches de neurones formels. L'information circule de la couche d'entrée vers la couche de sortie uniquement. Chaque couche est constituée d'un nombre variable de neurones et les neurones de la couche de sortie correspondant toujours aux sorties du système. Pour plus de détails sur ce type d'outil, le lecteur intéressé pourra se référer à la section 3 du chapitre 3.

Concernant les autres familles de *RNA*, on peut aussi citer celles des réseaux de neurones bouclés ou récurrents [Dreyfus 2004]. Contrairement aux réseaux de neurones non bouclés (comme les *PMC*) dont le graphe de connexion est acyclique, ils peuvent avoir une topologie de connexions quelconque, comprenant notamment des boucles qui ramènent aux entrées la valeur d'une ou plusieurs sorties [Elman 1990]. Pour qu'un tel système soit causal, il faut évidemment qu'à toute boucle soit associé un retard [Elman 1991]. La forme générale des

équations régissant un réseau de neurones bouclés est appelée forme canonique. En fait, tout réseau de neurones bouclé, aussi compliqué soit-il, peut être ramené à une forme canonique, comportant un réseau de neurones non bouclé dont certaines sorties (les variables d'état) sont ramenées aux entrées par des bouclages de retard unité (e.g. utilisation de l'opérateur retard *L*).

L'autre grande famille de réseau de neurones concerne les réseaux de fonctions radiales (ou *RBF* en anglais pour « Radial Basis Function »). Les fonctions radiales ont été introduites par Powell [1985] dans le cadre de l'interpolation, c'est-à-dire de la recherche de fonctions passant exactement par un nombre fini de points (appelés « points de collocation »). La recherche d'une fonction passant exactement par des points n'a de sens que si ces points ne sont pas entachés de bruit. L'article de référence des réseaux *RBF* est écrit par Broomhead et Lowe [1988], il propose d'utiliser ces réseaux pour l'approximation de fonctions non linéaires. La fonction recherchée est toujours une combinaison linéaire de fonctions radiales, mais leur nombre est beaucoup plus petit que le nombre de points et elles ne sont donc pas forcément centrées en ces points. La propriété d'approximateurs universels pour ces réseaux n'a été que très récemment prouvée pour des fonctions d'activation gaussienne. Ces réseaux ont été utilisés fréquemment comme outil de modélisation de type « boîte noire ». Cependant leur utilisation est plus dédiée aux processus de type entrée-sortie en automatisme qu'à la prédiction à horizon.

En règle générale, les réseaux de neurones sont couplés à un algorithme « d'entraînement » ou d'apprentissage qui consiste à modifier les poids synaptiques en fonction d'un jeu de données présentées en entrée du réseau. Le but de cet entraînement est de permettre au réseau de neurones « d'apprendre » à partir des exemples. Si l'entraînement est correctement réalisé, le réseau est capable de fournir des réponses en sortie très proches des valeurs d'origines du jeu de données d'entraînement. Les réseaux que l'on vient de présenter sont dit à apprentissages supervisés et bien d'autres modèles sont intégrés à cette grande thématique *RNA*. De plus, il existe aussi des réseaux à apprentissage non supervisés : le réseau est laissé libre de converger vers n'importe quel état final lorsqu'on lui présente une entrée. Pour n'en citer qu'un faisant partie de cette famille de *RNA*, il existe le réseau de Hopfield [1982] qui réalise une mémoire adressable par son contenu. Il s'agit d'un réseau récurrent (la propagation de l'information ne s'effectue pas de l'entrée vers la sortie) constitué de neurones formels à deux états (-1 et 1 ou 0 et 1), dont la loi d'apprentissage est la règle de

Hebb [1949]. Un poids synaptique améliore son activité si et seulement si l'activité des neurones adjacents est corrélée. Cela signifie que le poids ω_{ij} d'une connexion entre un neurone *i* et un neurone *j* augmente quand les deux neurones sont activés au même temps. La liste des familles de réseaux que nous venons d'exposer montre à quelpoint il est difficile de trouver celle qui va permettre de solutionner un problème posé, tant il existe de variétés différentes. Pour réussir à dégager les plus pertinents de chaque problématique, il est préconisé d'étudier la littérature sur le sujet choisi.

La prochaine section compare la pertinence des différents modèles que l'on vient d'énumérer (réseau de neurones, *SARIMA*, *k-NN*…) concernant des problèmes de prédiction de séries temporelles quelconques et plus particulièrement lors de la prédiction du rayonnement global horizontal.

4. Comparaison bibliographique des prédicteurs présentés

Avant d'entreprendre la comparaison des différents prédicteurs que l'on vient d'énoncer, il convient de présenter les principaux paramètres d'erreur permettant de comparer les approches. Outre les paramètres standard de comparaison, nous présenterons aussi les outils spécifiques à notre thématique et utilisés durant cette thèse, comme l'indice de similarité ou encore l'erreur de tendance.

4.1. Estimation de l'erreur de prédiction

Dans le cadre de la prédiction de séries temporelles de rayonnement global ou de puissance *PV*, il existe de nombreuses méthodes pour vérifier si un prédicteur est efficace [Ahlburg 1992 ; Kalogirou 2001 ; Mellit et Kalogirou 2008]. Durant cette thèse, la méthode utilisée est dite de « validation-croisée ». Ainsi, on a comparé deux à deux les prédicteurs en affectant à chacun une erreur de prédiction sur un échantillon test identique. Les outils que nous avons utilisés sont décrits ci-après :

- **le coefficient de détermination** (R^2) est un indicateur qui permet de juger la qualité d'une régression linéaire, simple ou multiple. D'une valeur comprise entre 0 et 1, il mesure l'adéquation entre le modèle et les données observées. Dans le cadre d'une régression

linéaire simple, c'est le carré du coefficient de corrélation. Le R^2 se définit comme la part de variance expliquée par rapport à la variance totale, autrement dit par [1 – (somme des carrés des résidus / variance totale)]. Ce coefficient (Equation 2-42) s'applique aussi bien à une régression simple qu'à une régression multiple.

$$R^2 = 1 - \frac{\sum_i (x_i - \hat{x}_i)^2}{\sum_i (x_i - <x>)^2} \qquad (2\text{-}42)$$

Il est possible d'utiliser le coefficient de détermination ajusté car ce dernier tient compte du nombre de variables[11]. Cependant, dans nos travaux comme le nombre de variables était très largement inférieur au nombre d'observation, il n'y avait pas grand intérêt à manipuler cette grandeur ;

- **l'erreur quadratique moyenne** (Mean Square Error ; *MSE*) est très utile pour comparer plusieurs estimateurs, notamment lorsque l'un d'eux est biaisé. On peut effectivement exprimer l'erreur quadratique moyenne en fonction du biais (*B*) et de la variance (σ^2), on a ainsi :

$$MSE = E[(\hat{x} - x)^2] = E(\hat{x} - x) + E\left[\left(\hat{x} - E(\hat{x})\right)^2\right] = B(\hat{x}) + \sigma^2(\hat{x}) \qquad (2\text{-}43)$$

Si les deux estimateurs à comparer sont sans biais, l'estimateur le plus efficace est simplement celui qui possède la variance la plus petite. Cette valeur de *MSE* n'est pas très significative pour une interprétation absolue de l'indicateur. En effet, on pourra utiliser cet indicateur pour hiérarchiser deux prédicteurs sur un problème donné, mais on ne pourra pas facilement donner un sens à cette erreur de prédiction. Pour manipuler une erreur basée sur la même échelle que la mesure, on peut utiliser le *RMSE* (Root Mean Square Error) défini par :

$$RMSE = \sqrt{MSE} \qquad (2\text{-}44)$$

Si l'on veut comparer des études faites sur deux sites géographiquement éloignés, le *RMSE* n'est pas utilisable. En effet, si la ressource solaire est plus abondante sur un site, on aura alors un *RMSE* qui sera aussi plus important, ce qui ne veut pas dire que le prédicteur utilisé est moins bon. Pour pallier à ce problème, il est possible de normaliser l'erreur par la valeur moyenne du rayonnement. On obtient ainsi un pourcentage qui n'est pas représentatif

[11] Pour *n* observations et *p* variables de régression : $R^2_{ajusté} = 1 - \frac{n-1}{n-p-1}.(1 - R^2)$.

de l'ordre de grandeur du rayonnement mesuré. Il existe plusieurs méthodes pour calculer ce *RMSE* normalisé (*nRMSE)*. Durant cette thèse, nous avons opté pour l'utilisation du rapport à la moyenne quadratique qui est en règle générale un critère robuste vis à vis des séries dont certains éléments peuvent être négatifs. La moyenne quadratique est nulle si et seulement si tous les éléments qui la composent sont nuls, concernant les autres moyennes cette propriété n'est plus applicable.

$$nRMSE = \sqrt{MSE/\langle x^2 \rangle} \qquad (2\text{-}45)$$

Notons que le fait d'utiliser cette définition tend à faire diminuer l'erreur nominale. En effet, cette définition du *nRMSE* donne des résultats inférieurs à ceux que l'on obtiendrait avec une définition faisant intervenir la moyenne arithmétique du type $nRMSE = \sqrt{MSE}/\langle x \rangle$. En règle générale on a la hiérarchisation suivante : $min[x(t)] < \langle x \rangle. < \sqrt{\langle x^2 \rangle} < max[x(t)]$ $\forall x \in \mathbb{R}$. Outre ce détail calculatoire lié aux différents *nRMSE*, il est possible de déterminer des *nRMSE* périodiques (= saisonniers) ou annuels, suivant l'intervalle sur lequel est calculé cette erreur ;

- **l'erreur moyenne absolue** (Mean Absolute Error ; *MAE*), est une quantité souvent utilisée pour mesurer l'écart des prédictions et des mesures. Sa formule est donnée par :

$$MAE = \langle |x - \hat{x}| \rangle \qquad (2\text{-}46)$$

Tout comme dans le cas du *nRMSE*, il est possible d'utiliser une forme relative de cette erreur, en divisant par la valeur moyenne du rayonnement. Le paramètre *MAPE* (Mean Absolute Pourcentage Error) que l'on a choisi d'utiliser dans cette étude est décrit par :

$$MAPE = MAE/\langle |x| \rangle \qquad (2\text{-}47)$$

Notons que l'utilisation du *MAPE* n'est pas identique à celle du *nRMSE*. En effet, la *MAE* est moins sensible aux erreurs importantes, alors que la *nRMSE* pénalise lourdement les gros écarts du fait de la mise au carrée de l'erreur. Ainsi, diminuer le *MAE* signifiera améliorer en moyenne la prédiction, alors que diminuer le *nRMSE* signifiera plutôt, limiter les grosses erreurs de prédiction ;

- le **biais moyen** (Mean Bias Error ; *MBE*), est défini comme étant l'écart algébrique moyen entre la simulation et la mesure, sa formule est :

$$MBE = \langle x - \hat{x} \rangle \qquad (2\text{-}48)$$

Ce paramètre se rapproche plus de l'erreur sur la prédiction du cumul de la ressource que de l'erreur commise à horizon. En effet, une valeur positive de *MBE* signifie que sur le long terme, le cumul des données de prédiction sous-estime le cumul des mesures réelles alors qu'une valeur négative signifiera l'inverse. L'utilisation de ce genre d'outil n'est pas triviale. En effet, il faut savoir qu'une valeur nulle de *MBE*, ne signifie pas forcément que la prédiction est parfaite. L'exemple qui illustre le mieux ce phénomène est sans doute le modèle classique basé sur la persistance (section 3-2 de ce chapitre, p 74). Avec ce système de prédiction, le *MBE* tend vers zéro, pour autant, les autres erreurs commises peuvent être élevées ;

- **l'erreur cumulée** (*EC*), qui représente non pas l'erreur de chaque prédiction, mais l'erreur globale sur le cumul des prédictions. Une valeur négative, indiquera que la prédiction va surestimer la mesure. Ce type de critère est valable pour représenter une erreur macroscopique, mais n'est pas adéquat pour estimer l'erreur de prédiction à horizon. Tout comme dans le cas précédent, la persistance induit une erreur cumulée faible, alors que l'erreur commise à chaque prédiction est élevée. L'expression de cette erreur est donnée par :

$$EC = \sum_i x_i - \sum_i \hat{x}_i \qquad (2\text{-}49)$$

Les deux estimations de l'erreur qui suivent (erreur de tendance et indice de similarité) ne sont pas conventionnelles. Nous avons ressenti le besoin de les construire durant les différentes manipulations effectuées. En effet, les estimateurs habituels tels que l'on vient de les définir, ne suffisent pas forcément pour hiérarchiser toutes les méthodologies de prédiction, il faut parfois développer de nouveaux outils. L'inconvénient inhérent à de tel processus, réside dans le fait qu'ils ne sont pas utilisés dans les travaux habituels de prédiction de séries temporelles, leur utilisation est donc délicate.

- **l'erreur de tendance** (Δ) qui matérialise si la tendance de la prédiction suit bien la tendance de la mesure, dans le sens où, si le rayonnement augmente, la prédiction augmente aussi, et si elle diminue, la prédiction diminue. Cette erreur est définie par :

$$\left.\begin{matrix}\Delta_m = x_{i+1} - x_i \\ \Delta_p = \hat{x}_{i+1} - x_i\end{matrix}\right\rangle \Delta = \frac{\Delta_m}{\Delta_p} \qquad (2\text{-}50)$$

Dans le cas de *n* prédictions, on obtient *n* valeurs de Δ. On peut classer ces dernières de la façon suivante :

– $\Delta < 0$; très mauvaises interprétations car cela signifie que $\Delta_m < 0 \; et \; \Delta_p > 0$ (ou l'inverse). Les tendances ne sont pas respectées ;

– $\Delta \in [0,1-\varepsilon[\cup [1+\varepsilon,+\infty[$; zone des erreurs de prédictions importantes mais avec une tendance respectée. Durant les travaux présentés ici et relatifs à cette thèse nous avons opté pour $\varepsilon = 0{,}2$ (une étude de sensibilité a motivé ce choix) ;

– $\Delta \in [1-\varepsilon,1+\varepsilon]$; partie qui représente les erreurs de prédictions faibles avec tendance respectée.

- **l'indice de similarité** (dit encore de fiabilité ; η) qui permet de vérifier si une valeur prédite est similaire à la valeur réelle. Cet outil est efficace pour encadrer une prédiction. Dans le cas d'un réseau de neurones, il est possible d'obtenir une série temporelle de η à l'issue de l'apprentissage. Durant la phase de prédiction, on peut accompagner la valeur simulée par son coefficient η. Concernant ce dernier, on aura 100 % si le résultat est sûr et 0 % si la prédiction n'est pas sûre du tout et que l'erreur attendue peut être supérieure à la valeur du rayonnement. Il est calculé par :

$$\eta_t = 100.\left(1 - \frac{|x_t - \hat{x}_t|}{x_t}\right) \qquad (2\text{-}51)$$

A partir de cette valeur, il est possible de déterminer un intervalle de confiance (*IC*) qui encadre la prédiction ($\hat{x}_t \mp IC_t$). Ainsi si $\eta_t = 100\%$ alors $IC_t = 0$ et si $\eta_t = 0\%$ alors $IC_t = \hat{x}_t$. De manière générale, *IC* se calcule par :

$$IC_t = \hat{x}_t.\left(1 - \frac{1}{100}\eta_t\right) \qquad (2\text{-}52)$$

Tous les critères que nous venons d'énumérer ont été utilisés dans ce manuscrit. Cependant, leurs sens et leurs interprétations peuvent différer en fonction du contexte de l'étude. Maintenant que les critères permettant de comparer des méthodologies de prédiction

sont établis, la prochaine section va dresser une comparaison bibliographique concise des méthodes de prédiction relatives au rayonnement global ou à la puissance *PV*.

4.2. Comparaison bibliographique

Certain auteurs ont dédié une grande partie de leur recherche à la prédiction de séries temporelles de rayonnement global ou de la puissance *PV*. C'est notamment le cas de Mellit qui est l'auteur de référence de cette thématique. Son domaine de prédilection est la manipulation des réseaux de neurones. Aussi dans sa revue bibliographique [Mellit et Kalogirou 2008], les résultats trouvés ont été repris et vérifiés dans de nombreux travaux. L'erreur de prédiction (*MAPE*) de rayonnement global mensuel induite par l'utilisation d'un *RNA* est estimée entre 0,2 % et 10,1 % suivant la ville et l'architecture considérée (médiane = 4 %). Les résultats présentés sont si disparates qu'ils paraissent incomparables. Cependant, il faut considérer que certaines localisations ne connaissent que très peu d'occurrences nuageuses alors que d'autres sont sujettes à des climats beaucoup moins cléments. Concernant les prédictions de rayonnement global horaires Sfetsos [Sfetsos et Coonick 2000] a montré que les réseaux de neurones généraient une erreur de 7 % et les méthodologies de type *ARMA* une erreur de 8 %. Il semble donc y avoir une égalité entre ces deux méthodologies, avec tout de même un léger avantage pour l'approche connexionniste. Il démontre aussi que les réseaux récurrents procurent de moins bon résultats que les *PMC*. Il faut toutefois relativiser ce résultat, car suivant l'étude considérée, les réseaux de neurones ne donnent pas toujours les meilleurs résultats, la localisation de l'étude et la méthodologie propre à chaque auteur semblent modifier le classement des prédicteurs. En règle générale, il s'avère mal aisé de dire qui de *RNA* ou *ARMA* est le plus efficace. De plus, il faut noter que la précédente étude est menée en Grèce et que la nébulosité est relativement faible sur la période considérée (prédiction réalisée sur 13 jours en été). A travers ces deux auteurs, une tendance semble se dégager : la synthèse de tous les travaux de prédictions basés sur les séries temporelles est assez délicate, car en règles générales, les approches, les estimations d'erreurs et les spécificités géographiques peuvent être trés différentes. Pour illustrer ce phénomène, Behrang et al. [2010] ont établi une liste des erreurs des prédictions faites avec des réseaux de neurones concernant le rayonnement global (Tableau 2-1).

Source	Localisation	*MAPE* (%)
Rehman et Mohandes [2008]	Arabie Saoudite	4,5
Mohandes et al. [1998]		12,6
Mohandes et al. [2000]		10,9
Behrang et al. [2010]	Iran	5,2
Azadeh et al. [2008]		6,7
Sozen et al. [2004]	Turquie	6,8

Tableau 2-1 : erreur de prédictions faites avec des *PMC* [Behrang et al. 2010]

On constate que pour des localisations pourtant identiques (voir le cas de l'Arabie-Saoudite), les erreurs suivant les auteurs peuvent doubler voire tripler. Les conclusions de ce travail sur les *PMC* peuvent se généraliser aux autres prédicteurs. D'après la bibliographie, les paramètres pouvant influencer la prédiction sont tellement nombreux, qu'il est difficile d'utiliser les résultats issus d'autres études. En l'occurrence, il peut être intéressant d'essayer des méthodes ou des paramètres même s'ils n'ont pas forcément fait leurs preuves dans d'autres études. C'est à ce titre que dans la suite de ce manuscrit nous avons essayé les modèles *SARIMA* et *ANN* car la littérature ne nous permettait pas de les hiérarchiser. Cependant pour le second, nous nous sommes limités au cas *PMC* qui semble faire l'unanimité des prévisions des chroniques de rayonnement global et de puissance *PV*. De plus, et d'après ce qui précède, tous les paramètres inhérents aux *PMC* et aux *SARIMA* doivent être étudiés pour chaque site testé, là aussi il existe des divergences suivant les auteurs. C'est dans cette optique de spécificités particulières des lieux d'études, que nous avons souhaité étudier des estimateurs peu, voire très rarement étudiés dans le domaine des énergies renouvelables. Nous avons donc essayé une méthodologie de prédiction basée sur les inférences Bayesiennes. Seule, cette méthode est très peu employée en prédiction de rayonnement global. On peut néanmoins citer de nombreux travaux visant à la coupler à d'autres prédicteurs comme les réseaux de neurones [Lauret et al. 2008] ou comme test discriminant de sélection de variables [Rivière et al. 2006]. Cependant, cette technique est très utilisée en économétrie, par le biais de publications très théoriques qui ne permettent pas vraiment de la comparer avec d'autres méthodes de prédiction. On peut surtout citer Xiang Fei [Fei et al. 2011], qui a montré que les inférences de Bayes permettent d'égaler une estimation *AR* dès lors que l'on étudie des variables non stationnaires. L'erreur sur la série

étudiée est proche de 10 % pour les deux modèles. Concernant les chaînes de Markov, elles sont aussi peu utilisées en énergétique, on estime d'après le papier de Hoacaoglu [2011] une erreur de prédiction de 6 % pour le rayonnement journalier et d'après Muselli [Muselli et al. 2001] une erreur sur l'énergie *PV* prédite sur surface horizontale de 10 %. D'après ces résultats, il semble logique d'intégrer ce type de prédicteur à notre étude, les erreurs énoncées sont proches de ce que donnent des techniques plus conventionnelles comme *ARMA* ou *RNA*. Les trois autres estimateurs que sont la persistance, *k-NN* et la moyenne, sont très faciles à mettre en œuvre. En effet, il n'y a pas de phase d'apprentissage (sauf pour la moyenne), et peu de contraintes sont nécessaires à leur utilisation (stationnarité, prétraitement, hypothèses, etc.). C'est pour cette raison que nous les avons qualifiés de « prédicteurs naïfs ». Il est logique que les méthodes élaborées donnent de meilleurs résultats, mais il faut toutefois avoir à l'esprit la balance entre complexité du modèle et qualité de prédiction. C'est pour cette raison qu'il est nécessaire de comparer les modèles évolués aux modèles dits « naïfs ».

5. Conclusion

Ce chapitre a permis de poser certaines définitions relatives aux séries temporelles, leur analyse et leur prédiction. Comme nous l'avons vu, le rayonnement global horizontal ou la puissance *PV*, peuvent être perçus comme une succession de mesures et ainsi être interprétés comme des séries temporelles « ordinaires » pour lesquelles tous les outils d'analyse (*ARMA*, *ANN*, Markov, etc.) sont potentiellement dédiés. Ce qui a été présenté ici, n'est bien entendu, pas une liste exhaustive de tout ce qui peut exister, ou peut être étudié. Nous nous sommes attaché à ne détailler que ce qui a été concrètement utilisé lors des différentes simulations. L'analyse bibliographique dressée dans ce chapitre permet de comprendre en quoi la prédiction de séries temporelles est complexe. Les résultats exposés dans ces articles sont tous évidemment pertinents, mais sont bien souvent exclusifs aux conditions des expérimentations. Il est généralement impossible de transposer les résultats sur un autre site ou à une méthodologie voisine avec de petites variantes. Cela explique les contradictions que l'on peut trouver entre certains articles. Concernant les prédictions basées sur les *RNA*, la principale famille d'architecture étudiée est le *PMC*. C'est de loin le plus utilisé, aussi nous nous focaliserons sur ce type d'outil, sans tenir compte des nombreux autres réseaux que l'on a énumérés durant ce chapitre.

Un des objectifs de cette thèse sera d'essayer de trouver une méthodologie de prédiction robuste qui s'adapte au site et à l'horizon de prédiction considéré. Un autre sera, toujours suivant l'horizon considéré, de hiérarchiser ces différents prédicteurs.

Comme nous l'avons vu lors du chapitre 1, outre sa « mise sous forme » de série temporelle, le rayonnement global bénéficie aussi de nombreuses années d'étude, non pas comme chronique, mais comme un « modèle de connaissance ». Le prochain chapitre va détailler la méthodologie de prédiction couplant ces deux approches, méthodologie que nous avons utilisée durant cette thèse. Nous verrons, entre autre, comment « hybrider » les différentes méthodes, et quels sont les outils statistiques qui permettent de stationnariser ou encore de sélectionner les variables intéressantes.

6. Bibliographie

Ahlburg D. Error measures and the choice of a forecast method. International Journal of Forecasting n°6;8(1), p 99-100, 1992.

Azadeh A, Maghsoudi A, Sohrabkhani S. An integrated artificial neural networks approach for predicting global radiation. Energy Conversion and Management, in press, 2010.

Badescu V.. Modelling Solar radiation at the earth surface, recent advances. Springer Ed, 2008.

Behrang MA, Assareh E, Ghanbarzadeh A, Noghrehabadi AR. The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data. Solar Energy, n°84(8), p 1468-1480, 2010.

Bourbonnais R, Terraza M. Analyse des séries temporelles, Dunod Ed., Paris, 2008.

Bourbonnais R. Analyse des séries temporelles en économie. Presses Universitaires de France – PUF, 1998.

Brockwell, PJ, Davis RA. Time series: theory and methods. Springer series in statistics, second edition, 2006.

Bouttier François. The Météo-France NWP system: description, recent changes and plans, CNRM, 2010.

Broomhead DS et Lowe D. Multivariable functional interpolation and adaptative network. Complex system, vol2 p 321-355, 1988.

Celeux G, Nakache JP. Analyse discriminante sur variables qualitatives. Polytechnica, Paris, 1994.

Cortez P, Rocha M, Neves J. Evolving Time Series Forecasting Neural Network Models. Proceeding of int. symposium on adaptive systems: evolutionary computation and probabilistic graphical models, 2001.

Cortez P, Sollari AF, Rocha M, Neves J,. Real-Time Forecasting by Bio-Inspired Models. Proceeding, 2002.

Crone SF. Stepwise Selection of Artificial Neural Networks Models for Time Series Prediction Journal of Intelligent Systems. Department of Management Science Lancaster University Management School Lancaster, United Kingdom, 2005.

De Gooijer JG et Hyndman RJ. 25 years of time series forecasting. International Journal of Forecasting 22, n°. 3, p 443-473, 2006.

De Jong P et Chu-chun-lin S. Satationarity and non stationarity state space models. Journal for time series analysis, vol 15, n°2, 1994.

Diday E, Lemaire L, Pouget J, Testu F. Éléments d'analyse de données, Dunod, Paris, 1982.

Dreyfus G, Idan Y. The canonical form of discrete-time non-linear models. Neural Computation, Vol.10, n°1, p 133-164, 1998.

Dreyfus G. Réseaux de neurones : méthodologie et appliquations. Eyrolles, 2004.

Elman JL. Finding structure in time, Cognitive Science, n°14, p 179-211, 1990.

Elman JL. Distributed representations, simple recurrent networks and grammatical structure. Machine Learning, n°7, p 195-224, 1991.

Faraway J, Chatfield C. Times series forecasting with neural networks: a case study. Research report 95-06 of the statistics group. University of Bath, 1995.

Fei X, Lu C-C, Liu K. A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction. Transportation Research Part C: Emerging Technologies, 2011.

Fröhlich C, Lean J. The Sun's total irradiance: Cycles and trends in the past two decades and associated climate change uncertainties. Geophys. Res. Let., n°25, p 4377-4380, 1998.

Hamilton JD. Times series analysis, 1994.

Hebb DO. The Organization of Behavior, New York, Wiley, 1949.

Hopfield JJ. Neural networks physical systems with emergent collective computational abilities, Proceedings of the National Academy of Sciences, n°79, p 2554-58, 1982.

Hopfield JJ. Neurons with graded responses have collective computational properties like those of two-states neurons, Proceedings of the National Academy of Sciences, n°81, p 3088-92, 1984.

Hocaoglu FO. Stochastic approach for daily solar radiation modeling. Solar Energy n°85(2), p 278-287, 2011.

Kalogirou S. Artificial neural networks in renewable energy systems applications: a review. Renewable and Sustainable Energy Reviews, n°12;5(4), p 373-401, 2001.

Kohonen T. Self-organized formation of topologically correct feature maps. Biological Cybernetics, n°43, p 59-69, 1982.

Lauret P, Fock E, Randrianarivony RN, Manicom-Ramsamy JF. Bayesian neural network approach to short time load forecasting. Energy conversion and management, n°49(5), p 1156–1166, 2008.

Leisch F, Trapelli T, Hornik K. On the stationary of autoregessive neural network models. Report series, adaptive information system and modeling economics and management science, 1998.

Logofet DO, Lesnaya EV. The mathematics of Markov models: what Markov chains can really predict in forest successions. Ecological Modelling n°126, p 285-298, 2000.

McCleland JL, Rumelhart DE. A distributed model of human learning and memory. Journal of experimental psychology, n°114, p 159-188, 1986 a.

McCleland JL, Rumelhart DE. and the PDP Research Group. Parallel Distributed Processing : Explorations in the Microstructures of Cognition, 2 vol., Cambridge, MIT Press, 1986 b.

Mc Culloch WS et Pitts WH. A logical calculus of the immanent in nervous activity. Bulletin of mathematical biophysics,n° 5, p 115-133, 1943.

Mellit A, Kalogirou SA. Artificial intelligence techniques for photovoltaic applications: A review. Progress in Energy and Combustion Science n°1-1, p 52-76, 2008.

Mellit A, Kalogirou SA, Hontoria L, Shaari S. Artificial intelligence techniques for sizing photovoltaic systems: A review. Renewable and Sustainable Energy Reviews n°13-2, p 406-419, 2009.

Minsky LM et Paper SA. Perceptron. The MIT Press, cambridge, 1969.

Mohandes M, Rehman S et Halawani TO, Estimation of global solar radiation using artificial neural networks, Renewable Energy n°14 (1–4), p 179–184, 1998.

Mohandes M, Balghonaim A, Kassas M, Rehman S et Halawani TO. Use of radial basis functions for estimating monthly mean daily solar radiation, Solar Energy n°68, 2000.

Mubiru J. Predicting total solar irradiation values using artificial neural networks. Renewable Energy, n°33-10, p 2329-2332, 2008.

Mueller RW, Dagestad KF, Ineichen P, Schroedter-Homscheidt M et al. Rethinking satellite-based solar irradiance modelling: The SOLIS clear-sky module. Remote Sensing of Environment 91, n°2, p 160-174, 2004.

Muselli M, Poggi P, Notton G, Louche A. First Order Markov Chain Model for Generating Synthetic 'Typical Days' Series of Global Irradiation in Order to Design PV Stand Alone Systems. Energy conversion and management, n°42-6, p 675-687, 2001.

Nilsson NJ. Learning Machines. Foundations cf trainable pattern-classifying system. Mcgraw-Hill, New-York, 1965.

Oussar Y. Réseaux d'ondelettes et réseaux de neurones pour la modélisation statique et dynamique de processus. Thèse de doctorat Paris VI, 1998.

Pole A, West M, Harrison J. Applied Bayesian forecasting and time series analysis. Chapman and Hall/CRC, 1994.

Powel MJD. Radial Basis Function for Multivariables interpolation : a review. IMA Conference on algoriths for the approximation of functions and data. RMCS Shrivenham, UK, 1985.

Rehman S et Mohandes M. Artificial neural network estimation of global solar radiation using air temperature and relative humidity, Energy Policy n°**63**, p 571–576, 2008.

Riviere C, Lauret P, Ramsamy JF, Page Y. A Bayesian neural network approach to estimating the energy equivalent speed. Accident Analysis & Prevention, n° 38(2), p 248–259, 2006.

Rosenblatt F. Principles of Neurodynamics : Perceptrons and the Theory of Brain Mechanisms, New York, Spartan Books, 1962.

Rumelhart DE, Hinton GE et Williams RJ. Learning Internal Representations by Error Propagation, letters to nature, n°323, p 533-536, 1986.

Sharif M, Burn DH. Simulating climate change scenarios using an improved K-nearest neighbor model. Journal of Hydrology, n°325 1-4, p 179-196, 2006.

Sfetsos A, Coonick AH. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. Solar Energy, n°68(2), p 169-178, 2000.

Simon G. Méthodes non linéaires pour série temporelles. Thèse de doctorat Université catholique de Louvain, 2007.

Sozen A, Arcaklioglub E, Ozalpa M et Caglar N. Forecasting based on neural network approach of solar potential in Turkey, Renewable Energy n°30, p 1075–1090, 2005.

Terceiro J, Casl J, Perez M, Serrano GR et al. Time series analysis using Matlab, including a complete Matlab Toolbox. User manuel E^4, téléchargeable sur www.ucm.es/info/icae/e4. 2000

Yakowitz S. Nearest neighbors method for time series analysis. Journal of Time Series Analysis n°8, p 235-247, 1987.

Widrow B et Hoff ME Jr. Adaptative switching circuits. IRE WESCON convention record, part 4, New York: IRE p 96-104, 1960.

Chapitre 3.

Prédiction de séries temporelles de rayonnement global et de production d'énergie photovoltaïque : notre approche

SOMMAIRE

3

1. Introduction

Dans les deux précédents chapitres, nous avons expliqué deux manières différentes d'appréhender le rayonnement global, ou la puissance *PV*. Ainsi ces grandeurs, peuvent être caractérisées, soit par des modèles de connaissance (Chapitre 1), soit par l'analyse des séries temporelles issues de mesures directes ou indirectes de ces phénomènes (Chapitre 2). La revue bibliographique faite dans le chapitre précédent a montré les différents estimateurs utilisables pour la prédiction du rayonnement global ou de la puissance *PV*. Deux d'entre eux semblent les plus dédiés à cette tâche, il s'agit des *RNA* et des processus *ARMA*. Dans ce chapitre, nous allons détailler la méthodologie développée durant cette thèse qui tient compte des deux approches distinctes : le modèle de connaissance et les séries temporelles. Tout d'abord, nous commencerons par donner les détails quant à l'utilisation des *PMC* pour la prédiction des séries temporelles. Ensuite, nous discuterons de l'efficacité de ce prédicteur d'après les travaux bibliographiques recensés. Enfin, nous élargirons et sortirons du cadre du *PMC*, pour annoncer les méthodologies de prédiction dans leurs globalités. Nous détaillerons ainsi la méthode d'optimisation choisie, celle concernant la sélection des variables et enfin les modes de stationnarisation que l'on a utilisés.

2. Le perceptron multicouche et les séries temporelles

La possibilité de créer une machine douée d'intelligence ; à même de reproduire la pensée humaine a toujours fasciné les êtres humains. Avec les avancés et les découvertes scientifiques, cette aspiration est devenue plus concrète et les recherches dans ce sens se sont développées. En 1637, le philosophe mathématicien René Descartes avait cependant prédit

qu'il ne serait jamais possible de faire une machine qui pense comme les humains (préface du traité du monde et de la lumière nommée « le discours de la méthode »). Pourtant, trois siècles plus tard, en 1950, le mathématicien britannique et pionnier de l'informatique, Alan Turing était lui persuadé « qu'un jour il y aurait une machine qui pourrait reproduire l'intelligence humaine dans tous les sens du terme » [Turing 1950]. Le calcul informatique simple ne simule en rien le cerveau humain, il peut certes le dépasser dans les calculs complexes, mais ne peut, en aucun cas, décider ou émettre un avis sur une situation donnée. Pour tenter de dépasser cette limite, des techniques dites d'Intelligence Artificielle (*IA*) ont été développées. Parmi elles, on peut citer les *RNA*, et plus particulièrement les *PMC*. Ce sont ces techniques qui ont souvent été utilisées pour réussir à effectuer la prédiction d'une série temporelle. Cette notion [Kalogirou 2001 ; Mellit et Kalogirou 2008 ; Mellit et al. 2009], dans son sens le plus large indique la capacité d'une machine ou d'un objet à effectuer les mêmes types de fonctions que celles caractérisant la pensée humaine. Dans les faits, l'utilisation de ces techniques pour la réalisation de prédiction de séries temporelles est encore loin d'égaler la pensée humaine. En fait, ces modèles sont confrontés à un ensemble de données jusqu'à ce qu'ils apprennent les modèles qui leur sont présentés [Zarzalejo et al. 2005]. Une fois constitués, de nouvelles conditions initiales peuvent leur être présentées en vue d'une estimation ou d'une prédiction [Balestrassi et al. 2009].

L'estimation des paramètres météorologiques nécessite souvent des codes informatiques numériques (Cf chapitre 1, section 5-2, p 58). Ces algorithmes sont en général très complexes. Ils ont besoin d'une quantité considérable de temps pour donner des prévisions précises [Yessad 2010]. Comme nous le détaillerons par la suite, les *PMC* sont une alternative séduisante à ces modèles sophistiqués. Cependant leur utilisation est spécifique à une localisation et à la prédiction d'un paramètre à la fois.

2.1. Le perceptron multicouche

En toute généralité, les réseaux de neurones artificiels sont des collections de petites unités, ou encore une juxtaposition d'éléments unitaires, fonctionnels et interconnectés [Bofinger et Heilscher 2006 ; Bosch et al. 2008] (Cf chapitre 2, section 3-8). Il existe une multitude d'agencements possibles [Jodouin 1994]. Cependant, dans le cadre des travaux présentés dans ce manuscrit, seul le Perceptron MultiCouche (*PMC*) a été utilisé (Cf chapitre

2, section 4, p 87). Ce choix est en accord avec les résultats bibliographiques, car il s'avère que c'est le type de réseau le plus utilisé en prédiction de séries temporelles [Jain et al. 1996 ; Qi et Zhang 2009 ; Mellit et Kalogirou 2008]. Il s'agit d'un réseau non bouclé formé de neurones formels interconnectés. Certaines applications peuvent nécessiter plusieurs sorties (dans le cas d'une classification à plusieurs classes par exemple). Dans notre cas, presque tous les réseaux utilisés ont une seule sortie : la configuration obtenue est certainement la plus classique. Dans cette architecture, les neurones sont organisés en couches comme le montre la Figure 3-1.

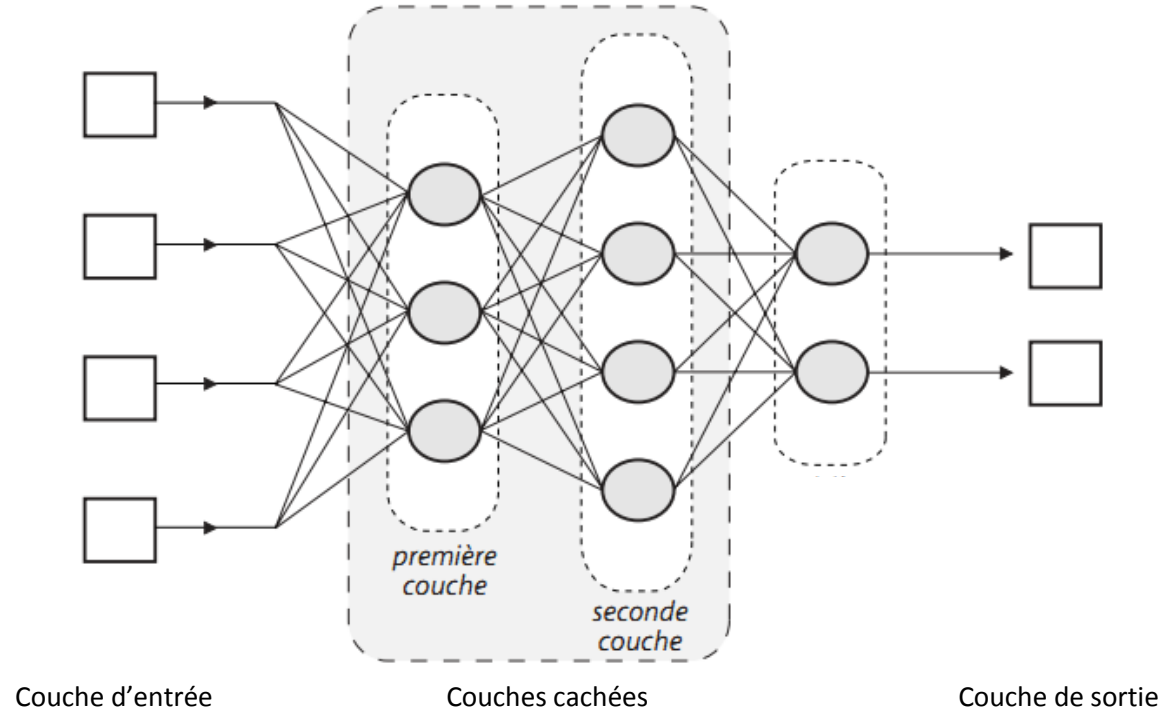


Figure 3-1 : exemple d'architecture *PMC* avec deux couches cachées et deux sorties

La couche intermédiaire située entre les entrées et les sorties est appelée couche cachée, il peut y en avoir plusieurs. Les connexions se réalisent d'une couche à la suivante sans qu'il y ait de connexion entre couches non adjacentes. Dans le cas du *PMC* avec une seule couche cachée, on qualifie également cette architecture « de réseau à deux couches » puisqu'il y a deux couches de poids ajustables : une qui relie les entrées aux neurones cachés et l'autre les neurones cachés aux sorties. Les neurones cachés sont ceux de la couche cachée. Notons que les fonctions d'activations de type sigmoïde (généralement présentes sur la ou les couches

cachées) sont les éléments principaux qui permettent aux *PMC* d'être qualifiés de modèles non-linéaires. Si l'on remplace ces fonctions d'activation sigmoïdes par des fonctions linéaires, on transforme le *PMC* en une régression linéaire. Dans le cadre de la prédiction des séries temporelles, on est alors en présence d'un modèle *AR* « masqué ». Habituellement la couche de sortie est pourvue d'une fonction d'activation linéaire. Si cette architecture est la plus décrite dans la littérature, cela est certainement lié au fait que les *PMC* ont la propriété générale d'être des approximateurs universels parcimonieux. La propriété d'approximation universelle a été initialement démontrée par Cybenko [1989] dans le cas de fonctions d'activation sigmoïdales. Elle fut ensuite généralisée par Hornik [1991], qui montra que cette propriété était plus liée à l'architecture générale du *PMC* qu'aux fonctions d'activation elles-mêmes. Ce théorème peut s'énoncer de la façon suivante :

« Toute fonction bornée suffisamment régulière peut être approchée uniformément, avec une précision arbitraire, dans un domaine fini de l'espace de ses variables, par un réseau de neurones comportant une couche de neurones cachés en nombre fini, possédant tous la même fonction d'activation, et un neurone de sortie linéaire. »

Cette propriété peut certainement expliquer le nombre d'études utilisant cette architecture [Mellit et al. 2009]. Comme le montre ce théorème, le nombre de neurones cachés doit être choisi convenablement pour obtenir la précision voulue. Lorsque l'on cherche à modéliser un processus à partir des données, on s'efforce toujours d'obtenir les résultats les plus satisfaisants avec un nombre minimum de paramètres ajustables : il s'agit du principe de parcimonie. Dans cette optique, Hornik et al. [1994] ont montré que :

« Si le résultat de l'approximation (c'est-à-dire la sortie du réseau de neurones) est une fonction non linéaire des paramètres ajustables, elle est plus parcimonieuse que si elle est une fonction linéaire de ces paramètres. »

De plus, pour des réseaux de neurones à fonction d'activation sigmoïdale, l'erreur commise dans l'approximation varie comme l'inverse du nombre de neurones cachés, et elle est indépendante du nombre de variables de la fonction à approcher [Hornik et al. 1994]. Par conséquent, pour une précision donnée, donc pour un nombre de neurones cachés donné, la quantité de paramètres du réseau est proportionnelle au nombre de variables de cette fonction. Ce résultat s'applique aux réseaux de neurones à fonction d'activation sigmoïdale puisque la sortie de ces neurones n'est pas linéaire par rapports aux poids synaptiques. Cette propriété

montre l'intérêt des réseaux de neurones par rapport à d'autres approximateurs comme les modèles régressifs linéaires dont la sortie est une fonction linéaire des paramètres ajustables : pour un même nombre d'entrées, le nombre de paramètres ajustables à déterminer est plus faible pour un réseau de neurones que pour un polynôme.

Une fois l'architecture d'un réseau de neurones choisie, il est nécessaire de déterminer les valeurs des poids et des biais des couches cachées et de sortie. Cette étape se nomme l'apprentissage : elle permet au réseau d'être apte à simuler correctement et de s'approcher autant que possible de l'objectif fixé. La section suivante présente ainsi comment s'utilisent les *PMC* dans le cadre de l'apprentissage de séries temporelles.

2.2. Le perceptron multicouche et la prédiction des séries temporelles

Nous venons de voir que les *PMC* sont des approximateurs universels. Ils peuvent identifier un processus, le mémoriser puis le restituer fidèlement. Pour ce faire, ils ne se contentent pas d'apprendre par cœur les données d'apprentissage. En fait, ils s'imprègnent des caractéristiques intrinsèques des signaux. Grâce à cette caractéristique ils peuvent aussi être utilisés pour l'estimation, la modélisation et la prédiction de séries temporelles. La méthodologie d'utilisation du *PMC* pour une étude de prédiction de séries temporelles de rayonnement global est présentée sur la Figure 3-2. Les notations sont celles utilisées lors de la description du neurone formel dans le chapitre 2 (section 3-8, p 82). Dans ce cas, on voit que la sortie correspond à la prédiction de la série temporelle au temps *t* ($\hat{x}_t$). On notera que la méthodologie de la prédiction dans le cas de la puissance *PV*, est similaire à celle présentée ici.

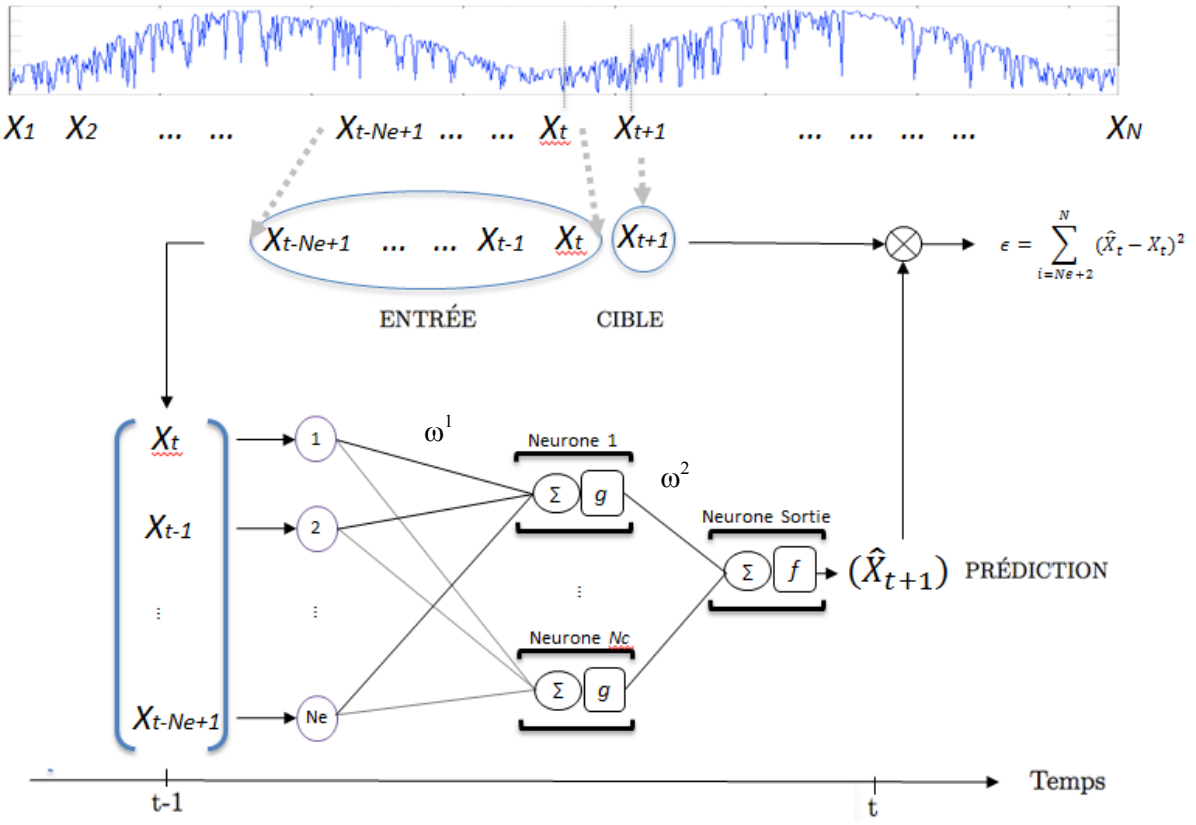


Figure 3-2 : construction d'un *PMC* utilisable dans le cas de prédiction de série temporelle de rayonnement (*Ne* neurones d'entrée, *Nc* neurones cachés, ε erreur de prédiction ou fonction coût)

Cette forme développée du réseau peut se résumer en une expression analytique. En effet, si l'on considère les différentes sorties de chaque neurone caché (y_i), la prédiction du rayonnement global est égale à :

$$\hat{X}_{t+1} = f(\sum_{i=1}^{Nc} y_i \, \omega_i^2 + b^2) \qquad (3\text{-}1)$$

En se référant à la définition de y_i de la section 3-8 du chapitre 2 (p 84), on obtient :

$$y_i = g(\sum_{j=1}^{Ne} X_{t-j+1} \, \omega_{ij}^1 + b_i^1) \qquad (3\text{-}2)$$

Sous cette forme développée, on s'aperçoit que le *PMC* n'est autre qu'un modèle de régression non linéaire des valeurs antérieures de rayonnement global. L'utilisation d'un *PMC* avec un seul neurone de sortie et une fonction d'activation de sortie de type linéaire : $y = x$ (type d'architecture fréquemment utilisé) permet de faire une régression comme décrit dans l'Equation 3-3.

$$\hat{X}_{t+1} = \sum_{i=1}^{Nc}(g(\sum_{j=1}^{Ne} X_{t-j+1}\,\omega_{ij}^{1} + b_i^1))\,\omega_i^2 + b^2 \qquad (3\text{-}3)$$

Où *Nc* est le nombre de neurones cachés, ω et *b* sont les poids et biais du réseau, *g* la fonction de transfert et *Ne* le nombre de neurones d'entrée du réseau. Pour un tel réseau, durant l'apprentissage, le nombre de paramètres à estimer est de $N_e.(N_c+2)+1$. Il faut être vigilent à ne pas trop augmenter le nombre de neurones cachés ou d'entrée sous peine de trop complexifier le modèle. Cet effet peut, dans certain cas, dégrader la qualité de la prédiction. Par exemple pour 10 neurones d'entrée et 10 neurones cachés, il y a 121 coefficients de régressions à estimer au cours de l'apprentissage. C'est une étape importante de la prédiction, elle consiste à trouver des valeurs adéquates des poids et des biais. Suivant les conditions initiales choisies (*Nc, g, Ne, etc*.), l'optimisation aboutira à des résultats potentiellement très différents [Brockwell et Davis 1991]. Elle s'effectue grâce à la minimisation d'une fonction, appelée « fonction de coût ». Elle est calculée à partir des exemples de la base d'apprentissage et de leurs prédictions issues de la sortie du réseau de neurones. En règle générale, la fonction utilisée est la fonction erreur qui peut se calculer avec les formules du *MSE*, du *MAE*, du *MAPE*, etc. (chapitre 2, section 4-1, p 87). Il existe là aussi de nombreux algorithmes qui permettent d'effectuer un apprentissage supervisé. Durant cette thèse, nous avons toujours utilisé l'algorithme de Levenberg-Marquardt (*LM*) [Hu etHwang 2002 ; Costa et al. 2007]. Contrairement à l'algorithme d'optimisation de type retropropagation du gradient, c'est une méthode dite de second ordre. La méthode de *LM* est un standard pour l'optimisation de l'erreur quadratique due à ses propriétés de convergence rapide et à sa robustesse. De plus, après quelques tests, il est aisé de remarquer que c'est celle qui allie le mieux, pour notre type d'étude, rapidité et qualité. Son principe est basé sur une descente de gradient itérative de la famille quasi-Newton. Elle s'appuie sur les techniques des moindres carrés non-linéaires et de l'algorithme de Gauss-Newton à voisinage restreint, conférant une garantie rapide de la convergence vers un minimum [Hu etHwang 2002]. La méthode de *LM* se déduit de l'Equation 3-4, concernant l'optimisation des poids (ω) entre le pas d'apprentissage k et le pas k+1 d'un *PMC*.

$$\Delta\omega = (\omega)^{k+1} - (\omega)^k = \pm[J^T(\omega)^k J(\omega)^k + \mu^k I]^{-1} J^T(\omega)^k e(\omega)^k \qquad (3\text{-}4)$$

Avec $J(\omega)$ la matrice jacobienne (Equation 3-5 pour *N* éléments d'apprentissage), $J^T(\omega)$ sa transposée et $e(\omega)$ l'erreur entre la simulation et la mesure (Equation 3-6 pour N éléments d'apprentissage).

$$J(\omega) = \begin{pmatrix} \frac{\partial e_1(\omega)}{\partial \omega_{11}^1} & \cdots & \frac{\partial e_1(\omega)}{\partial \omega_{Nc}^2} \\ \vdots & \ddots & \vdots \\ \frac{\partial e_N(\omega)}{\partial \omega_{11}^1} & \cdots & \frac{\partial e_N(\omega)}{\partial \omega_{Nc}^2} \end{pmatrix} \quad (3\text{-}5)$$

$$e(\omega) = \begin{pmatrix} e_1(\omega) \\ \vdots \\ e_N(\omega) \end{pmatrix} = \begin{pmatrix} X_1 - \hat{X}_1 \\ \vdots \\ X_N - \hat{X}_N \end{pmatrix} \quad (3\text{-}6)$$

Pour de petites valeurs de pas (μ^k) la méthode de *LM* s'approche de celle de Newton. Inversement pour de grandes valeurs, l'algorithme est équivalent à l'application de la règle du gradient simple. On constate que l'inversion du produit des deux matrices jacobiennes (habituellement nommée la hessienne) nécessaire à l'utilisation de cet algorithme, dépend directement du nombre de paramètres à optimiser. En pratique, des simplifications sont souvent nécessaires pour garantir des temps de calcul convenables [Oussar 1998]. Une itération totale de l'algorithme représente une époque d'apprentissage.

Nous venons d'expliciter les caractéristiques fondamentales des *PMC*. Ces bases vont nous permettre de mieux interpréter les conclusions relatives aux travaux de recherche en rapport avec la prédiction des séries temporelles et l'utilisation des *PMC*. Ces conclusions font suite dans ce rapport.

2.3. Efficacité des perceptrons multicouches pour la prédiction du rayonnement global et de la puissance photovoltaïque

Dans ce paragraphe, nous allons présenter une revue de la littérature sur les prédictions de rayonnement global et de puissance *PV*. Une utilisation optimale de l'énergie renouvelable nécessite une bonne caractérisation et un bon potentiel prédictif en vue de dimensionner les détecteurs ou d'estimer le potentiel énergétique des centrales électriques [Cao et Cao 2006 ; Chaouachi et al. 2009]. En termes de prédiction, les fournisseurs d'électricité sont intéressés par divers horizons afin de pouvoir estimer les économies de combustible fossile, gérer les centrales installées et augmenter l'intégration de systèmes d'énergies renouvelables sur le réseau électrique (section 3-1 de l'introduction). De plus, la prédiction de la production photovoltaïque passe souvent par la caractérisation de la ressource solaire. Concernant la prédiction à horizon du rayonnement solaire, on peut citer les travaux de Mellit [Mellit et

Kalogirou 2008 ; Mellit et al 2009] dans lesquels il est possible de retrouver une grande partie de tout ce qui relève du couplage des *RNA* avec le rayonnement global. Outre ces travaux, il en existe d'autres en rapport avec la prédiction de données météorologiques telles que le rayonnement solaire [Kemmoku 1999 ; Kalogirou 2001 ; Hocaoglu et al. 2008 ; Hocaoglu 2011]. Les réseaux de neurones ont été étudiés sur de nombreux sites et ces travaux de recherche ont montré la capacité de ces techniques à prévoir correctement les séries chronologiques de données météorologiques [Hontoria et al. 2002]. Comparés à des algorithmes classiques basés sur des modèles linéaires (comme *ARMA*), les *RNA* offrent une alternative intéressante en fournissant des modèles non-linéaires paramétriques (chapitre 2, section 4-2, p 92). De plus, si l'on se réfère à ces études, il convient d'utiliser principalement le *PMC*, lequel est le *RNA* le plus utilisé dans le domaine des énergies renouvelables [Mellit et Kalogirou 2008 ; Mellit et al 2009] et de manière générale dans la prévision des séries chronologiques [Jain et al. 1996 ; Crone 2005 ; Hwang et al. 2006]. Si l'on se réfère à des travaux plus anciens, on peut citer Negnevitsky [1995] qui a eu l'idée de combiner un système expert et un *RNA*. La finalité globale de l'étude n'était pas la détermination du rayonnement global, mais l'évaluation de l'indice de protection thermique. Cependant, le *RNA* a été utilisé pour déterminer l'irradiance solaire horaire en fonction des conditions astronomiques et météorologiques. En 1998, Al-Awi et Al-Hinai [1998] ont utilisé un *PMC* pour prédire le rayonnement solaire dans les zones non couvertes par des instruments de mesure. Les données d'entrée du réseau sont l'emplacement géographique, le mois, la pression moyenne, la température moyenne, la pression de vapeur moyenne, l'humidité relative moyenne, la vitesse moyenne du vent et la durée moyenne d'ensoleillement. Cette étude montre que le modèle *PMC* prédit le rayonnement solaire avec une précision de 93 % et un *MAPE* = 7,3 %. Il ne faut cependant pas confondre ce genre d'étude (modélisation par régression multivariée), avec la prédiction à horizon d'une série chronologique. En effet, les *PMC* sont assez régulièrement utilisés pour leur aspect « approximateur universel », capables de faire une régression non linéaire sur des paramètres d'entrée bien choisis. En 1998, apparaissent des travaux qui utilisent les *PMC* comme prédicteurs non-exclusifs à un site [Mohandes et al. 1998]. La méthodologie permet de prédire le rayonnement global sur une zone très vaste. Les auteurs ont utilisé les données de 41 stations de collecte en Arabie Saoudite. Les valeurs d'entrée du réseau sont la latitude, la longitude, l'altitude et la durée d'ensoleillement. Le *MAPE* moyen pour les stations étudiées est de 12,6 % indiquant la viabilité de cette approche pour la modélisation spatiale du rayonnement solaire. Un an plus tard, en 1999, pour la première fois

un auteur communique sur la prédiction à horizon de séries temporelles de rayonnement global via *PMC*. C'est Kemmoku [1999] qui utilise une méthode basée sur les *PMC* pour prédire l'ensoleillement de la journée suivante. Les données d'entrée du réseau sont la pression atmosphérique moyenne prévue par un autre réseau, et différentes données météorologiques de la journée précédente. Les résultats obtenus montrent une erreur de prévision (*MAPE*) de 18,5 % en été et 21,8 % en hiver. Le Tableau 3-1 dresse l'historique des années 2000 concernant les méthodes *RNA* appliquées à la prédiction du rayonnement global. Pour un historique sur les années antérieures à l'année 2000, le lecteur intéressé pourra aussi se référer à l'article de Mellit [Mellit et Kalogirou 2008] qui a présenté les premiers travaux de prédiction de rayonnement global avec les *RNA*. On constate que le *PMC* est un bon prédicteur, utilisable pour différentes localisations et différents horizons. Dans tous ces articles, on constate que les erreurs liées aux prédictions (mensuelles, journalières, horaires et minutes) sont en moyenne comprises entre 5 % et 10 %. Cependant, comme nous l'avions remarqué dans la section 4-2 du chapitre 2 (p 92), les méthodologies de prédictions sont si différentes, qu'elles sont difficiles à comparer. De plus, les méthodologies de calcul des erreurs sont hétérogènes : erreur de prédiction sur certains jours échantillonnés ou sur une large période, cumul des prédictions sur une période journalière ou horaire, paramètres d'erreur non usuels. On constate cependant que les *PMC* peuvent être utilisés avec des paramètres exogènes ou des prédicteurs couplés (Markov, ondelettes, etc.). Ces variantes semblent améliorer la précision des modèles. Dans l'article de Mellit et Kalogirou [2008], on constate que 79% des méthodes *d'IA* utilisées en prédiction de données météorologiques sont fondées sur une approche connexionniste (*RNA*). On peut aussi citer l'utilisation de logique floue (5 %), de réseaux de type hybride *ANFIS* (Adaptative Neuro-Fuzzy Inferences Systems ; 5 %), les réseaux couplant la décomposition en ondelettes et les *RNA* (8 %) et le mix *RNA* / chaîne de Markov (3 %). En somme, l'utilisation des *RNA*, et plus particulièrement du *PMC*, représente une large majorité des travaux de recherches. Il s'agit, et de loin, de la technique la plus usitée. Les autres méthodes ne sont utilisées que très sporadiquement.

Historiquement, en prédiction de rayonnement global au sol, « le concurrent direct » des techniques *RNA* est le modèle précurseur *ARMA* qui n'est pas une technique *d'IA*. Ainsi, en remontant en 1990, Akinoglu et Ecevit [1990] avaient déjà développé des modèles statistiques adéquats pour le rayonnement solaire horaire (*ARMA*).

Auteurs	Sujet	Localisation	Horizon	Erreur	Remarques
Almonacid [2010]	Estimation de l'énergie *PV*	Espagne	mensuel	*MAPE* = 7,3 %	*Le PMC* est beaucoup plus performant que les modèles classiques (diode, Osterwald, etc)
Behrang et al. [2010]	Modélisation du rayonnement global par différents *RNA*	Iran	j+1	*MAPE* = 5,2 %	*Le PMC* avec variables exogènes est très performant
Benghanem et mellit [2010]	Modélisation du rayonnement global avec *RBF*, *PMC* et regression conventionnelle	Arabie Saoudite	j+1	R^2=0,98	*Le RBF* est le modèle le plus performant, de plus la méthodologie est validée sur un système *PV*
Mellit et Pavan [2010]	Prédiction à horizon du rayonnement global avec *RNA*	Italie	h+24	R^2>94 %	L'utilisation de plusieurs *PMC* donne de bons résultats validés sur un cas réel *PV*
Azadeh et al. [2009]	Modélisation du rayonnement global par *RNA*	Iran	mensuel	Précision= 94 % (erreur de 6 %)	*Le PMC* est plus performant que le modèle de Angström
Chaabene et Ben Ammar [2008]	Prédiction à horizon du rayonnement global avec une méthode hybride mélant *PMC* inférence, floue, *ARMA* et filtre de Kallman	Tunisie	h+1 m+5	*nMBE*=0,2 % *nRMSE* < 10 %	Plusieurs horizons de prédiction sont considérés avec cette méthodologie, l'approche est dynamique
Jiang [2008]	Prédiction du rayonnement diffusé avec *PMC*	Chine	mensuel	Précision de 95 %	La méthodologie est validée sur l'ensemble du territoire chinois
Mubiru et Banda [2008]	Modélisation du rayonnement global avec *PMC*	Uganda	j+1	*RMSE* = 107 Wh/m²	*Le PMC* est plus performant que les méthodes empiriques
Bosch et al. [2008]	Modélisation du rayonnement global	Espagne	j+1	*nRMSE* = 6 %	Il est possible d'utiliser les PMC pour la prédiction en zones montagneuse avec une erreur acceptable
Elminir et al. [2007]	Prédiction à horizon du rayonnement diffusé avec *PMC*	Egypte	h+1 j+1	Standart error = 4,2 % Standart error = 9 %	*Le PMC est* plus performant que les modèles de régression
Mellit et al. [2006]	Prédiction du rayonnement global par *PMC* et ondettes	Algérie	j+1	*MAPE* < 6 %	La méthodologie est aussi validée pour la puissance *PV* et pour d'autres données météorologiques
Cao et Cao [2005]	Prédiction à horizon du rayonnement global avec *RNA* récurrent et décomposition en ondelettes	Chine	j+1	*nRMSE* = 8 % (avec ondelette) et 35 % sans ondelette	La décomposition en ondelette améliore énormément la prédiction
Mellit et al. [2005]	Modélisation du rayonnement global avec *PMC* et matrices de transition de Markov	Algérie	j+1	*nRMSE* = 8 %	*Le PMC* est plus performant que *AR*, *ARMA* et chaîne de Markov
Sozen et al. [2004]	Modélisation du rayonnement global avec *PMC*	Turquie	j+1	*MAPE* < 7 %	L'entrainement et le test sont délocalisés, montrant la robustesse de la prédiction par *PMC*
Reddy et Manish [2003]	Modélisation du rayonnement global mensuel et journalier avec *PMC*	Inde	h+1	*MAPE* = 4 %	*Le PMC* est plus performant que tous les modèles de régression testés
Sfetsos et Coonick [2000]	Prédiction à horizon du rayonnement global par *PMC*	Grèce	h+1	RMSE = 27,6 W/m²	L'utilisation de données exogènes améliore la qualité des résultats du *PMC*

Tableau 3-1 : liste non exhaustive des applications *RNA* pour la prédiction du rayonnement global

La même année, Hokoi et al. [1990] ont développé un modèle stochastique du rayonnement solaire horaire pour les mois d'été. Après une série de transformations, les auteurs avaient constaté qu'un *ARMA*(3,3) était le modèle présentant les meilleurs résultats. Si l'on se réfère à l'historique des techniques issues de *l'IA* et utilisées pour la prédiction du rayonnement global [Hontoria et al. 2002 ; Kemmoku et al. 1999 ; Mellit et Kalogirou 2008], on constate que, d'une part, elles ont été développées avec succès pour plusieurs échelles temporelles (mensuelle, journalière, horaire, etc.) et d'autre part les résultats montrent une amélioration vis-à-vis des approches statistiques existantes [Omer Faruk 2010 ; Zarzalejo et al. 2005]. De plus, elles ont aussi permis d'estimer les données météorologiques dans les sites où les instruments de mesure ne sont pas disponibles (régression sur la latitude et la longitude).

Comme nous venons de le voir, le *PMC* est un outil de choix pour la prédiction du rayonnement global. De plus, en regardant les méthodes utilisées par les différents auteurs, il ne semble pas y avoir de consensus concernant l'optimisation de l'architecture. D'après l'analyse des publications présentées ici, le nombre et la nature des données connectées aux neurones d'entrée diffèrent d'une étude à l'autre, supposant une optimisation « site-dépendant ». Concernant l'apprentissage, les études citées précédemment convergent vers le fait qu'il faut stationnariser le rayonnement global pour avoir des résultats optimaux et utiliser l'algorithme *LM* pour augmenter la rapidité de convergence.

Tous les éléments que l'on vient d'énumérer, vont servir par la suite pour construire notre méthodologie de prédiction. La prochaine section va permettre de présenter la méthodologie appliquée tout au long de cette thèse. Nous allons détailler les différents outils qui ont permis de sélectionner ou de valider notre approche ainsi que le type de réseau de neurones utilisé.

3. La méthodologie de prédiction

Comme le montre la section précédente, les *PMC* sont la principale architecture de *RNA* dédiée à la prédiction des séries temporelles de rayonnement global ou de puissance *PV*. L'approche que l'on propose ici, nécessite l'utilisation de ce type de réseau. De nombreux paramètres doivent être optimisés pour utiliser correctement les réseaux de neurones, et plus

particulièrement les perceptrons multicouches. Dans ce chapitre, nous allons détailler ces différentes étapes puis décrire les outils employés pour valider nos choix.

3.1. Optimisation des paramètres du PMC

En début du second chapitre, nous avions détaillé la prédiction à horizon avec le formalisme des séries temporelles [Faraday et Chatfield 1998]. Dans la section précédente, nous avons explicité ce que sont les *PMC* et leurs caractéristiques intrinsèques. Notre méthodologie vise à utiliser le formalisme des séries temporelles dans le cas d'une approche avec *PMC*. La prédiction à horizon opérée avec un *PMC* (une seule couche cachée), se formule avec l'Equation 3-7.

$$X_{t+1} = \hat{X}_{t+1} + \epsilon_{t+1} = \sum_{i=1}^{Nc} \left(g\left(\sum_{j=1}^{Ne} X_{t-j+1}\, \omega_{ij}^{1} + b_i^{1} \right) \right) \omega_i^{2} + b^{2} + \epsilon_{t+1} \qquad (3\text{-}7)$$

Pour chaque type d'étude menée, il est nécessaire de réaliser une optimisation spécifique des paramètres modifiables non dépendants de l'apprentissage (*Nc*, *Ne*, *g*, etc.). En effet, tous ces paramètres seront en théorie à modifier en fonction de la localité, du type de rayonnement considéré (incliné, global, direct, etc.), du nombre de mesures à disposition, de l'horizon considéré, et de la stationnarisation préalable. Ainsi pour chaque problème il est nécessaire d'étudier chaque paramètre séparément pour dégager le réseau « idéal ». Dans le domaine des énergies renouvelables, l'utilisation des *PMC* est assez fréquente (paragraphe précédent), il est donc possible d'utiliser des résultats communs à bons nombres d'études [Bosch et al. 2008 ; Kemmoku 1999 ; Mellit et Kalogirou 2008]. La pratique et ses lectures bibliographiques nous ont amené à étudier les paramètres listés ci-après, qu'il faut optimiser pour chaque type d'étude et qui sont susceptibles de modifier les résultats :

1. Type de réseau ;
2. Nombre de neurones d'entrée ;
3. Prétraitement des données d'entrées (normalisation et stationnarisation) ;
4. Nombre de couches et de neurones cachés ;
5. Choix des fonctions de transfert ;
6. Algorithme d'apprentissage (sans oublier les paramètres propres à chaque algorithme) ;

7. Mode on-line ou batch (on-line = les poids sont ajustés après chaque utilisation d'éléments d'apprentissage ; batch = les poids sont ajustés après le passage de tous les éléments d'apprentissage, soit, après une époque complète) ;
8. Initialisation des poids (au hasard ou fixe) ;
9. Arrêt de l'apprentissage (early stopping, nombre d'époques, seuil d'arrêt, etc.) ;
10. Fonction de coût utilisée durant l'apprentissage ;
11. Fonction de comparaison utilisée durant le test ;
12. Répartition des données (apprentissage, validation et test).

En pratique, il existe tout de même des paramètres qui ne nécessitent pas d'optimisation. Ainsi, dans nos études, nous avons fixé certaines conditions initiales, pour ne pas multiplier les manipulations superflues, et nous concentrer sur celles plus aptes à modifier les résultats de prédiction. Ainsi, d'après les articles bibliographiques le mode « batch » est le seul utilisé (rapidité de convergence et résultats quasi-identiques, de plus avec l'algorithme *LM*, c'est le seul mode utilisable), l'initialisation des poids est aléatoire (meilleure prise en compte des minima locaux), les réseaux utilisés sont uniquement de type *PMC* pour toutes les études et il n'y a qu'une seule couche cachée (en accord avec la littérature [Mellit et Kalogirou 2008 ; Omer Faruk 2010 ; Zervas et al. 2008] et avec des manipulations préalables [Voyant et al. 2011]). De plus, pour chaque étape d'optimisation, nous avons choisi de lancer entre six et huit simulations afin de nous affranchir le plus possible des phénomènes de minima locaux. L'erreur de prédiction relevée est la moyenne des erreurs issues de simulations de *PMC* « éduqués » séparément. Tous les autres paramètres sont étudiés indépendamment pour chaque localisation, horizon et nombre de mesures disponibles. Le mode d'optimisation retenu ici consiste à figer tous les paramètres sauf celui étudié. Cette hypothèse « naïve » implique qu'ils soient tous orthogonaux. L'optimisation consiste donc à trouver parmi l'ensemble des 12 paramètres (neuf si l'on considère les trois fixés initialement d'après la bibliographie) présentés précédemment ($I=\{i_1,i_2,\ldots,i_{12}\}$), les valeurs particulières I_0 ($=\{i_{01},i_{02},\ldots,i_{012}\}$) qui minimisent l'erreur de prédiction (*nRMSE* dans notre cas) comme le montre l'Equation 3-8.

$$I_0 = \underset{I}{argmin}(nRMSE) \xRightarrow[hyp\ naïve]{} \begin{cases} i_{01=} \underset{i_1}{argmin}(nRMSE) \\ \vdots \\ i_{012=} \underset{i_{12}}{argmin}(nRMSE) \end{cases} \qquad (3\text{-}8)$$

Du fait du grand nombre de paramètres à étudier, cette étape d'optimisation nécessite un nombre important d'expérimentations. En ce qui concerne l'interprétation de l'optimisation de chaque élément de *I*, l'analyse graphique peut être suffisante. Si l'on prend l'exemple du nombre de neurones cachés, l'optimisation revient à rechercher la configuration qui induit l'erreur de prédiction la plus faible (e.g. minimum global sur l'intervalle considéré de la fonction coût) comme le montre la Figure 3-3. Dans cet exemple précis, on choisira dix neurones cachés car c'est la valeur qui donne l'erreur de prédiction la plus faible. Notons, que concernant le nombre de neurones cachés, il existe des règles empiriques qui permettent de déterminer leur nombre en fonction du nombre de neurones positionnés sur la couche d'entrée. On peut ainsi prendre un nombre égal à celui de la couche d'entrée [Wierenga et Kluytmans 1994], soit 75 % de celle-ci [Venugopal et Beats. 1994], soit un nombre égal à la racine carrée du produit du nombre de neurones d'entrée et de sortie [Shepard 1990]. Pour un PMC avec 10 neurones d'entrée et 1 de sortie, ces trois règles donnent les résultats respectifs suivants : 10, 7 et 3 neurones cachés. Vu la grande disparité de ces résultats, nous avons choisi de ne pas utiliser ces règles, mais plutôt une analyse exhaustive comme introduite en début de paragraphe. Comme le montre les résultats précédents, en pratique, le nombre de neurones cachés n'est pas supérieur au nombre de neurones d'entrée. Ainsi nous nous sommes fixés un nombre maximum de neurones cachés correspondant au nombre de neurones d'entrée + 10%. Par exemple pour dix neurones d'entrée, on teste donc de 1 à 11 neurones sur la couche cachée. Concernant toujours les neurones cachés, on peut estimer que le fait d'augmenter leur nombre va tendre à améliorer l'erreur de prédiction sur l'espace d'apprentissage (apprentissage par cœur), mais par contre dès que la prédiction sera faite sur un autre intervalle (hors apprentissage), les résultats seront moins performants. En fait, en réduisant le nombre de degrés de liberté laissés au réseau (= diminuer le nombre de neurones cachés), sa capacité d'adaptation sur l'échantillon d'apprentissage diminue mais, sa capacité de généralisation et sa stabilité vont croître. Ce phénomène est aussi celui qui régit la différence engendrée par la taille de l'échantillon d'apprentissage. Si ce dernier est trop petit, le *PMC* risque d'apprendre par cœur la dynamique du système et ainsi avoir une très faible erreur d'apprentissage, mais par contre une erreur de simulation importante. En somme, un faible nombre d'éléments lors de l'apprentissage peut induire une perte de robustesse du *PMC*.

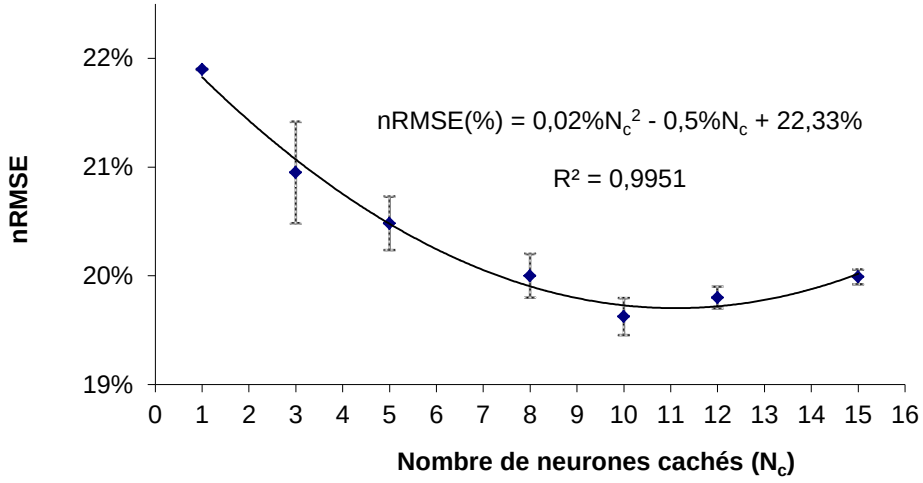


Figure 3-3 : exemple d'optimisation du nombre de neurones positionnés sur la couche cachée d'un *PMC* avec 14 neurones d'entrée. L'intervalle de confiance (95%) est calculé après huit simulations

Outre la nature du réseau considéré (en l'occurrence le *PMC* dans notre cas), l'élément essentiel de l'optimisation est le nombre de neurones d'entrée du réseau, ainsi que les variables à lier à chaque neurone (endogènes ou exogènes). Il est toujours possible d'appliquer une méthode d'analyse exhaustive (toutes les valeurs sont testées) en considérant à chaque fois toutes les valeurs possibles. Cependant, ce type d'étude, souvent très long, pose le problème de la plus grande valeur considérée. Pour le nombre de neurones d'entrée, si l'on applique la méthode exhaustive pour 1-20 neurones, pourquoi négliger le 21^{ième} neurone et ceux qui suivent ? Il semble logique de faire intervenir des critères de choix plus objectifs pour faciliter cette tâche d'optimisation du nombre de neurones d'entrée [Pollock 1999]. La présentation de ces critères fait l'objet du prochain paragraphe.

3.2. La sélection des variables d'entrée

La prédiction des séries temporelles météorologiques avec un *PMC* nécessite de sélectionner des variables de régression (ou d'entrée de réseau) opportunes [Lopez et al. 2005]. De plus, d'après le principe de parcimonie (e.g. de petites modifications peuvent engendrer des effets importants), il ne faut utiliser que les bonnes variables, et uniquement

celles-ci, sous peine de dégrader la qualité des prédictions. La Figure 3-4 décrit le cas général de l'utilisation du *PMC* pour la prédiction à horizon réalisé dans cette thèse. La couche de pré-entrée (ou pré-couche) du réseau est constituée d'un large nombre de neurones. Ces derniers peuvent être soit de même nature que ceux de la sortie (variables endogènes), soit de nature différente (variables exogènes ; pression, nébulosités, prévisions externes, etc.).

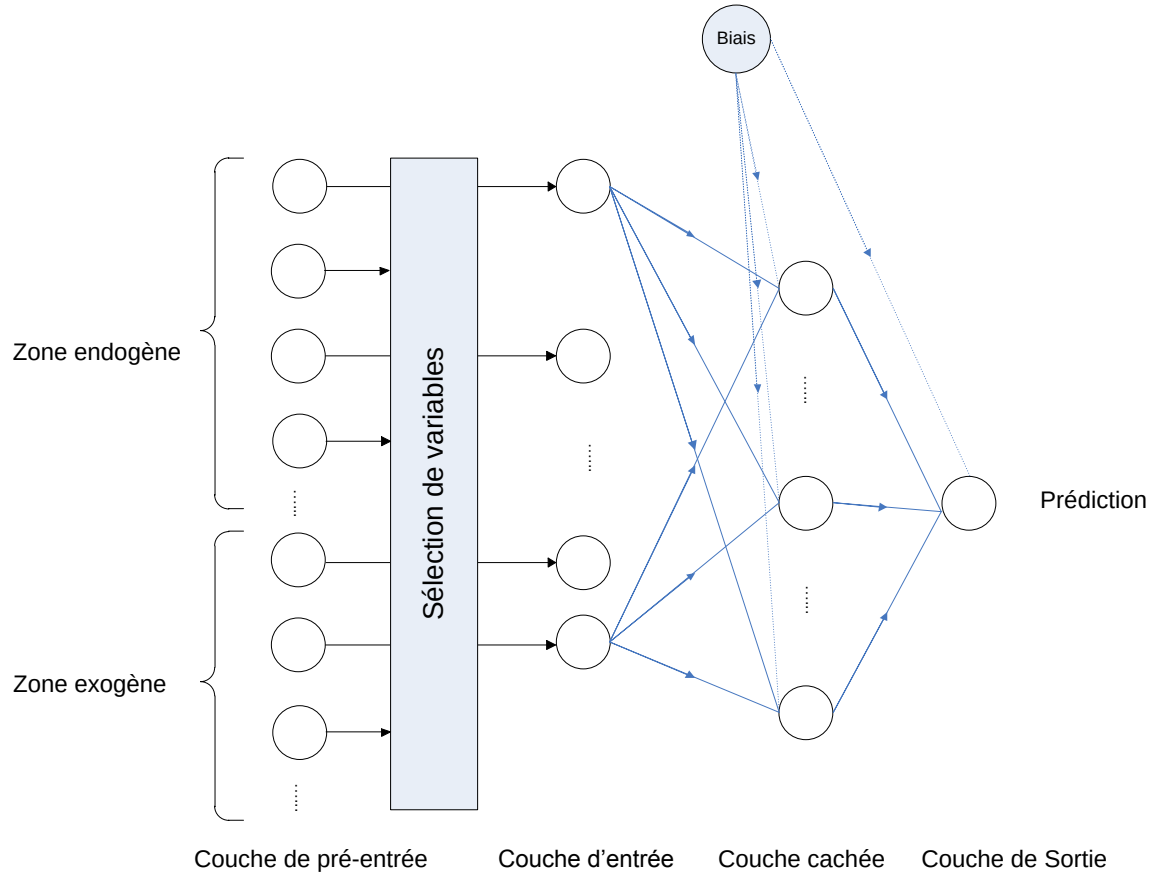


Figure 3-4 : méthodologie de prédiction basée sur l'utilisation d'un PMC avec dénomination des couches

Une multitude de données de pré-entrée doit être potentiellement considérée, mais il n'est pas nécessaire de toutes les intégrer à l'entrée du réseau. Comme on l'a vu précédemment, une augmentation du nombre de neurones d'entrée implique une augmentation du nombre de neurones cachés. Le système se complexifie énormément augmentant les temps de calcul, le nombre de minima locaux, et plus généralement diminuant la qualité des résultats. Cette diminution du nombre de neurones d'entrée est le rôle de l'étape de sélection des variables qui doit statuer sur la pertinence des variables avant de les intégrer à

la couche d'entrée. Plusieurs méthodes existent pour réaliser cette sélection, nous allons détailler celles que nous avons testées. Tout d'abord une méthode dédiée aux variables endogènes (basée sur les autocorrélations), une aux variables exogènes (basée sur les corrélations), puis une plus globale qui permet de considérer les deux types de variables simultanément (basée sur la régression linéaire et une estimation des « moindres carrées »). Seule la prédiction du rayonnement global est considérée, cependant la méthodologie reste également valable pour la puissance *PV*.

a. Utilisation des coefficients d'autocorrélation : cas endogène

Pour les variables endogènes, il est possible d'utiliser la méthode dérivant du principe de modélisation *ARMA* décrite par Box and Jenkins [Hamilton 1994]. Cette approche est la plus utilisée concernant ces modèles linéaires. Elle consiste à utiliser les coefficients de corrélation partielle (*PACF* pour « Partial AutoCorrelation Function » en anglais) qui sont une extension des coefficients de corrélation (*ACF* pour « AutoCorrelation Function » en anglais) (Cf chapitre 2, section 3-7a, p 78). Ils sont utilisés pour identifier le nombre de lags[12] nécessaires au modèle autorégressif [Bourbonnais et Terraza 2008]. L'utilisation des *PACF* permet de sélectionner uniquement en entrée du réseau les lags qui sont corrélés avec les valeurs futures. En pratique, les valeurs inférieures au seuil de significativité montrant une non-corrélation avec *x(t+1)* ne doivent pas être positionnées dans le réseau. Le seuil est déterminé par utilisation de l'intervalle de confiance à 95 % (*IC*) du *PACF* (identique à celui de *l'ACF*) [Sfetsos et Cornick 2000]. Ainsi, si graphiquement une valeur de *PACF* est trop proche (inférieure à l'*IC)* de l'axe des abscisses, il ne sera pas possible de dire que cette valeur est différente de zéro. Pour calculer cet intervalle de confiance sur *n* mesures, on utilise l'Equation 3-9 :

$$IC = \mp 1{,}96.\sqrt{\frac{1}{n}} \qquad (3\text{-}9)$$

Sur la Figure 3-5, nous pouvons observer un exemple de courbe de *PACF* concernant le rayonnement global journalier pour les sites de Bastia et d'Ajaccio (avec un intervalle de confiance à 95%). On voit que pour Bastia le premier lag non significativement différent de

[12] De time-lag signifiant le délai qu'il y a entre deux évènements : l'unité correspond au temps qui sépare deux observations, ainsi un lag 2 pour le cas horaire, signifie 2 heures d'écart entre les deux informations.

zéro correspond au 5ième et au 3ième pour Ajaccio. Ainsi dans l'optique de sélection de variables d'entrée d'un *PMC* et pour la prédiction de $\hat{x}(t+1)$, il faudra deux neurones d'entrée endogènes sur Ajaccio ($x(t)$ et $x(t-1)$) et quatre sur Bastia ($x(t)$, $x(t-1)$, $x(t-2)$ et $x(t-3)$).

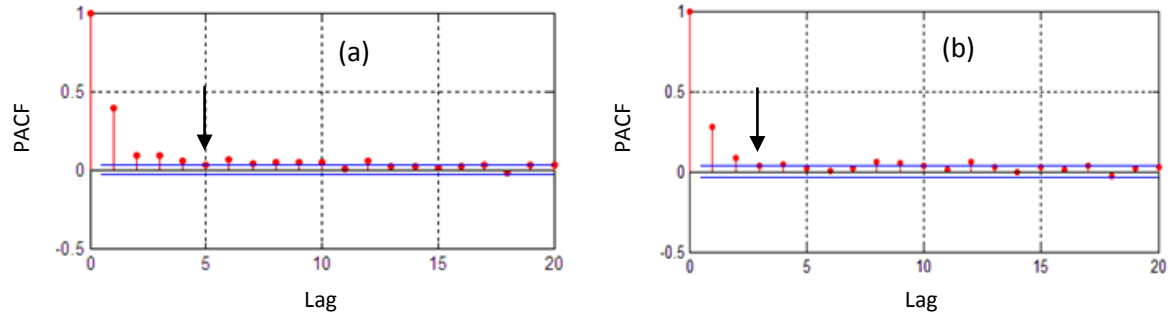


Figure 3-5 : exemple de *PACF* d'une série temporelle journalière de rayonnement global (1998-2007). (a) : Bastia ; (b) : Ajaccio

b. Utilisation des coefficients de corrélation : cas exogène

Pour les variables exogènes la méthodologie est sensiblement équivalente. Elle consiste à calculer la corrélation croisée de Pearson entre les variables exogènes (y) et le rayonnement global (X) au *lag* $t+1$[13]. Le coefficient de corrélation R [Edward 1976] entre ces deux grandeurs se calcule d'après l'Equation 3-10.

$$R = \sum_{k=1}^{N}(X_k - \langle X\rangle)(y_k - \langle y\rangle) / \sqrt{\sum_{k=1}^{N}(X_k - \langle X\rangle)^2 \sum_{k=1}^{N}(y_k - \langle y\rangle)^2} \qquad (3\text{-}10)$$

Généralement, si ce coefficient est compris entre -0,5 et 0,5 le lien entre les deux variables est faible. Pour vérifier la significativité des résultats, il est possible d'utiliser le test de Student en comparant le paramètre t calculé à l'aide de l'Equation 3-11, avec le t de Student empirique (relatif à n évènements ou $n-2$ degrés de liberté).

$$t = \frac{|R|}{\sqrt{\frac{1-R^2}{n-2}}} \qquad (3\text{-}11)$$

En pratique, il n'est pas possible d'utiliser ce critère, car il est trop permissif. En effet, pour 1000 éléments et un seuil de décision (valeur critique) à 5%, la limite de significativité

[13] Dans le cas de l'utilisation de l'indice ciel clair, il conviendra de remplacer le rayonnement global par l'indice ciel clair.

est de 0,06. Cette valeur étant très faible, elle annihile toute sélection. En effet, beaucoup de variables étudiées seraient significativement différentes de zéro comme le montre la Figure 3-6. Dans ce cas précis il y a aurait 5 lags liés à la pression atmosphérique à prendre en compte. Cela serait équivalent quel que soit le paramètre étudié. Le nombre total de variables exogènes étant trop important, il est nécessaire d'affiner la sélection. En fait, suivant l'horizon considéré, nous avons établi une règle de sélection qui revient à considérer une valeur limite de significativité pour *R*. Dans le cas montré dans la Figure 3-6 le seuil fut fixé à 20 %. Une étude de sensibilité a motivé ce choix. En effet, en dessous de ce seuil, la sélection était beaucoup trop permissive ; au-dessus, quasi aucune variable ne passait le test (à partir d'un seuil à 30%). Dans cet exemple, la méthodologie de sélection permettrait de ne considérer que le premier lag exogène.

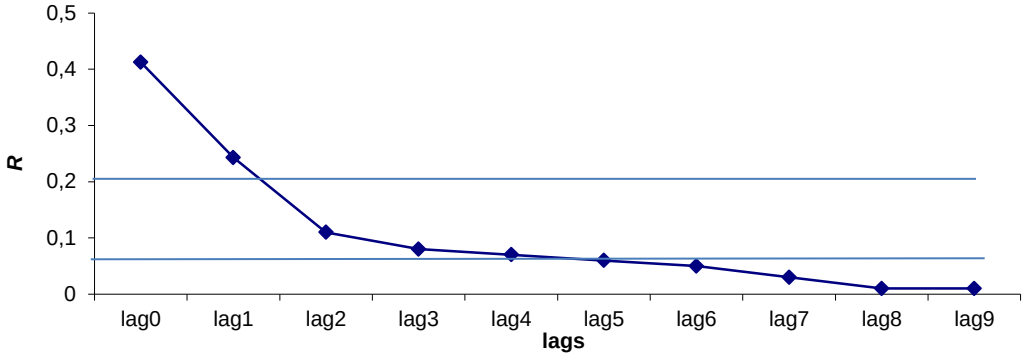


Figure 3-6 : exemple de corrélation entre le rayonnement global et une variable exogène journalière (pression atmosphérique) sur le site d'Ajaccio (1998-2007). Les deux traits pleins représentent les limites des seuils de significativité à 5% et 20%

Il est à noter que dans le cas horaire, l'étude de sensibilité montre des résultats différents, la limite des 20 % n'est plus utilisable mais la méthodologie reste identique. Le détail de la sélection est donné dans la partie résultat (chapitre 4, section 3-1, p 172).

c. Utilisation de la régression linéaire : cas endogène et exogène

Pour lever la part de subjectivité décrite plus haut sur le seuil de significativité à appréhender, nous avons développé une autre méthode de sélection de variables basée sur un modèle de régression linéaire. Dans ce dernier, le passage entre la pré-couche et la couche d'entrée se fait à l'aide d'un système externe aux réseaux de neurones. Dans la pré-couche,

sont référencées toutes les variables disponibles. Il convient tout de même de ne pas trop en ajouter car au-delà d'un certain nombre de lag, l'intérêt à utiliser toutes les variables se restreint (si la série est correctement stationnarisée, le *PACF* tend vers 0 quand le nombre de lag augmente). Les *PACF* ou coefficients d'autocorrélation peuvent être un bon moyen pour déterminer le nombre total de lags de type endogène à ajouter dans la pré-couche. Concernant les lags exogènes positionnés sur cette pré-couche, le coefficient de corrélation de Pearson peut aussi donner une bonne estimation du maximum de neurones de type exogène à placer dans la couche cachée. Ainsi les deux méthodes décrites dans les paragraphes précédents (section 3-2a et 3-2b) permettent de construire le modèle présenté ici. La méthodologie de sélection entre la pré-couche et cette couche, consiste à appliquer un test de Student sur des coefficients de régression afin de vérifier s'ils sont significativement différents de zéro. La régression se formule pour N_e variables endogènes (X) et Me variables exogènes (E) par :

$$X_{t+1} = \hat{X}_{t+1} + \epsilon_{t+1} = \sum_{j=1}^{Ne} \omega_j^{RL} X_{t-j+1} + \sum_{p=1}^{Me} \omega_{Ne+p}^{RL} E_{t-p+1} + b^{RL} + \epsilon_{t+1} \qquad (3\text{-}12)$$

Cette équation peut se synthétiser sous la forme d'un produit de matrice. Ainsi, si Y est un vecteur formé d'éléments X_{t+1} et S' une matrice avec les données endogènes et exogènes, on obtient :

$$Y = S'W^{RL} + \epsilon \qquad (3\text{-}13)$$

L'écriture développée correspondant à N_T mesures endogènes servant à l'estimation des paramètres de régression est donnée par l'Equation 3-14.

$$\begin{pmatrix} X_{t+1} \\ \vdots \\ X_{t+2-N_T} \end{pmatrix} = \qquad (3\text{-}14)$$

$$\begin{bmatrix} 1 & X_t & \cdots & X_{t-Ne+1} & E_t & \cdots & E_{t-Me+1} \\ 1 & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{t-N_T+1} & \cdots & X_{t-Ne-N_T+2} & E_{t-N_T+1} & \cdots & E_{t-Me+N_T+2} \end{bmatrix} . \begin{pmatrix} b^{RL} \\ \omega_1^{\mathrm{RL}} \\ \vdots \\ \omega_{Ne+Me}^{\mathrm{RL}} \end{pmatrix}$$

$$+ \begin{pmatrix} \epsilon_{t+1} \\ \vdots \\ \epsilon_{t+2-N_T} \end{pmatrix}$$

La méthode consiste à résoudre cette équation, afin de déterminer les coefficients b^{RL} et ω^{RL}. L'optimisation peut être faite par l'utilisation de l'algorithme des moindres carrés. L'estimateur classique est défini par (S'^{T} est le transposé de la matrice S') :

$$\widehat{W}^{RL} = (S'^{T}S')^{-1}S'^{T}Y \qquad (3\text{-}15)$$

L'étape qui suit concerne la vérification des poids calculés. Il est impératif qu'ils soient différents de zéro. Pour cela nous appliquons le test de Student sur chaque poids. La statistique utilisée est $t_j = \omega_j^{\mathrm{RL}}/\sigma_j$, où σ_j^2 est la variance du paramètre *j* disponible sur la diagonale de la matrice variance-covariance des paramètres estimés qui est calculé par $\langle\epsilon^2\rangle(S'^{T}S')^{-1}$. Le poids n'est pas significativement différent de zéro si $|t_j| > 1{,}96$ (alpha de 5%). Notons que ce test peut être remplacé par un autre plus commode qui consiste à utiliser l'intervalle de confiance des poids considérés qui est défini par $\omega_j^{RL} \in [\omega_j^{RL} - 1{,}96.\sigma_j \ , \ \omega_j^{RL} + 1{,}96.\sigma_j]$. Si les signes des deux bornes de l'intervalle sont différents, cela signifie que la valeur zéro est dans l'intervalle de confiance. On peut dès lors considérer que le poids n'est pas statistiquement différent de zéro. Une variante totalement équivalente revient à calculer le signe du produit des deux bornes représenté par l'Equation 3-16.

$$(\omega_j^{RL} + 1{,}96.\sigma_j).\ (\omega_j^{RL} - 1{,}96.\sigma_j) = ((\omega_j^{RL})^2 - 1{,}96^2\ \sigma_j^2)) \qquad (3\text{-}16)$$

Ainsi, si $sign((\omega_j^{\mathrm{RL}})^2 - 1{,}96^2\ \sigma_j^2))$ est strictement positif alors on considère que le poids est différent de zéro ($\omega_j^{RL} \neq 0$). Les variables associées à un poids ω_j qui n'est pas statistiquement différent de zéro, seront moins corrélées avec le rayonnement global au lag t+1. C'est ce critère de sélection qui sert de base à la transition entre la pré-couche et la couche d'entrée du réseau de neurone. La règle de sélection des neurones d'entrée du *PMC* considérant les poids issus de la régression linéaire est la suivante :

$$si\ \omega_j^{RL} = 0\ alors\ \omega_{i,j}^{1} = 0\ \forall i. \qquad (3\text{-}17)$$

Cette méthode de sélection basée sur la régression et le test de Student cumule trois avantages vis-à-vis des deux autres méthodes présentées précédemment dans ce paragraphe (purement endogène ou exogène) :

- les variables endogènes et exogènes sont traitées simultanément, sous la forme d'une seule manipulation ;

- il est possible de sélectionner des lags profonds sans tenir compte des précédents. Les lags d'une même grandeur exogène sont considérés indépendants. Dans les premières méthodes, dès que le PACF ou les coefficients de corrélations étaient nuls, les lags plus profonds étaient supprimés ;

- il n'y a plus de règles subjectives concernant la sélection des variables (si ce n'est le nombre maximum de neurones à tester). Le test de Student sélectionne les variables essentielles à la prédiction.

Une fois la sélection de variable effectuée, la couche d'entrée du *PMC* est dotée de variables endogènes et exogènes. Dans la suite du manuscrit un tel réseau prendra la forme canonique suivant : $(Endo^{Ne*}E^{Me*})xN_cx1$ avec *Ne** et *Me** un groupe de nombres compris respectivement entre 0 et *Ne* et 0 et *Me*, qui représentent les lags issus de la sélection de variables entre la pré-couche et la couche d'entrée. Dans le paragraphe suivant, nous allons détailler les processus de stationnarisation des séries temporelles utilisées durant cette thèse, ainsi que les méthodes de validation associées.

3.3. La stationnarité

Comme nous l'avons détaillé dans le chapitre 2, la stationnarisation des variables tient une place importante dans l'analyse et la prévision des séries temporelles [Kim et al. 2004]. Ce résultat, initialement montré pour les méthodes type *ARMA* [Bourbonnais et Terraza 2008] reste également correct pour l'étude et la prédiction avec réseau de neurones [Zhang et Qi 2005]. En fait, tous les réseaux de neurones artificiels peuvent être considérés comme des approximateurs de fonctions continues à support compact[14]. En outre, les *PMC* standards, avec au moins une couche cachée, sont asymptotiquement stationnaires (e.g. avec les notations utilisées dans l'équation 2-1 du chapitre 2 et de la section 2-1 p 68 ; $lim_n f_n = f$). Ils ne sont pas caractérisés par un comportement divergent ou bien par une augmentation de la

[14] Si au moins une couche est dotée de fonction de transfert saturante (g), alors si $f(x)$ est la sortie du réseau ($=\sum_{i=1}^{Nc} \omega_i^2 . g\left(\sum_j^{Ne} \omega_{ij}^1 x\right)$), une fonction h est définie sur un espace compact $h \in C(I^n)$ et $\epsilon > 0$ alors $|f(x) - h(x)| < \epsilon$ sur I^n.

variance avec le temps [Hornik et al. 1994]. En pratique, pour pouvoir utiliser un *PMC*, il faut donc que les données étudiées soient stationnaires [Ito 1991 a et b ; Ito 1992], ou qu'elles soient régies par une variation lente. Toutefois, un réseau *PMC* peut tout à fait simuler un processus non stationnaire, si et seulement si, cela est réalisé sur un intervalle de temps fini. Ainsi, la prédiction hors échantillon d'apprentissage et les performances de ces systèmes seront médiocres car le réseau ne peut pas saisir certaines caractéristiques importantes du processus étudié. Sans prétraitement, le *PMC* (tout comme *ARMA*) peut être inapplicable à beaucoup de problèmes de non-stationnarité rencontrés en pratique [Pollock 1999]. Une des solutions pour surmonter cet obstacle consiste à transformer la chronique initiale en une série qui soit stationnaire (« faiblement » ou si possible « strictement », chapitre 2, section 3-7, p 78) afin que la modélisation *PMC* puisse être appliquée [Connor et Martin 1994]. Dans le chapitre 1, nous avons vu que le rayonnement global présente des périodicités annuelles et journalières. Ces périodicités rendent la série temporelle non stationnaire. On remarquera par exemple, une tendance ascendante en début d'année et début de journée et tendance descendante en fin d'année et de journée.

a. Méthode de stationnarisation

Les modélisations de type ciel clair ou hors atmosphère parviennent à expliquer clairement les phénomènes périodiques, c'est pour cette raison que de nombreux auteurs [Mellit et Pavan 2010 ; Mubiru 2008 ; Mubiru et Banda 2008] tendent à utiliser ces modélisations pour stationnariser les séries chronologiques de rayonnement global. Cette manipulation permet de modifier l'approche connexionniste de type « boîte noire » en une approche de type « boîte grise ». Notons que dans le cas du rayonnement global, la tendance extra-annuelle peut être considérée comme nulle (la moyenne annuelle est constante). La méthodologie de stationnarisation que nous avons utilisée est la suivante :

1. modéliser le site avec un **modèle ciel clair** (modèle Solis [Mueller et al. 2004] décrit au chapitre 1 section 2, p 40), comme le décrit l'Equation 3-18.

$$H_{gh}(t) = H_0(t).exp\left(\frac{-\tau}{sin^b(h(t))}\right).sin(h(t)) \qquad (3\text{-}18)$$

2. effectuer un **rapport à la tendance** afin de s'affranchir de la périodicité, le résultat obtenu est l'indice ciel clair (Clear Sky Index en anglais ; *CSI*) [Badescu 2008].

$$CSI_t = X_t / H_{gh}(t) \qquad (3\text{-}19)$$

Il est possible d'utiliser une variante de cet indice, en ne considérant que le rayonnement hors atmosphère, on obtient de cette façon l'index de clarté (*k*) [Badescu 2008].

$$k_t = X_t / H_0(t) \qquad (3\text{-}20)$$

En pratique, les deux rapports à la tendance que l'on vient d'expliciter ne permettent pas de s'affranchir totalement de la périodicité. Pour remédier à ce problème nous avons complété notre méthode par l'utilisation des coefficients périodiques [Bourbonnais 1998 ; Bourbonnais et Terraza 2008] comme énoncé ci-après.

3. réaliser une **moyenne mobile** de l'index considéré en appliquant le fait que $2\Delta T$+1 soit au moins égal à la périodicité que l'on veut étudier. Dans le cas horaire et de la périodicité de 9 lags (voir chapitre 1, section 4-1) Δt = 4.

$$MM_t = \langle CSI_t \rangle_{t \in [t-\Delta t, t+\Delta t]} \qquad (3\text{-}21)$$

4. opérer un **nouveau rapport à la tendance** afin de générer les coefficients périodiques C_t :

$$CP_t = CSI_t \,/\, MM_t \qquad (3\text{-}22)$$

5. calculer la moyenne des CP_t (e.g. E[C(t)]), pour obtenir une année type (ou moyenne) de coefficients périodiques CP_t. La nouvelle série des *CSI* (ou des *k*) désaisonnalisées se nomme *CSI** (ou *k**) et s'exprime par le **rapport à la moyenne** des CP_t :

$$CSI_t^* = CSI_t \,/ E[CP_t] \qquad (3\text{-}23)$$

La Figure 3-7 décrit ces différentes étapes de stationnarisation lors de la prédiction du rayonnement global X_t. On constate que ce traitement nécessite une recoloration de la série (appelé post-traitement) afin d'obtenir la prédiction de la grandeur souhaitée.

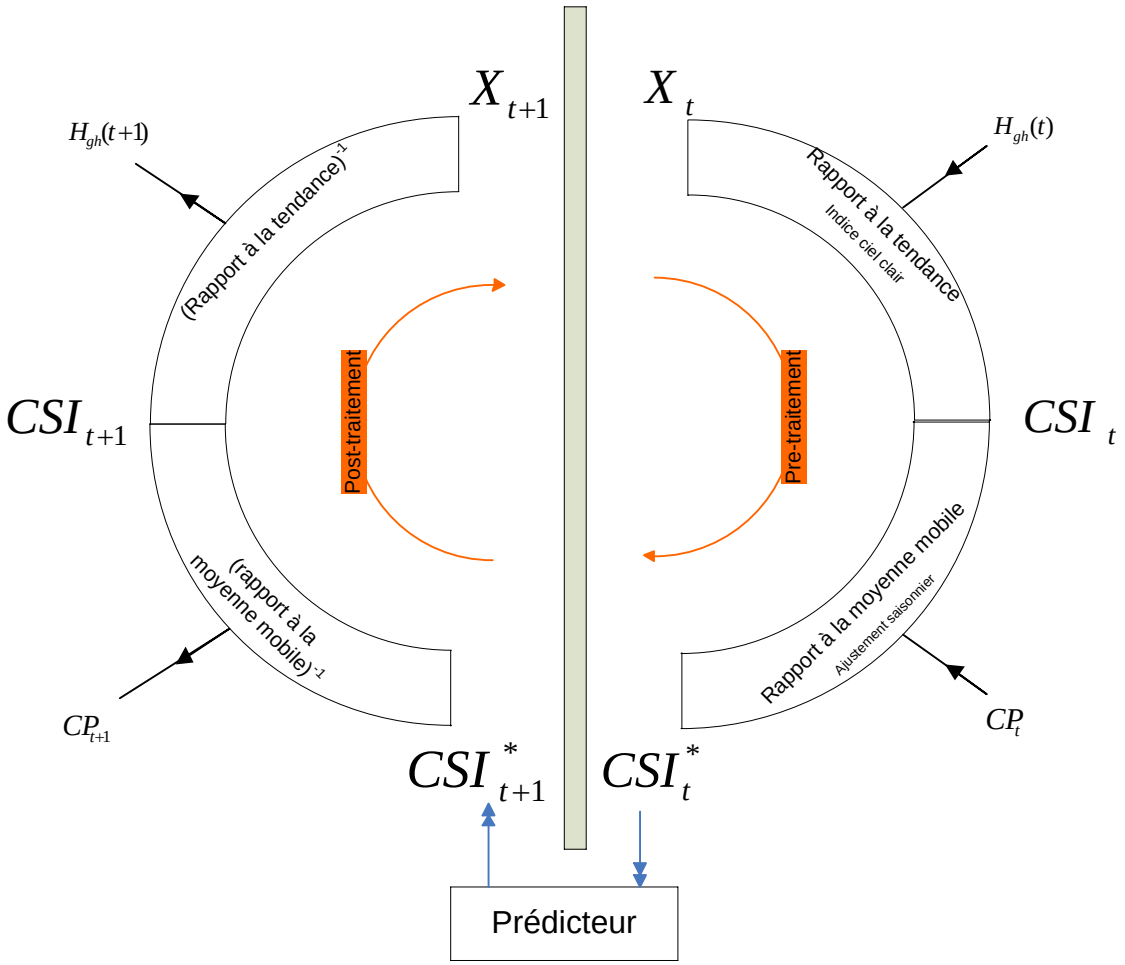


Figure 3-7 : méthodologie globale de prédiction du rayonnement global X_t

b. Validation de la stationnarisation

Après avoir stationnarisé les séries temporelles, il convient de vérifier que le traitement proposé corrige le phénomène de périodicité. Deux approches sont utilisées pour cette validation. La première méthode consiste à utiliser le coefficient de variation défini par :

$$VC_x = \sqrt{E[(x_t - \mu)^2]}/E[x_t] = \sigma/\mu \qquad (3\text{-}24)$$

Cet outil est relativement facile à utiliser et est dédié à la comparaison croisée de deux méthodologies. Cependant il ne permet pas de valider absolument la stationnarisation. Pour le corriger, il existe de nombreuses méthodes de test de stationnarité : racines unitaires, Dickey-Fuller, KPSS, pour en citer quelques-unes [Hamilton 1994 ; Pollock 1999 ; Bourbonnais et Terraza 2008]. La multiplication de ces méthodes et les différences de méthodologie entre

elles, nous ont conduit à choisir un test efficace et plus simple à mettre en œuvre. Il s'agit d'un test très classique basé sur l'analyse de la variance : le test de Fisher [Bourbonnais 2008]. Ce test est dédié aux séries sans tendance extra-saisonnière, comme c'est le cas pour le rayonnement global (moyenne annuelle approximativement constante, voir chapitre 2 section 2-2). Dans le cas d'effets périodiques significatifs, nous considérons que la série temporelle est pourvue d'une composante saisonnière. Le test est construit pour deux types de saisonnalité : journalière et annuelle. Pour construire ce test, il est nécessaire de considérer la série temporelle de la façon suivante (*a* et *b* sont respectivement relatifs à la saisonnalité journalière et annuelle, i la période et j la position de la mesure dans la période) :

$$CSI^*_{i,j} = m_{i,j} + \epsilon_{i,j} \ (residu = \epsilon_{i,j} \rightarrow N(0,\sigma^2)), \qquad (3\text{-}25)$$

$$m_{i,j} = a_{i,j} + b_{i,j} \qquad (3\text{-}26)$$

Le test est construit à partir de l'hypothèse nulle (nommée H0) signifiant qu'aucune influence périodique n'est observée. L'hypothèse alternative (nommée H1) signifie que la série temporelle est périodique. Le principe du test de Fisher appliqué à la stationnarisation est expliqué par la méthodologie suivante :

1. Calcul de la statistique de Fisher empirique $F_c = V_p/V_R$ où V_p et V_R sont les variances sur la période et la variance du résidu ; *p* est le nombre de mesures par période et *N* le nombre de périodes. Les expressions de V_p et V_R sont :

$$V_p = \frac{1}{p-1}.\sum_{j=1}^{p} N.(\langle CSI^*_{i,j}\rangle_i - \langle CSI^*_{i,j}\rangle_{i,j})^2 \qquad (3\text{-}27)$$

$$V_R = \frac{1}{(p-1)(N-1)}.\sum_{i=1}^{N}\sum_{j=1}^{p}(CSI^*_{i,j} - \langle CSI^*_{i,j}\rangle_j - \langle CSI^*_{i,j}\rangle_i + \langle CSI^*_{i,j}\rangle_{i,j})^2 \qquad (3\text{-}28)$$

$$\begin{cases} \langle CSI^*_{i,j}\rangle_j = \left(\frac{1}{p}\right)\sum_{j=1}^{p} CSI^*_{i,j} \\ \langle CSI^*_{i,j}\rangle_i = \left(\frac{1}{N}\right)\sum_{i=1}^{N} CSI^*_{i,j} \\ \langle CSI^*_{i,j}\rangle_{i,j} = \left(\frac{1}{Np}\right)\sum_{i=1}^{N}\sum_{j=1}^{p} CSI^*_{i,j} \end{cases} \qquad (3\text{-}29)$$

N et *p* sont différents suivant si l'on considère la série *a* (journalière) ou la *b* (annuelle).

2. Détermination de la valeur seuil lue dans la table de Fisher [Bourbonnais 2008] et correspondant aux degrés de liberté $v_1=(p-1)$ et $v_2=(N-1).(p-1)$ pour $\alpha = 0{,}05$.

$$F_{v_1;v_2}^{0,05} = F_{seuil} \qquad (3\text{-}30)$$

3. Si $F_c > F_{seuil}$ alors l'hypothèse H0 est rejetée et H1 est acceptée. Si F_c calculé est supérieur à la valeur critique de la distribution (pour alpha=0,05), la série temporelle est décrite comme saisonnière. Par ailleurs, plus le F_c est important plus la saisonnalité est marquée.

Pour certaine manipulation il n'est pas possible d'utiliser les méthodes de stationnarisation et de validation que nous venons d'exposer. C'est notamment le cas, lorsque la tendance ne peut être bien approximée par un modèle de type ciel clair (problème de masque solaire, d'inclinaison des panneaux dépassant les limites des modèles étudié, de localisation, etc.). Ce sera aussi le cas dès que l'on intègre dans un *PMC* des données qui ne sont pas du rayonnement global standard (énergie ou puissance *PV* pré ou post onduleur, etc.). De ces cas précis, il devient impossible de déterminer et d'utiliser une expression analytique de la tendance de la série temporelle. Une alternative existe, il s'agit d'une méthode issue de l'économétrie et qui se nomme la régression sur variables indicatrices [Bourbonnais et Terraza 2008]. Dans le cadre des *PMC*, elle consiste à introduire dans un (ou deux) neurone (s) d'entrée du *PMC*, un indice temporel [Sfetsos et Coonick 2000]. Cet indice permet d'implémenter l'ordre d'apparition des évènements, on peut ainsi introduire un indice allant de 1 à 365 (à une constante de normalisation près) pour référencer le jour considéré, ou encore allant de 1 à 24 (là aussi à une constante près) pour référencer l'heure considérée. Le réseau est laissé libre d'établir et de modéliser la tendance en fonction de ces indices. Contrairement à une approche basée sur l'indice de clarté ou sur l'indice ciel clair, la tendance ne pourra être obtenue que sous forme d'une régression non linéaire sur le jour et l'heure. De plus, le fait d'augmenter les neurones d'entrée entraîne une augmentation du nombre de neurones cachés et donc complexifie le système. Pour ces raisons, nous pensons qu'il convient de n'utiliser cette technique que lorsque les autres techniques standards de stationnarisation ne fonctionnent plus.

Cette section a permis de décrire comment le phénomène de stationnarisation a été intégré durant les manipulations. Nous avons ainsi expliqué ce que signifient les termes *CSI*, *k* *CSI** et k^*. Il nous a semblé important d'en détailler les différentes étapes ainsi que les

vérifications associées. Par la suite, nous allons voir qu'il est possible de coupler les différents prédicteurs afin d'augmenter la qualité des prévisions.

4. Vers une approche hybride

La plupart des études de prédiction ne nécessitent que des approches traditionnelles de modélisations. Cependant bon nombre de séries temporelles de la vie courante sont extrêmement complexes et ne sont pas forcément bien modélisées avec des modèles simples, en particulier lorsqu'une grande précision est requise [Diazrobles et al. 2008 ; Mellit et al 2009 ; Tseng 2002]. Plusieurs études préconisent des modèles hybrides combinant les modèles *ARIMA* et les *RNA* [Li et Su 2010 ; Valenzuela et al. 2008]. Les résultats obtenus ont montré que l'hybridation des modèles produit de meilleures prévisions par comparaison au modèle *ARIMA* seul, ou au modèle *RNA* seul [Ömer Faruk 2010]. En 2003, Zhang [Zhang 2003] a proposé un modèle combinant *ARIMA* et *ANN* en argumentant que l'hybridation permettait d'utiliser les caractéristiques intrinsèques (surtout les points forts) des deux techniques et que le modèle combiné était un moyen efficace d'accroître la qualité de prédiction quel que soit la chronique étudiée. Il existe d'autres couplages possibles, ainsi :

- Cao et Cao [2005] puis Mellit et al [2006] proposèrent l'utilisation conjointe d'ondelettes et de RNA ;

- Sfetsos et Coonick [2000] puis Chaabene et Ben Ammar [2008] étudièrent le couplage inférences floues - ANN, nommée Neuro-Fuzzy Inference Scheme (ANFIS) ;

- Mellit et al [2005] travaillèrent sur l'hybridation avec les chaînes de Markov ;

- Lauret et al [2008] proposèrent l'utilisation des inférences bayésiennes avec les *PMC*.

Ces études permettent de juger de la pertinence de l'hybridation des méthodologies. Elles nous ont conduits à envisager de coupler plusieurs approches de prédiction. Comme il n'est pas raisonnable d'essayer toutes les combinaisons imaginables, nous avons continué avec les prédicteurs présentés plus en amont de ce rapport de thèse. Nous avons essayé de regrouper sous forme d'un modèle hybride les prédictions issues de modèles de la famille

SARIMA et de la famille *PMC*. La Figure 3-8 montre dans le cas du rayonnement global le principe de l'hybridation de ces deux modèles.

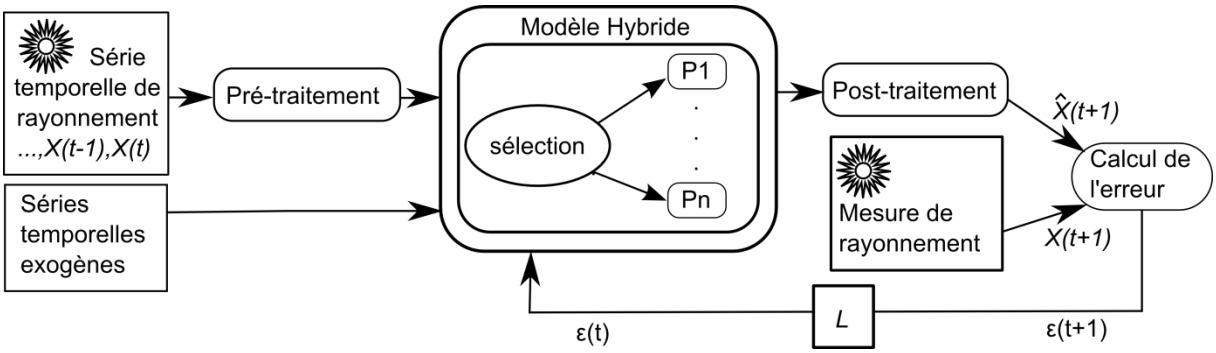


Figure 3-8 : principe de la méthodologie hybride

Le challenge quant à l'utilisation de ce type de prédicteur, est le choix du critère de sélection entre les différents modèles à coupler (P1, P2, .., Pn). Suivant la saison, l'horizon de prédiction, ou le lieu considéré, les résultats peuvent différer totalement. Ainsi, des résultats affichant le fort potentiel d'une méthodologie dans un cas particulier peuvent s'avérer totalement décevants pour une autre configuration. Les manipulations réalisées durant cette thèse ont permis de dégager plusieurs critères de sélection qui ont pu être comparés sur des échantillons tests et qui seront présentés dans le chapitre 4. Les méthodologies d'hybridation testées sont toutes de type sélection entre deux modèles. Cependant il est aussi possible de combiner les résultats des modèles (moyenne des deux, combinaison linéaire, pondération dynamique, etc.). Une étude est actuellement en cours (Stage de M^{lle} Rodler en partenariat scientifique avec l'université de Corse et l'université de la Réunion) pour tester la pondération des résultats issus des modèles *PMC*, *ARMA* et persistance. Cette étude tend à valider l'apport des inférences Bayésiennes dans la sélection et l'hybridation des *PMC* et des modèles *ARMA*. La méthodologie utilisée repose sur le calcul, et l'interprétation de l'évidence (vraisemblance marginale ou intégrale) des *PMC* et du Bayesian Information Criterion (*BIC*) des modèles *ARMA*. En cours d'expérimentation, ces manipulations ne pourront pas être présentées dans le cadre de ce manuscrit.

5. Conclusion

Dans ce chapitre, nous avons posé les fondements de notre approche de prédiction de séries temporelles avec *PMC*. Nous avons axé notre travail sur trois grands thèmes :

- l'optimisation des paramètres du réseau avec la méthodologie fixant un à un les paramètres ajustables ;

- la sélection des variables d'entrée avec la méthode basée sur les *PACF*, les coefficients de corrélation de Pearson et la régression linéaire ;

- la stationnarisation et les nouvelles grandeurs décrites *CSI*, CSI^*, k et k^*.

Peu d'études synthétisent ces trois thématiques au sein d'un seul document, c'est pourquoi il nous a paru primordial de proposer une méthodologie qui ne se limite pas à l'utilisation d'un *PMC* mais plutôt de l'intégrer comme un maillon de la chaîne de prédiction. Le prochain chapitre va décrire les expérimentations que nous avons effectuées durant ce travail doctoral. Elles reprennent pour des horizons de prédiction différents, les méthodes énoncées ici. Les horizons que nous développerons sont : le pas journalier (j+1), le pas horaire (h+1), le pas horaire à horizon 24 heures (h+24) ainsi que le pas 5 min (m+5).

6. Bibliographie

Akinoglu BG et Ecevit A. A further comparison and discussion of sunshine-based models to estimate global solar radiation, Energy n°**15**, p 865–872, 1990.

Al-Alawi SM, Al-Hinai HA. An RNA-based approach for predicting global radiation in locations with no direct measurement instrumentation. Renewable Energy, n°14(1-4), p 199-204, 1998.

Almonacid F, Rus C, Hontoria L, et Muñoz FJ. Characterisation of PV CIS module by artificial neural networks. A comparative study with other methods. Renewable Energy, vol. 35, no. 5, p 973-980, 2010.

Azadeh A, Maghsoudi A, et Sohrabkhani S. An integrated artificial neural networks approach for predicting global radiation. Energy Conversion and Management, vol. 50, no. 6, p 1497-1505, 2009.

Badescu V. Modeling Solar radiation at the earth surface, recent advances. Ed Springer 2008.

Balestrassi P, Popova E, Paiva A, Marangonlima J. Design of experiments on neural network's training for nonlinear time series forecasting. Neurocomputing, n°72(4-6), p 1160-1178, 2009.

Behrang MA, Assareh E, Ghanbarzadeh A, Noghrehabadi AR. The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data. Solar Energy, vol. 84, no. 8, p 1468-1480, 2010.

Benghanem M et Mellit A. Radial Basis Function Network-based prediction of global solar radiation data: Application for sizing of a stand-alone photovoltaic system at Al-Madinah, Saudi Arabia. Energy, vol. 35, no. 9, p 3751-3762, 2010.

Bofinger S, Heilscher G. Solar electricity forecast : approach and first results. european photovoltaic solar energy conference (EU-PVSEC) in Germany, 2006.

Bosch JL, López G, Batlles FJ. Daily solar irradiation estimation over a mountainous area using artificial neural networks. Renewable Energy, vol. 33, no. 7, p 1622-1628, 2008.

Bourbonnais R. Analyse des séries temporelles en économie. Presses Universitaires de France - PUF; 1998.

Bourbonnais R et Terraza M. Analyse des séries temporelles : application à l'économie et à la gestion. 2 éd. Paris: Dunod; 2008.

Brockwell PJ et Davis RA. Time series: theory and methods. 2 éd. New York: Springer-Verlag, 1991.

Cao J et Cao S. Study of forecasting solar irradiance using neural networks with preprocessing sample data by wavelet analysis. Energy, n°31(15), p 3435-3445, 2006.

Cao S et Cao J. Forecast of solar irradiance using recurrent neural networks combined with wavelet analysis. Applied Thermal Engineering, vol. 25, no. 2-3, p 161-172, 2005.

Chaabene M, Ben Ammar M. Neuro-fuzzy dynamic model with Kalman filter to forecast irradiance and temperature for solar energy systems. Renewable Energy, n°33(7), p 1435-1443, 2008.

Chaouachi A, Kamel RM, Ichikawa R, Hayashi H et al. Neural Network Ensemble-Based Solar Power Generation Short-Term Forecasting. World Academy of Science, Engineering and Technology. 2009.

Connor JT, Martin RD. Recurrent Neural Networks and Robust Time Series Prediction. IEEE Transactions on Neural Networks, n°5(2), 1994.

Costa MA, de Pádua Braga A, et Rodrigues de Menezes B. Improving generalization of MLPs with sliding mode control and the Levenberg-Marquardt algorithm. Neurocomputing 70, n°. 7,p 1342-1347, 2007.

Crone SF. Stepwise Selection of Artificial Neural Networks Models for Time Series Prediction. Journal of intelligent systems, n°14, p 23, 2005.

Cybenko G. Approximation by Superposition of a Sigmoidal Function. Mathematics of control, signals and systems, Vol. 2, p 303-314, 1989.

Diazrobles L, Ortega J, Fu J, et al. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. Atmospheric Environment, n°42(35), p 8331-8340, 2008.

Edwards AL. The Correlation Coefficient. Ch. 4 in An Introduction to Linear Regression and Correlation. San Francisco, CA: W. H. Freeman, p 33-46, 1976.

Elminir HK, Azzam YA, Younes FI. Prediction of hourly and daily diffuse fraction using neural network, as compared to linear regression models. Energy, vol. 32, no. 8, p 1513-1523, 2007.

Faraday J, Chatfield C. Times Series Forecasting with Neural Networks: A Case Study. 1998;

Hamilton J. Time series analysis. Princeton N.J. Princeton University Press, 1994.

Hocaoglu F, Gerek O, Kurban M. Hourly solar radiation forecasting using optimal coefficient 2-D linear filters and feed-forward neural networks. Solar Energy, n°8;82(8), p 714-726, 2008.

Hocaoglu FO. Stochastic approach for daily solar radiation modeling. Solar Energy, n°85(2), p 278-287, 2011.

Hokoi S, Matsumoto M, Ihara T. Statistical time series models of solar radiation and outdoor temperature -- Identification of seasonal models by Kalman filter. Energy and Buildings, n°15(3-4), p 373-383, 1990.

Hontoria L, Aguilera J, Zufiria P. Generation of hourly irradiation synthetic series using the neural network multilayer perceptron. Solar Energy, n° 72(5), p 441-446, 2002.

Hornik K. Approximation Capabilities of Multilayer Feedforward Networks. Neural Networks, n°4(2), p251–257, 1991.

Hornik K, Stinchcombe M, White H. Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. Neural computation, vol 6, p 1262-1275, 1994.

Hu YH, Hwang JN. Handbook of neural network signal processing, Ed CRC press, 2002.

Hwang S, Basawa I, Yoon Kim T. Least squares estimation for critical random coefficient first-order autoregressive processes. Statistics & Probability Letters, n°1;76(3), p 310-317, 2006.

Ito Y. Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory. Neural Networks , n°4(3), p 385-394, 1991.

Ito Y. Approximation of functions on a compact set by finite sums of a sigmoid function without scaling. Neural Networks, n°4(6), p 817-826, 1991.

Ito Y. Approximation of continuous functions on Rd by linear combinations of shifted rotations of a sigmoid function with and without scaling. Neural Networks, n°5(1), p 105-115, 1992.

Jain K, Jianchang M, Mohiuddin KM. Artificial neural networks: A tutorial, IEEE Computer, n°29-3, p 31–44, 1996.

Jiang Y. Prediction of monthly mean daily diffuse solar radiation using artificial neural networks and comparison with other empirical models. Energy Policy, vol. 36, no. 10, p 3833-3837, 2008.

Jodouin JF. 2 vol. : Les réseaux de neurones, principes et définitions et Les réseaux neuromimétiques, Paris, Hermès, 1994.

Kalogirou S. Artificial neural networks in renewable energy systems applications: a review. Renewable and Sustainable Energy Reviews, n°12;5(4), p 373-401, 2001.

Kemmoku Y. daily insolation forecasting using a multi-stage neural network. Solar Energy, n°66(3), p 193-199, 1999.

Kim TY, Oh KJ, Kim C, Do JD. Artificial neural networks for non-stationary time series. Neurocomputing, n°61, p 439-447, 2004.

Lauret P, Fock E, Randrianarivony RN, Manicom-Ramsamy JF. Bayesian neural network approach to short time load forecasting. Energy Conversion and Management, vol. 49, no. 5, p 1156-1166, 2008.

Li K et Su H. Forecasting building energy consumption with hybrid genetic algorithm-hierarchical adaptive network-based fuzzy inference system. Energy and Buildings, n° 42(11), p 2070-2076, 2010.

López G, Batlles FJ, Tovar-Pescador J. Selection of input parameters to model direct solar irradiance by using artificial neural networks. Energy, n° 30(9), p 1675-1684, 2005.

Mellit A, Benghanem M, Arab AH, Guessoum A. A simplified model for generating sequences of global solar radiation data for isolated sites: Using artificial neural network and a library of Markov transition matrices approach. Solar Energy, vol. 79, no. 5, p 469-482, 2005.

Melli A, Benghanem M, Kalogirou SA. An adaptive wavelet-network model for forecasting daily total solar-radiation. Applied Energy, vol. 83, no. 7, p. 705-722, 2006.

Mellit A, Kalogirou S. Artificial intelligence techniques for photovoltaic applications: A review. Progress in Energy and Combustion Science n°10;34(5), p 574-632, 2008.

Mellit A, Kalogirou SA, Hontoria L, Shaari S. Artificial intelligence techniques for sizing photovoltaic systems: A review. Renewable and Sustainable Energy Reviews, n°13(2), p 406-419, 2009.

Mellit A, Pavan AM. A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy. Solar Energy, n°Mai;84(5):807-821, 2010.

Mohandes M, Rehman S, Halawani TO. Estimation of global solar radiation using artificial neural networks. Renewable Energy, n° 14(1–4), p 179–184, 1998.

Mubiru J. Predicting total solar irradiation values using artificial neural networks. Renewable Energy, n°10;33(10), p 2329-2332, 2008.

Mubiru J et Banda E. Estimation of monthly average daily global solar irradiation using artificial neural networks. Solar Energy, n° 2;82(2), p 181-187, 2008.

Negnevitsky M. An Expert System Application for Clearing Overloads, International Journal of Power and Energy Systems, n°15, p 9-13, 1995.

Ömer Faruk D. A hybrid neural network and ARIMA model for water quality time series prediction. Engineering Applications of Artificial Intelligence, n° 6;23(4), p 586-594, 2010.

Oussar Y. Réseaux d'ondelettes et réseaux de neurones pour la modélisation statique et dynamique de processus. Thèse de doctorat Paris VI, 1998.

Pollock D. A handbook of time-series analysis, signal processing and dynamics. San Diego Calif; London Academic, 1999.

Qi M et Zhang GP. Trend Time Series Modeling and Forecasting With Neural Networks. IEEE Transactions on Neural Networks 19, n°5, 2009.

Reddy KS et Ranjan M. Solar resource estimation using artificial neural networks and comparison with other correlation models. Energy Conversion and Management, vol. 44, no. 15, p 2519-2530, 2003.

Shepard RN. Neural nets for generalization and classification: comment on Staddon and Reid. Psychol Rev, n°97(4), p 579-580, 1990.

Sfetsos A et Coonick AH. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. Solar Energy, vol. 68, no. 2, p 169-178, 2000.

Sözen A, Arcaklioglu E, Özalp M. Estimation of solar potential in Turkey by artificial neural networks using meteorological and geographical data. Energy Conversion and Management, vol. 45, no. 18-19, p. 3033-3052, 2004.

Tseng F. Combining neural network model with seasonal time series ARIMA model. Technological Forecasting and Social Change, n°69(1), p 71-87, 2002.

Turing Alain. Computing Machinery and intelligence. Republié dans Collected Works of A. M. Turing, volume Mechanical Intelligence, ed. Darrel Ince, 1950.

Valenzuela O, Rojas I, Rojas F, Guillen A et al. Hybridization of intelligent techniques and ARIMA models for time series prediction. Fuzzy Sets and Systems, n°159(7), p 821-845, 2008.

Venugopal V et Baets W. Neural networks and statistical techniques in marketing research: A conceptual comparison. Marketing Intelligence & Planning, n°12 (7), p 30-38, 1994.

Voyant C, Muselli M, Paoli C, Nivet M. Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation. Energy, n°36(1), p 348-359, 2011.

Wierenga B. et Kluytmans J. Neural nets versus marketing models in time series analysis: a simulation study. Proceeding of the 23th annual conference of the European marketing academy, p 1139-1153, Ed Maastricht: University of Limburg, 1994.

Yessad K. Basics about ARPEDE/IFS, *ALADIN* and AROME in the 37 of the cycles of *ARPEGE*/IFS, météo-France/CNRM/GMAP/ALGO, research report, 2010.

Zarzalejo LF, Ramirez L, Polo J. Artificial intelligence techniques applied to hourly global irradiance estimation from satellite-derived cloud index. Energy, n°30(9), p 1685-1697, 2005.

Zervas P, Sarimveis H, Palyvos J, Markatos N. Prediction of daily global solar irradiance on horizontal surfaces based on neural-network techniques. Renewable Energy, n°33(4), p 1796-1803, 2008.

Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, vol. 50, p 159-175, 2003.

Zhang GP et Qi M. Neural network forecasting for seasonal and trend time series. European Journal of Operational Research, n°1;160(2), p 501-514, 2005.

Chapitre 4

Résultats

SOMMAIRE

4

1. Introduction

Les méthodologies présentées dans les trois chapitres précédents nous ont permis d'entreprendre une série de manipulations en vue de la prédiction à horizon de séries temporelles. Dans le chapitre 1, nous avons défini les modèles qui régissent la dynamique solaire et plus particulièrement le rayonnement global au sol, avec les modèles *SOLIS* pour la partie modèle ciel clair, *CLIMED2* pour l'obtention de la fraction de diffusé et Klucher pour l'inclinaison du rayonnement diffusé. Ce chapitre fut aussi l'occasion de présenter les principes des systèmes photovoltaïques conventionnels et les données qui ont été utilisées durant ces manipulations. Dans le chapitre 2, le formalisme des séries temporelles a été introduit et les principaux prédicteurs associés énoncés. Le chapitre 3 a exposé une méthodologie générale de prédiction dont les résultats des manipulations effectuées lors de cette thèse sont l'objet de ce chapitre. Cette méthodologie est relative aux notions d'analyse multivariée, de processus de stationnarisation, de *PMC* et d'approche hybride. Les méthodes présentées sont applicables à différents horizons de prédiction comme j+1, h+24, h+1 ou encore m+5.

Le plan choisi pour exposer les résultats ne suit pas exactement l'ordre chronologique des manipulations. Cependant pour une plus grande lisibilité, nous avons souhaité analyser séparément chaque horizon de prédiction.

La prochaine section de ce chapitre est dédiée à l'étude du cas journalier. Cet horizon est certainement celui pour lequel la prédiction a le moins d'impact tant pour un gestionnaire de réseau que dans le cadre d'une étude de foisonnement. Cependant, ce cas relativement simple (une seule périodicité, cumul annihilant le bruit des données, etc.) nous a permis de

tester différentes approches et surtout différents outils de comparaison et d'évaluation de l'erreur. Une application sur le cas réel du mur photovoltaïque du laboratoire est aussi présentée. Nous parlerons également très brièvement de la possibilité de délocalisation géographique entre les données prédites avec le *PMC* et celles utilisées durant l'apprentissage.

Nous traiterons ensuite le cas de la prédiction horaire, qui est certainement celui qui requiert la plus grande attention, tant à cause des difficultés de mise en œuvre que pour l'intérêt qu'il suscite chez les industriels. Outre l'étude sur les prédicteurs et les variables stationnarisées, nous introduirons l'usage des variables exogènes et décrirons l'apport de la modélisation multivariée sur la prédiction du rayonnement global.

La section suivante décrira aussi le cas horaire, mais sous un horizon 24 heures. La méthodologie, bien que relativement proche du cas précédent, est plus complexe et a nécessité une comparaison entre l'utilisation d'un réseau de *PMC* et d'un *PMC* à sorties multiples.

Le dernier cas présenté sera la prédiction à très court terme de l'horizon 5 minutes. Nous verrons qu'en l'occurrence, ce que nous avons pu réaliser et démontrer pour les horizons plus lointains, ne s'applique pas de façon aussi satisfaisante.

2. Le cas journalier : *j+1*

Bien que la prédiction journalière de rayonnement global ou d'énergie *PV* ne soit pas l'horizon prioritaire pour les gestionnaires de réseau, c'est l'étude de cet horizon que nous avons choisi pour développer et tester nos techniques de prédiction. Il nous est apparu comme le cas le plus simple à résoudre. Dans ce paragraphe, nous allons détailler le cheminement suivi. Cheminement qui a orienté nos choix et qui nous a permis d'affiner notre technique de prédiction pour les autres horizons. Nous présenterons en premier lieu, les résultats relatifs au rayonnement global, puis viendront ceux liés à la puissance *PV*. Tous les résultats liés aux prédicteurs énoncés dans le chapitre 2 seront exposés incluant ceux tenant compte des variables exogènes météorologiques.

2.1. Prédiction de rayonnement global

La série temporelle utilisée concerne le site d'Ajaccio pour les années comprises entre 1971 et 1989. Le but des manipulations initiées dans cette section est de sélectionner les modèles les plus aptes à prédire le rayonnement global pour le jour d'après. Les modèles évalués sont les *SARIMA*, les inférences Bayesiennes, les chaînes de Markov, les *k-NN* et les *PMC* (avec variables endogènes). Chaque technique est soumise à une optimisation spécifique afin de ne retenir que les meilleurs modèles pour cette étude (chapitre 2, section 3). Pour tous ces modèles, aucune stationnarisation n'est effectuée. Après optimisation, les modèles effectifs sont :

- la moyenne, qui ne nécessite pas réellement d'optimisation, le descriptif de ce prédicteur est donné au chapitre 2 (section 3-3, p 75) ;

- la chaîne de Markov (voir chapitre 2, section 3-6, p 77) avec 50 classes et 3 lags (h = 3) de rayonnement. Résultats d'optimisation issus de tests sur la série temporelle considérée. La littérature ne révèle pas de consensus quant à la taille de cette matrice ;

- les inférences Bayesiennes (voir chapitre 2, section 3-5, p 76), avec J = 3 et 50 classes de rayonnement ($k \in [1,50]$). Cette optimisation fait suite à des tests sur des données de rayonnement global sur Ajaccio. Dans la littérature, il ne semble pas y avoir de consensus ;

- *k-NN* avec k = 10 (voir chapitre 2, section 3-4, p 75) ;

- *ARMA* avec p = 8 et q = 0 (voir chapitre 2, section 3-7, p 80) ;

- le *PMC* (voir chapitre 3, section 3, p 84) avec, 8 neurones sur la couche d'entrées (X_t … X_{t-7}) normalisées sur [0,1], 1 couche cachée avec 3 neurones et la fonction d'activation Gaussienne, 5000 époques avec l'algorithme *LM* (le paramètre d'amortissement $\mu = 0{,}5$), tous les autres paramètres sont pris par défaut (Cf l'aide de la toolbox ANN disponible sur http://www.mathworks.fr/help/toolbox/nnet/index.html). En accord, avec ce qui a été expliqué au chapitre 3, la sélection de variables entre la pré-couche et la couche d'entrée est réalisée avec la méthode des *PACF*. Avec cette configuration, la prédiction du rayonnement global *X* est modélisée par :

$$\hat{X}_{t+1} = \sum_{i=1}^{3} e^{(\sum_{j=1}^{8} X_{t-j+1}\omega_{ij}^{1} + b_i^{1})^2}\, \omega_i^2 + b^2 \qquad (4\text{-}1)$$

Pour tous les modèles étudiés, les années 1971-1987 ont servi de base d'apprentissage et les deux années 1988-89 concernent le test de prédiction. La comparaison des prédicteurs est donnée sur le Tableau 4-1 qui matérialise les erreurs de prédiction sur l'échantillon de test. On note que la forme canonique des *PMC* est la suivante : $(Endo^{Ne}E^{Me})xN_cx1$ avec *Ne* et *Me* le nombre de neurones endogènes et exogènes (voir chapitre 3 section 3-2-c).

Méthodologie	**Caractéristiques**	***nRMSE* (%)±Intervalle de confiance à 95%**
Moyenne	-	26,00 ± 0
Chaîne de Markov	ordre 3	25,11 ± 0
Inf. Bayésiennes	ordre 3	25,16 ± 0
k-NN	ordre 10	25,20 ± 0
SARIMA	AR(8)	21,18 ± 0
ANN	$Endo^{1\text{-}8}$x3x1	20,97 ± 0,15

Tableau 4-1 : évaluation de toutes les méthodologies de prédiction étudiées (*nRMSE* et *IC* obtenus après huit simulations). Prédiction faites à horizon j+1, sur Ajaccio pour les deux années 1988 et 1989

Les quatre premières méthodes (moyenne, chaîne de Markov, *k-NN* et inférences Bayésiennes) sont à peu près équivalentes, le *nRMSE* moyen est proche de 25,5 %. La moyenne est, sur l'échantillon testé, le prédicteur le moins performant. Concernant les modèles *ARMA* et *PMC,* ils semblent, dans les conditions du test, présenter une efficacité légèrement supérieure aux autres avec des *nRMSE* inférieures à 22 %. Ce premier résultat, est important, car sans stationnarisation, ou méthode de sélection de variables sophistiquée, ces deux prédicteurs sont équivalents et surpassent les autres approches. Pour cette raison, dans la suite du manuscrit, nous nous concentrerons sur la comparaison de ces deux méthodes de prédiction. Pour ce faire, nous allons essayer d'exploiter toutes les connaissances décrites dans les chapitres précédents afin de les optimiser séparément. Fort de ce qui a déjà été exposé pour accroitre la précision de la prédiction, *ARMA* et *PMC* peuvent, dans un premier temps, bénéficier d'une stationnarisation des données d'entrée. Les coefficients périodiques, l'indice de ciel clair et l'indice de clarté sont les outils que l'on va utiliser (présentés au

chapitre 3 dans la section 3-3, p 123). Concernant les deux années de tests, la Figure 4-1 présente les résultats de l'application de notre méthodologie globale de stationnarisation et son incidence sur l'allure de la série. Avant tout traitement (étape 1), le coefficient de variation (chapitre 3, section 3-3, p 126) de la série temporelle est élevé ($CV \sim 0{,}5$), alors qu'aux étapes 2 et 3, ce coefficient est réduit de moitié ($CV \sim 0{,}3$). Ce coefficient et l'allure des courbes tendent à montrer que l'on a une meilleure stationnarisation à l'issue des étapes 2 et 3.

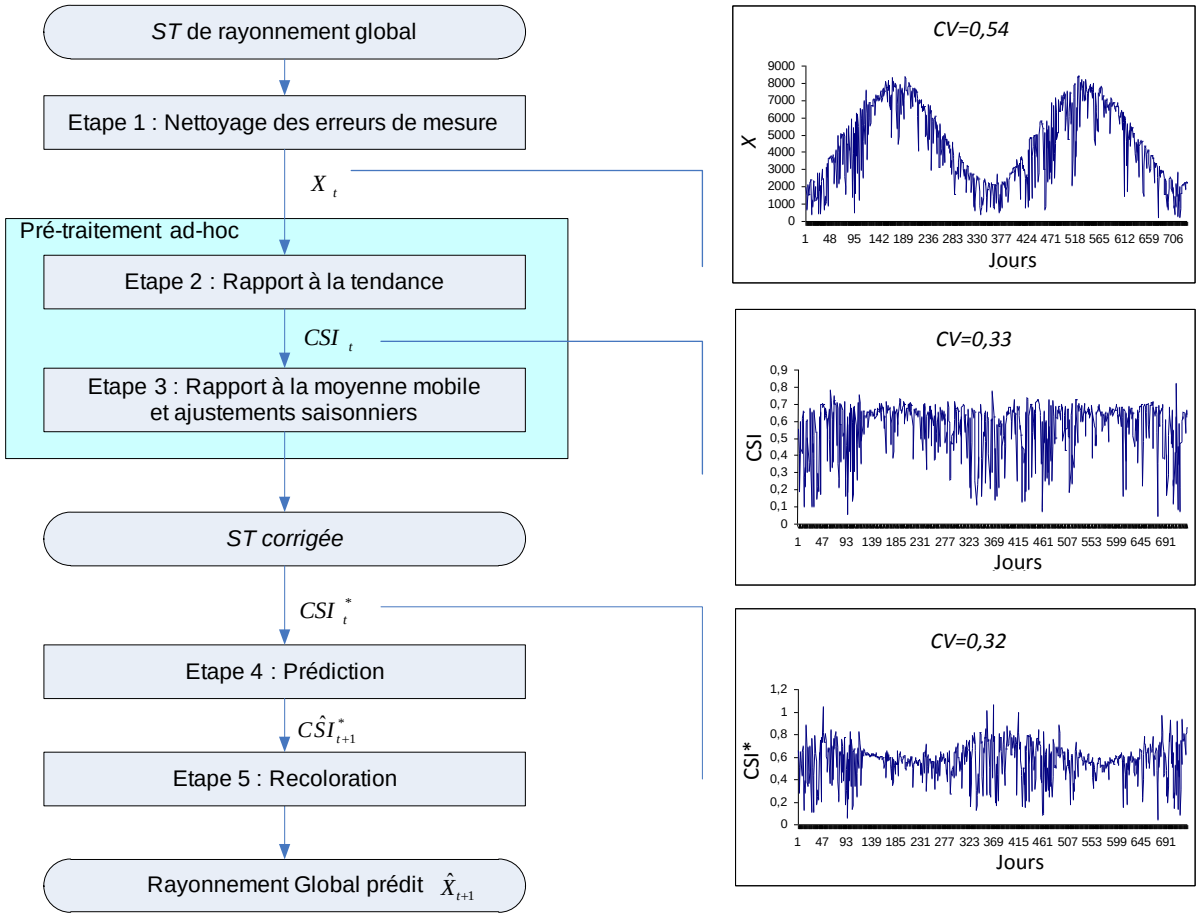


Figure 4-1 : méthodologie de prédiction couplant la stationnarisation et l'utilisation des *PMC*

L'étape 2 peut être effectuée par deux méthodes relativement proches : soit l'utilisation de l'indice ciel clair CSI^*, soit par l'utilisation de l'indice de clarté k^* (Cf chapitre 3, section 3-3). Nous allons expliciter les avantages et inconvenants de chacune. Pour cela, en ne retenant que les modèles *ARMA* et *PMC*, nous avons testé l'impact de ces deux modes de stationnarisation sur l'erreur de prédiction concernant les deux mêmes années que précédemment. Les résultats sont rapportés dans le Tableau 4-2. Le passage par k^* et CSI^*

conduit à une nouvelle phase d'optimisation pour ces deux prédicteurs. En effet, il faut adapter les modèles et les architectures à la nouvelle dynamique du signal. L'optimisation pour *ARMA* conduit à utiliser le modèle *ARMA*(2,2), alors que pour le *PMC*, la configuration reste inchangée.

	Sans stationnarisation	**Méthodologie k^***	**Méthodologie CSI^***
ARMA	*21,18 ± 0%* *AR(8)*	**20,31 ± 0%** *ARMA*(2,2)	20,32 ± 0% *ARMA*(2,2)
PMC	20,97 ±0,15% *endo*$^{1\text{-}8}$x3x*1*	**20,17 ± 0,1%** *Endo*$^{1\text{-}8}$x3x*1*	20,25 ± 0,1% *endo*$^{1\text{-}8}$x3x*1*

Tableau 4-2 : erreur de prédiction pour *ARMA* et *PMC* (*nRMSE* ± *IC95*%). Prédictions à j+1, sur Ajaccio concernant les deux années 1988-1989. Les chiffres en gras représentent les meilleures valeurs de *nRMSE*

Les deux méthodologies de prédiction (*PMC* et *ARMA*) semblent équivalentes pour l'horizon considéré. En effet, que ce soit sans stationnarisation ou en utilisant les deux indices, il est très difficile de les départager. On pourrait effectuer un test de Student pour affiner la comparaison. Cependant l'interprétation des intervalles de confiance, est suffisante, dans ce cas précis, pour aider à conclure. Sans stationnarisation, la limite supérieure de l'intervalle de confiance du *nRMSE* vaut 21,12 % (= 20,97+0,15), cette valeur est inférieure à l'erreur donnée par le prédicteur *ARMA*. A ce titre, on peut considérer qu'il est possible d'effectuer une hiérarchisation entre ces deux prédicteurs. Il semble, dans le cas particulier que l'on vient d'étudier, que le *PMC* soit la méthode de prédiction la plus adéquate, bien que les résultats obtenus avec *ARMA* soient aussi convaincants. Concernant la comparaison entre les deux méthodologies de stationnarisation, il n'est pas possible de conclure, les moyennes de ne sont pas significativement différentes. Cependant, le fait de stationnariser améliore l'erreur de prédiction pour *ARMA et PMC*.

L'utilisation d'autres paramètres d'estimation de l'erreur pourrait certainement permettre de conclure et d'affiner la comparaison des effets engendrés par les deux indices. C'est dans cette optique que dans le Tableau 4-3 sont recensées les erreurs *nRMSE*, *RMSE*, *MAE*, *MBE*, R^2 et la *nRMSE* mensuelle moyenne, toutes liées à la manipulation précédente et définies dans la section 4-1 du chapitre 2, p 87.

	Méthodologie k^*	**Méthodologie CSI^***
nRMSE (%)	**20,17 ± 0,1%**	20,25 ± 0,1%
RMSE (Wh/m².jour)	**998**	1002
MBE (Wh/m².jour)	-104,24	**-88,92**
MAE (Wh/m².jour)	**737,80**	740,90
R^2	0,801	**0,790**
Erreur mensuelle moyenne (%)	**3,9**	4,1

Tableau 4-3 : erreurs annuelles calculée avec le *PMC* sur deux années de prédiction (1988 et 1989). La prédiction concerne Ajaccio à l'horizon j+1, les chiffres en gras pointent les valeurs des coefficients statistiques les plus faibles

Les écarts entre toutes ces erreurs sont minimes et pour la plupart non significatifs. La différence de *RMSE* est inférieure à 5 Wh/(m².jour), ce qui est totalement négligeable d'un point de vue énergétique. De plus, en première approximation et concernant le *nRMSE*, les deux intervalles de confiance à 95 % se chevauchent, d'un point de vue statistique les moyennes sont donc identiques. De plus, le *RMSE* et le *MAE* n'apportent pas d'éléments supplémentaires, par contre le *MBE* est plus important avec l'indice de clarté. Cela signifie que sur le long terme, il y a une sous-estimation du rayonnement global. L'erreur mensuelle moyenne, n'est là qu'à titre indicatif, et représente l'erreur sur le cumul mensuel de la ressource solaire, dans ce cas-là, il y a une différence de 0,2 % en faveur de l'indice de clarté. Cet écart ne permet pas de départager les deux modes de stationnarisation. En fait, plus la moyenne est calculée sur un intervalle de temps important, plus l'erreur nominale diminue. En effet, elle passe de 20 % pour l'erreur journalière à 4 % pour l'erreur sur le cumul mensuel.

Le potentiel solaire n'étant pas le même tout au long de l'année, la prochaine étape sera de vérifier saison par saison l'évolution des erreurs les plus pertinentes (*MSE*, *nRMSE* et *MBE*). Le prédicteur correspond à celui utilisé dans la manipulation précédente ($Endo^{1\text{-}8}$x3x*1*) avec l'utilisation de l'indice de clarté. La Figure 4-2 représente la répartition de ces erreurs. On constate que les mois d'été sont ceux où les prédictions sont les plus faciles (*nRMSE* ~ 14

% et *RMSE* ~ 850 Wh/m²). Le printemps semble une période assez difficile à appréhender. Même si le *nRMSE* est faible (20%), la valeur du *RMSE* est très importante (~1350 Wh/m²), induisant souvent une grosse sous-estimation de la prédiction (*MBE* négatif) pouvant être, sur le long terme, dommageable pour un gestionnaire de réseau.

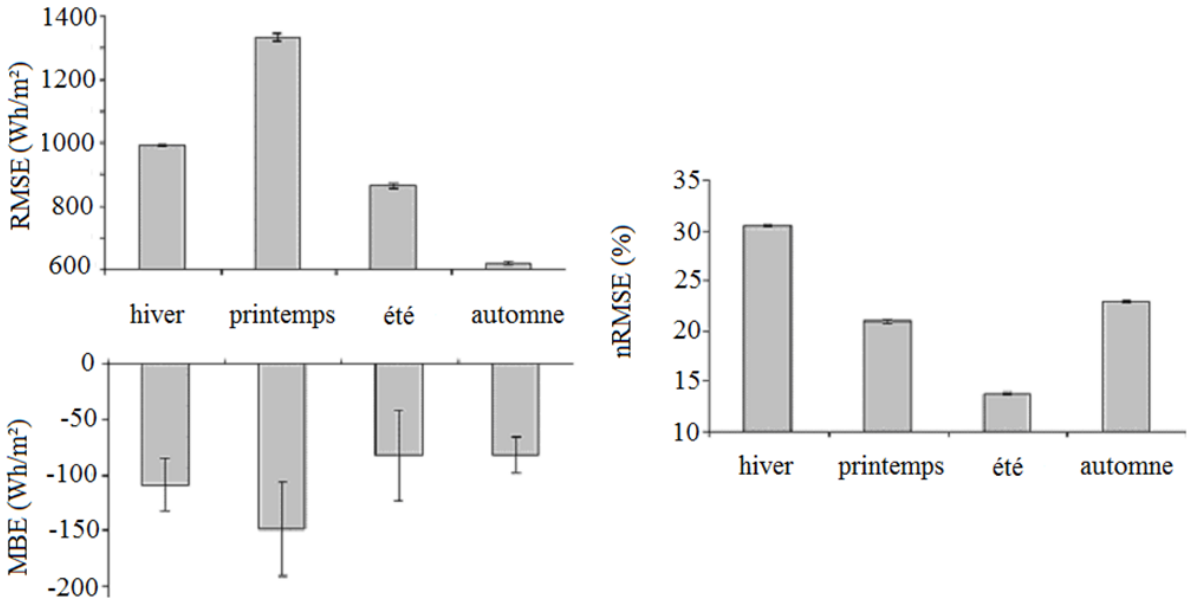


Figure 4-2 : erreurs saisonnières de prédiction j+1 avec intervalle de confiance à 95%. Ajaccio, année 1988-89

Cette première étude sur le rayonnement global d'Ajaccio a permis de tester différentes techniques de prévision. Seulement deux (*ARMA* et *PMC*) semblent être en mesure de prédire correctement. Bien que le modèle utilisant le *PMC* procure les meilleurs résultats, ceux obtenus avec la prédiction *ARMA* sont quasi-équivalents. D'autres manipulations sont à envisager pour permettre de mieux les hiérarchiser. L'autre point remarquable vient de la stationnarisation des séries temporelles. Les deux modes testés (k^* et CSI^*) donnent des résultats similaires, mais améliorent ceux obtenus sans stationnarisation.

Dans la prochaine section, nous proposons d'appliquer la démarche de prédiction utilisée ici, sur le cas réel d'un champ *PV*.

2.2. Prédiction d'énergie PV

Pour cette première étude visant à prédire l'énergie *PV*, nous nous sommes intéressés au pan orienté plein sud et incliné à 80° composé de modules *PV* issus de la technologie mono-Si décrits dans le chapitre 1 section 5-3, p 59. Cela correspond au cas le plus simple, car les modules sont ceux possédant les meilleurs rendements et car l'azimut zéro facilite l'élaboration du modèle de ciel clair. L'intérêt recherché de cette section, réside dans la confection d'une méthodologie qui sera utilisable en partie, voire adaptable au cas horaire. En effet, il est important de tester la méthodologie de prédiction sur le cas journalier (qui est plus simple) avant de passer au cas horaire (beaucoup plus complexe).

Dans le cas journalier, l'énergie *PV* peut, en première approximation être calculée par une relation linéaire simple reliant le rayonnement global (X) à l'énergie électrique (E_{dc}). Le coefficient de proportionnalité n'est autre que le produit entre le rendement des cellules et la leurs surfaces utiles. Dans le cas du *mono-Si*, le rendement est élevé et vaut η_{PV} ~ 15,3% (R^2 = 0,9967) comme on peut le voir sur la Figure 4-3.

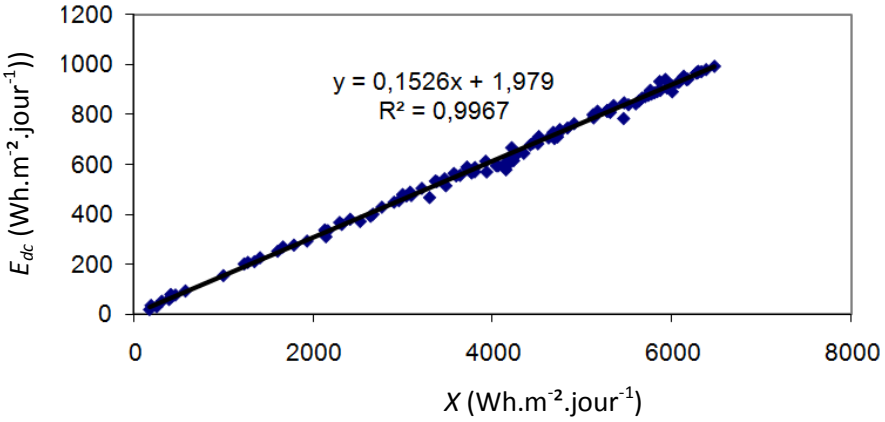


Figure 4-3 : courbe de rendement du mur *PV* testé. Relation entre le rayonnement global et l'énergie électrique surfacique produite

L'énergie électrique suit la relation 4-2 où X est le rayonnement global incliné de 80° et S la surface utile du mur valant 10,125 m².

$$E_{dc}\ [\text{Wh.jour}^{-1}] = \int P_{dc}dt = \eta_{PV}.S.X\ [\text{Wh.m}^{-2}.\text{jour}^{-1}] \qquad (4\text{-}2)$$

Nous nous sommes employés à prédire les énergies électrique *AC* produites par les modules *PV*. Cette manipulation a suivi de peu l'installation des modules. Aussi les seules données disponibles couvraient la période allant du 15 janvier 2009 au 15 Juin 2009. Compte tenu du faible nombre de données mesurées sur le champ *PV* (6 mois), les données de rayonnement global horizontal de Météo-France de la station synoptique de Campo dell'Oro distante de 10 km du laboratoire ont été utilisées pour les apprentissages de nos réseaux de neurones ou de nos modèles *ARMA*. Les données d'énergie *PV* de 2009 ont alors servi à la comparaison avec l'énergie électrique *AC* prédite. Pour utiliser cette méthodologie expérimentale, il a fallu être capable de transformer le rayonnement incliné en rayonnement horizontal, puis le rayonnement solaire en énergie électrique. L'apprentissage des modèles *ARMA* et *PMC* se fait avec du rayonnement global horizontal et la prédiction avec la méthodologie présentée sur la Figure 4-4. Les modèles *SOLIS*, *CLIMED2* et Klutcher ont contribué à pouvoir appliquer cette méthodologie (respectivement : modélisation par ciel clair pour le modèle *SOLIS*, et l'inclinaison du rayonnement global pour les deux autres).

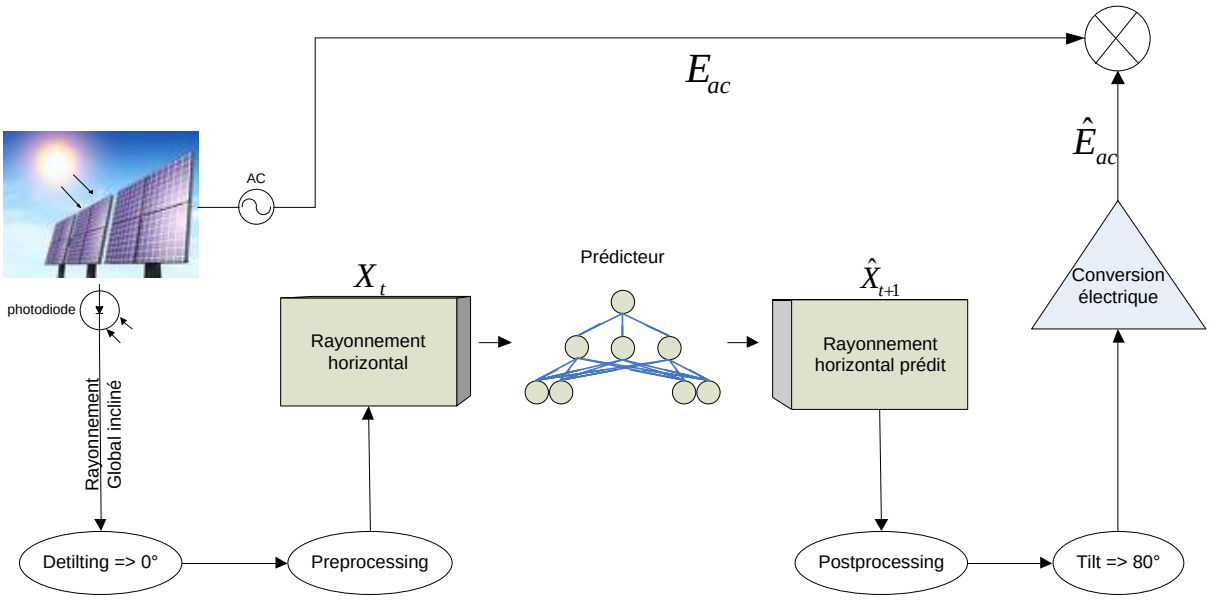


Figure 4-4 : méthodologie de prédiction de l'énergie *PV* d'un champ mono-Si, basée sur une simulation *ANN* et *ARMA* de rayonnement global horizontal

Pour l'inclinaison (notée « detilting » sur la figure) du rayonnement horizontal, l'erreur d'inclinaison commise pour passer de 80° à 0° est de 14 % sur la période couverte (calcul *nRMSE* entre le rayonnement horizontal mesuré et calculé à partir du rayonnement incliné).

La méthodologie utilisée est celle décrite au chapitre 1 section 3-2, p 44. Pour convertir le rayonnement global en énergie *PV* post-onduleur (E_{AC}, partie 2 de la Figure 4-5), il est nécessaire d'utiliser le coefficient *PR* (Performance Ratio, chapitre 1, section 4-3). Les résultats de prédiction doivent être corrigés afin de prendre en compte les composants comme l'onduleur, la température, l'énergie perdue, etc. Ces derniers peuvent modifier considérablement la production d'électricité au cours du temps. Le *PR* est calculé durant la période des mesures par le rapport entre les rendements mesurés et théoriques. Nous obtenons dans notre cas une valeur minimale de 0,71 en hiver, maximale de 0,80 en été et une moyenne annuelle de 0,76. Ce paramètre permet de déterminer l'énergie *AC* produite. Le diagramme en bloc de l'étude du mur *PV* est décrit par la Figure 4-5.

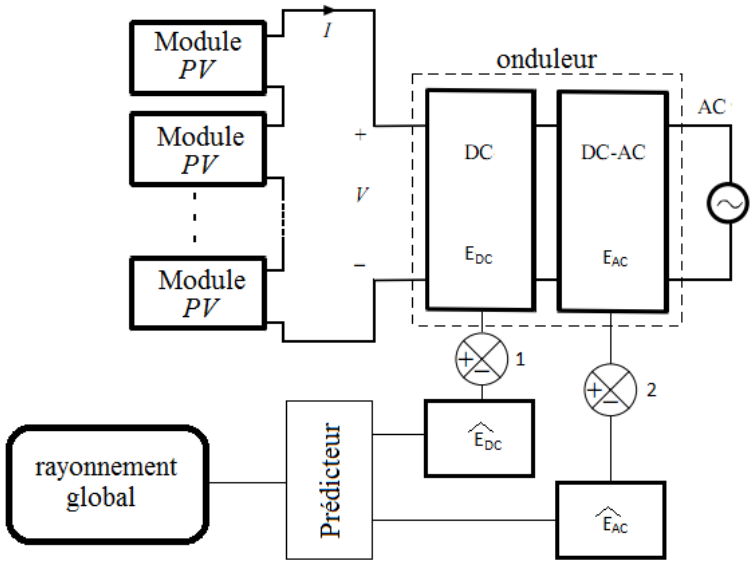


Figure 4-5 : schéma de la méthodologie de prédiction de l'énergie *PV*. En (1) la comparaison pré-onduleur et en (2) la comparaison post-onduleur

Pour ne pas entraver la visualisation et la compréhension des résultats, cette étude ne concerne que les signaux post onduleur ($\hat{E}_{ac}$). Ce signal est le plus important, car c'est souvent cette énergie (E_{ac}) qui est stockée, redistribuée ou utilisée directement.

Pour les premiers mois de prédiction, l'écart type mesuré concernant le courant *AC* est de 227 Wh, alors que sa moyenne vaut 506 Wh (Tableau 4-4). Le coefficient de variation atteint donc environ 50 %, ce qui prouve la haute variabilité de la production *PV* durant cette période. Lors de cette étude, sept modèles ont été testés : les deux meilleurs prédicteurs dans

le cas de la prédiction du rayonnement global (*PMC* et *ARMA*) ainsi que trois prédicteurs naïfs. Outre la prédiction par valeur moyenne, nous avons testé lors de cette expérience, la modélisation par ciel clair et la persistance. Le descriptif de ces modèles est :

- la modélisation par **ciel clair** (A). Le modèle *SOLIS* permet d'acquérir le rayonnement horizontal, puis la méthodologie d'inclinaison détaillée plus haut permet d'incliner ce dernier. La conversion électrique fait intervenir le rendement et le *PR* ;
- l'utilisation de la **moyenne** (B). La moyenne du rayonnement global horizontal est estimée à partir des données relatives aux 19 années disponibles sur Campo del'Oro. L'inclinaison et la conversion correspondent à ce qui a été détaillé dans la méthodologie A ;
- **Persistance** (C). La persistance du rayonnement du jour d'avant comme décrit dans le chapitre 2 section 3-2 ;
- ***PMC* sans stationnarisation** (D). L'optimisation conduit à utiliser $Endo^{1-8}$x3x1 ;
- ***PMC* couplé avec l'indice de clarté** (E). Les coefficients périodiques sont utilisés. L'optimisation conduit à choisir $Endo^{1-8}$x*3*x*1* ;
- ***PMC* couplé avec l'indice ciel clair** (F). Les coefficients périodiques sont utilisés. L'optimisation conduit à préférer $Endo^{1-8}$x*3*x*1* ;
- ***ARMA* couplé avec l'indice ciel clair** (G). Les coefficients périodiques sont utilisés. L'optimisation conduit à se servir d'une modélisation *ARMA(2,2).*

Les quatre derniers prédicteurs sont basés sur les principes énoncés par la Figure 4-4. D'après les fondements de cette étude sur la puissance *PV*, les prédicteurs sont construits pour effectuer une prédiction de rayonnement global. De ce fait, les résultats du paragraphe précédent concernant l'optimisation de chaque modèle (*PMC* ou *ARMA*) ont été repris. Le Tableau 4-4 synthétise les résultats relatifs à toutes ces techniques. Contrairement au cas horizontal, l'énergie électrique produite par les modules *PV* est maximale en hiver. Ce phénomène est une résultante de l'inclinaison des panneaux à 80°. C'est aussi la période où le signal est le plus bruité (écart type de 227 $Wh.jour^{-1}$ en hiver Vs 58 $Wh.jour^{-1}$ au printemps). Malgré cette différence importante, on constate que les conclusions faites dans le cas du rayonnement global horizontal restent applicables au cas de l'énergie *PV*. En effet, la stationnarisation améliore les résultats (*nRMSE* de 34,4 % en moyenne pour les méthodes E et F utilisant les *PMC* et de 35,6 % pour les *PMC* sans stationnarisation) et les deux modes

testés sont relativement équivalents (écart de *nRMSE* inférieur à 0,6 %). Comme on pouvait s'y attendre, les trois premiers modèles (ciel clair, moyenne et persistance) sont ceux qui donnent les résultats les moins performants (*nRMSE* supérieures d'environ 10 %). Il faut cependant relever que leur utilisation est très simple à mettre en œuvre et que les résultats obtenus, restent en première approximation exploitables. Durant les mois de mai et juin, la persistance génère une erreur inférieure à 20 % alors qu'elle est de presque 50 % les mois d'hiver. Ce résultat fait de ce prédicteur naïf un modèle de choix pour les prédictions durant les mois à faible nébulosité ou dès que les mesures antérieures sont inaccessibles. De plus, c'est le plus compétitif des trois modèles testés. Concernant le modèle *ARMA*, sur cette étude, il est totalement équivalent à ce que l'on peut observer avec le *PMC*. Jusqu'ici, les conclusions issues de cette manipulation sont identiques à celles élaborées dans le paragraphe précédent.

Le *MBE* est très souvent négatif avec les modèles complexes. Ce paramètre montre que la prédiction sous-estime la grandeur solaire, surtout dans le cas D, où l'on a *MBE* ~ -29 Wh.jour^{-1} en hiver. Le modèle qui génère les plus grandes valeurs de *MBE* est le modèle A. Cela s'explique par le fait que, par définition, ce dernier ne considère pas la couverture nuageuse, ce qui induit une surestimation conséquente (+137,5 Wh.jour^{-1} sur six mois). Comme nous l'avons vu dans le chapitre 2 section 4-1 ce paramètre d'erreur permet une interprétation proche de ce que l'on aurait dans le cas de la comparaison entre le cumul de la prédiction et de la mesure. Ce type de comparaison pourrait permettre de départager l'approche basée sur le *PMC* et celle basée sur *ARMA*.

	A ciel clair	B moyenne	C persistance	D PMC	E *PMC* clarté	F *PMC* ciel clair	G *ARMA* ciel clair
Jan-Fev			Moyenne AC: 506 Wh-Ecart Type : 227 Wh				
nRMSE (%)	56,4	57,1	47,0	41,7	38,7	**37,9**	38,4
RMSE (Wh/jour)	312,1	315,9	260,2	230,7	214,2	**209,7**	212,4
MBE (Wh/jour)	220,3	217,3	**-1,3**	-28,8	-1,6	4,3	-9,5
MAE (Wh/jour)	223,9	236,7	195,1	208,5	190,6	**184,1**	189,2
Mars-Avril			Moyenne AC: 480 Wh- Ecart Type : 202 Wh				
nRMSE (%)	49,3	45,1	47,1	38,3	38,3	37,9	**37,7**
RMSE (Wh/jour)	257,3	235,4	245,6	199,9	199,9	197,8	**196,8**
MBE (Wh/jour)	166,0	122,1	1,9	-2,7	-5,4	-3,6	**0,8**
MAE (Wh/jour)	167,8	159,9	179,1	160,7	162,9	162,3	**158,5**
Mai-Juin			Moyenne AC: 412 Wh- Ecart Type : 58 Wh				
nRMSE (%)	17,9	18,9	18,8	15,5	**16,2**	16,4	16,4
RMSE (Wh/jour)	74,7	78,9	78,1	64,7	**67,6**	68,5	69,2
MBE (Wh/jour)	41,6	16,0	**2,4**	-6,6	-13,9	-19,6	-24,8
MAE (Wh/jour)	140,1	167,5	140,2	133,8	132,4	**130,8**	131,2
Jan-Juin			Moyenne AC: 464 Wh- Ecart Type : 178 Wh				
nRMSE (%)	46,6	47,6	42,3	35,6	34,7	**34,1**	34,2
RMSE (Wh/jour)	233,9	238,9	210,2	178,7	174,1	**171,2**	172,0
MBE (Wh/jour)	137,5	134,7	**1,34**	-9,9	-7,3	-7,0	-9,9
MAE (Wh/jour)	140,1	167,5	140,2	133,8	132,4	**130,7**	131,2

Tableau 4-4 : erreur (*nRMSE* en %) de prédiction d'énergie E_{ac} entre le 15 janvier 2009 et Juin 2009 à Ajaccio. Les meilleurs résultats sont en gras

La Figure 4-6 reprend les quatre meilleurs prédicteurs en intégrant les prédictions et les mesures sur la période de test couverte.

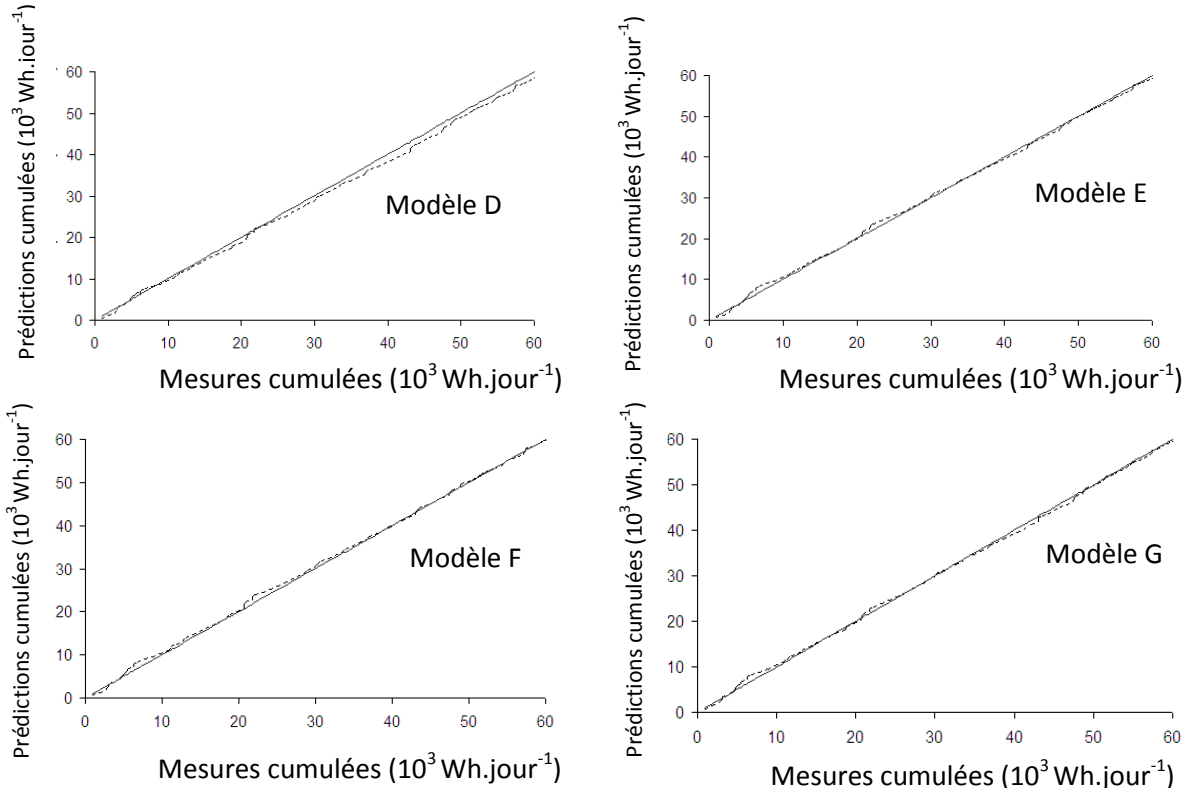


Figure 4-6 : cumul ou intégration des prédictions D (*PMC*), E (*PMC* et indice de clarté) F (*PMC* et indice de ciel clair) et G (*ARMA* et indice de ciel clair) par rapport aux mesures (prédiction de Janvier à Juin 2009 sur Ajaccio)

Sur cette figure, tous les modèles approximent bien les mesures. Seul le modèle D (*PMC* sans stationnarisation) a tendance à sous-estimer l'énergie électrique. Sans pouvoir l'expliquer en détail, ce résultat conduit à confirmer que la stationnarisation améliore aussi bien l'erreur de prédiction que l'erreur sur le cumul de la prédiction. Les autres modèles sont similaires d'après ce critère comparatif basé sur l'erreur d'intégration. Ce genre de courbe ne permet pas de départager les modèles *PMC* et *ARMA*. *Le* R^2 est identique pour les trois courbes et vaut 0,99. Le cumul des prédictions de janvier à juin diffère peu du cumul réel. En effet, après les six mois d'intégration, les erreurs sont très faibles : on obtient 2,1 % pour le modèle D et 1,5 % pour les trois autres prédicteurs utilisant une stationnarisation préalable des données.

A l'issue de cette série d'expériences, on retiendra que l'erreur donnée par le critère *nRMSE*, est certainement la plus intéressante pour tester une estimation à cet horizon. Les autres paramètres d'erreur contribuent aussi à l'analyse des phénomènes, mais ne peuvent vraisemblablement pas être utilisés seuls : ils complètent les analyses. Concernant le *nRMSE*, et la manière dont il est calculé dans cette thèse, des précautions sont à prendre. En effet, les représentations de l'erreur qu'il véhicule et l'aspect macroscopique de cette erreur, peuvent biaiser les interprétations qui en découlent. Elle est formée d'une accumulation d'erreurs, chacune indépendantes. Pour parfaire la comparaison et permettre l'export des résultats obtenus, il convient de l'étudier en détail. Le Tableau 4-5 représente la décomposition de cette erreur dans le cas du prédicteur F (*PMC* avec indice ciel clair) qui est le plus souvent le meilleur prédicteur lors de la manipulation précédente (7 fois sur 16 pour F et 4 fois pour G).

	nRMSE
Méthode globale de prédiction	19%
Spécificité de la période couverte	14%
Conversion électrique	~ 1%
Erreur totale observée	**34%**

Tableau 4-5 : décomposition du *nRMSE* pour la prédiction de l'énergie électrique

La deuxième ligne du tableau correspond à la spécificité de la période du test. Pour calculer cette erreur nous avons utilisé des données de rayonnement global horizontal relatives à cette période. L'erreur totale de prédiction dépassait de 14 % ce que l'on a habituellement sur le même site et sur une année. Cet écart est dû au climat particulièrement inclément de l'année 2009 et au fait que la période estivale (souvent facile à appréhender), n'est pas intégrée à l'étude sur les modules *PV*. L'erreur liée à la conversion électrique provient de l'utilisation du rendement des panneaux pour calculer l'énergie *PV*. Le terme d'erreur intitulée « méthode globale de prédiction » est une erreur par défaut, sachant l'erreur totale et les erreurs spécifiques détaillées ci-avant. L'estimation de l'erreur sur une année totale normale est donc estimée à 19 %. Ce chiffre peut être comparé aux *nRMSE* établis lors de l'étude du rayonnement global. En toute logique, il faudrait affiner les résultats issus du tableau précédent, en décomposant mieux les termes d'erreur, et réussir ainsi à isoler des éléments et les décorréler les uns des autres (pas de compensation). Ce genre de

décomposition n'est pas possible, on peut seulement affirmer que l'inclinaison des données avec notre méthodologie et les spécificités de l'année 2009, induisent toutes deux une *nRMSE* d'environ 14 %. Cela ne veut certainement pas dire que l'erreur à imputer au *PMC* est uniquement de 5 % (19 %-14 %) mais que plusieurs phénomènes se compensent.

L'analyse du Tableau 4-4 préconise d'utiliser principalement le *PMC* avec l'indice ciel clair (modèle F). Concernant les autres prédicteurs (hors naïfs), la hiérarchisation est difficile tant les différentes erreurs ne permettent pas de les comparer. Pour essayer de les départager, nous avons développé une approche basée sur la tendance de la prédiction et l'utilisation des facteurs Δ (nommée erreur de tendance au chapitre 2, section 4-1, p 90). Le tableau 4-6 présente ces résultats. Le modèle de ciel clair, la moyenne et la persistance, qui donnaient les moins bons résultats jusqu'ici, n'ont pas été intégrés à cette étude. Il faut noter qu'un test de Student sur les deux paramètres Δ_m et Δ_p pour *PMC* (modèle F) et *ARMA* (modèle G) indique que les prédictions réalisées avec *PMC* sont les meilleures. En effet, t = 0,364 et p = 0,71 pour *ANN et* t = 0,584 et p = 0,56 pour *ARMA*, où *p* représente la probabilité que la moyenne de la prédiction soit identique à la moyenne de la mesure et *t* le paramètre calculé du test. En moyenne, le *PMC* tend à mieux approximer les mesures que *ARMA*.

	***PMC* (D)**	***PMC-k* (E)**	***PMC-CSI** (F)**	***ARMA-CSI** (G)**
$\Delta < 0$	33 (24 %)	32 (24 %)	30 (22 %)	32 (24 %)
$0 < \Delta < 0,8$	54 (40 %)	53 (39 %)	56 (41 %)	53 (39 %)
$0,8 < \Delta < 1,2$	11 (8 %)	5 (4 %)	7 (5 %)	8 (6 %)
$\Delta > 1,2$	38 (28 %)	46 (34 %)	43 (32 %)	43 (32 %)

Tableau 4-6 : distribution des valeurs de Δ pour les processus D, E, F et G

La première ligne représente le nombre d'occurrence où la tendance n'est pas respectée. Cette méthodologie de comparaison permet de conclure que ces méthodologies sont différentes. Le modèle D qui jusqu'ici donnait les moins bons résultats, est celui qui permet d'obtenir ici, le plus de prédictions « très satisfaisantes » (ligne 3 du tableau), c'est aussi celui qui engendre le plus grand nombre de non-respects de la tendance (ligne 1 du tableau). Le paramètre ε est fixé pour cette étude à 20 % (chapitre 2, section 4-1, p 91). Les résultats étant conditionnés par ce choix, il n'est pas évident d'établir de règle objective pour ce paramètre. Une alternative consiste à étudier la répartition des prédictions suivants toutes les valeurs de Δ et non à partir des quatre intervalles définis plus haut. La Figure 4-7 permet de visualiser cette distribution.

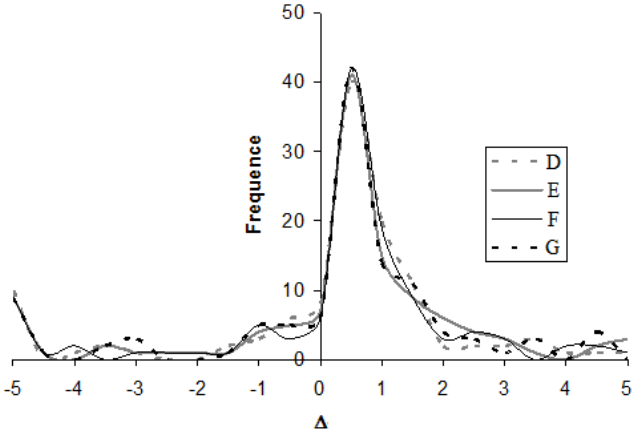


Figure 4-7 : distribution des valeurs de Δ pour les quatre meilleurs modèles *PMC* et *ARMA*

En fait, il est difficile de départager ces modèles sur un critère d'erreur à la tendance. Le seul élément qui permet de révéler une différence entre les modèles, est le test de Student utilisé sur Δ_m et Δ_p. Cependant, comme ce genre de test n'est pas conventionnel et que l'utilisation des Δ n'apporte pas de valeur ajoutée à la comparaison. Dans la suite des manipulations, nous ne retiendrons plus cet estimateur d'erreur comme discriminent.

A l'issue des deux premières expériences menées, nous avons opté pour l'étude : de la persistance comme prédicteur naïf, de la méthodologie basée sur le *PMC* et *ARMA* avec la stationnarisation *CSI* et l'utilisation des coefficients périodiques (*CSI**). En effet ces configurations nous ont permis d'obtenir les meilleurs résultats lors de cette première série de manipulation. La hiérarchisation de tous les prédicteurs testés correspond à : F*(PMC-CSI***)* ~ G(*ARMA-CSI***)* > E(*PMC*-k) > D(*PMC*) >> C(persistance) > A(ciel clair) ~ B(moyenne). Dans la prochaine partie, nous essayerons de quantifier l'impact des données non endogènes sur le système étudié. Ainsi, dans un premier temps, nous allons voir s'il est intéressant d'utiliser l'analyse multivariée en prédiction du rayonnement global journalier et de l'énergie *PV*.

2.3. Utilisation de variables exogènes

Nous avons vu dans le paragraphe précédent que les prédictions de rayonnement global réalisées en utilisant uniquement, en entrée du *PMC*, des données de nature identique à la nature de la sortie, ne permettent pas de faire mieux que la méthodologie *ARMA*. Nous allons

essayer d'établir si l'utilisation d'entrées de natures exogènes modifie cette conclusion, et améliore ainsi la prédiction du rayonnement global horizontal et de l'énergie *PV*. Plus l'information en entrée du réseau de neurones est corrélée avec l'information future, plus la performance du réseau est supposée s'améliorer. Des paramètres météorologiques tels que la pression, la température, les précipitations ou la nébulosité ont une incidence directe sur le rayonnement solaire. Ces variables dénommées « exogènes » peuvent être intégrées au vecteur d'entrée.

a. Prédiction de rayonnement global

La méthode que nous avons développée (chapitre 3 et de la section 3, p 112) consiste à utiliser des variables météorologiques en plus des entrées endogènes comme le montre la Figure 4-8. A partir de la variable température *T*, on définit quatre sous-variables qui sont la température moyenne (*Ta*), la température maximale (*TM*), la température minimale (*Tm*) et la température de la nuit à 3 heures du matin (*Tn*). De plus, à partir de la pression *P* on définit le gradient de pression journalier ($DGP = P_t - P_{t-1}$). Les données exogènes présentées sur cette figure sont détaillées dans le chapitre 1, section 5-1.

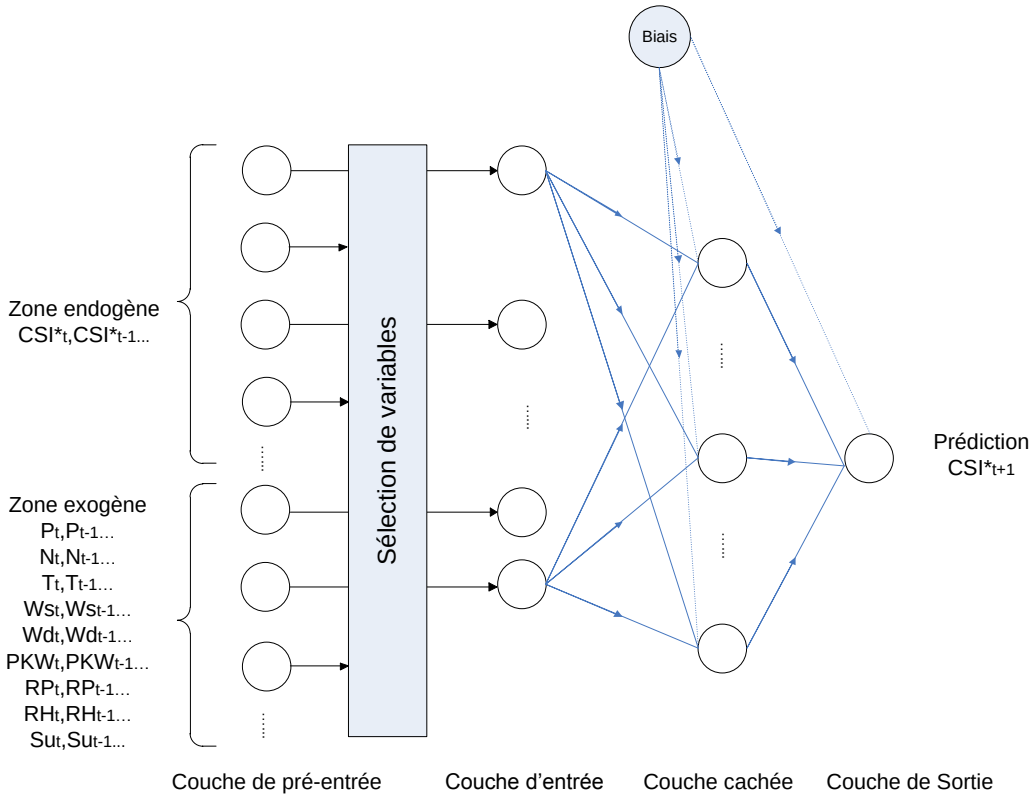


Figure 4-8 : détail du système de prédiction multivariée avec le *PMC* dédié à l'horizon j+1

Durant cette manipulation, la sélection du nombre de variables endogènes est réalisée avec la méthode des *PACF* et le nombre de variables exogènes avec la méthode des cross-corrélations. La sélection des variables et la méthode d'optimisation globale est décrite dans le chapitre 3, section 3, p 118. Pour cette étude, on utilise les données allant de janvier 1998 à décembre 2007 concernant les sites d'Ajaccio et de Bastia. Les huit premières années sont dédiées à l'apprentissage et les deux dernières aux tests. La Figure 4-9 présente graphiquement les corrélations qui peuvent exister entre le CSI^* et les variables exogènes utilisées. On constate que certaines variables ne sont certainement pas propices à l'analyse multivariée du rayonnement global.

Pour Ajaccio, on constate que seulement quatre grandeurs exogènes sont réellement corrélées avec le rayonnement global (*Su*, *P*, *N* et *DGP*) sur la base du seuil choisi. Concernant Bastia, il y a trois variables qui se démarquent de par leur corrélation (*Su*, *RH et N).* Toutes ces corrélations ne concernent que le premier lag. Aussi, en reprenant les conventions de la Figure 4-8 pour Ajaccio, la zone exogène d'entrée comprend ($Su^1P^1N^1DGP^1$), ce qui correspond à Su_t, P_t, N_t et DGP_t. Concernant le nombre de neurones endogènes d'entrée, l'étude des *PACF* conduit à considérer 4 neurones pour Bastia (nommée Endo$^{1\text{-}4}$, soit $CSI_t^*, \ldots, CSI_{t-3}^*$) et 2 pour Ajaccio (nommée Endo^{1,2}, soit CSI_t^*, CSI_{t-1}^*)

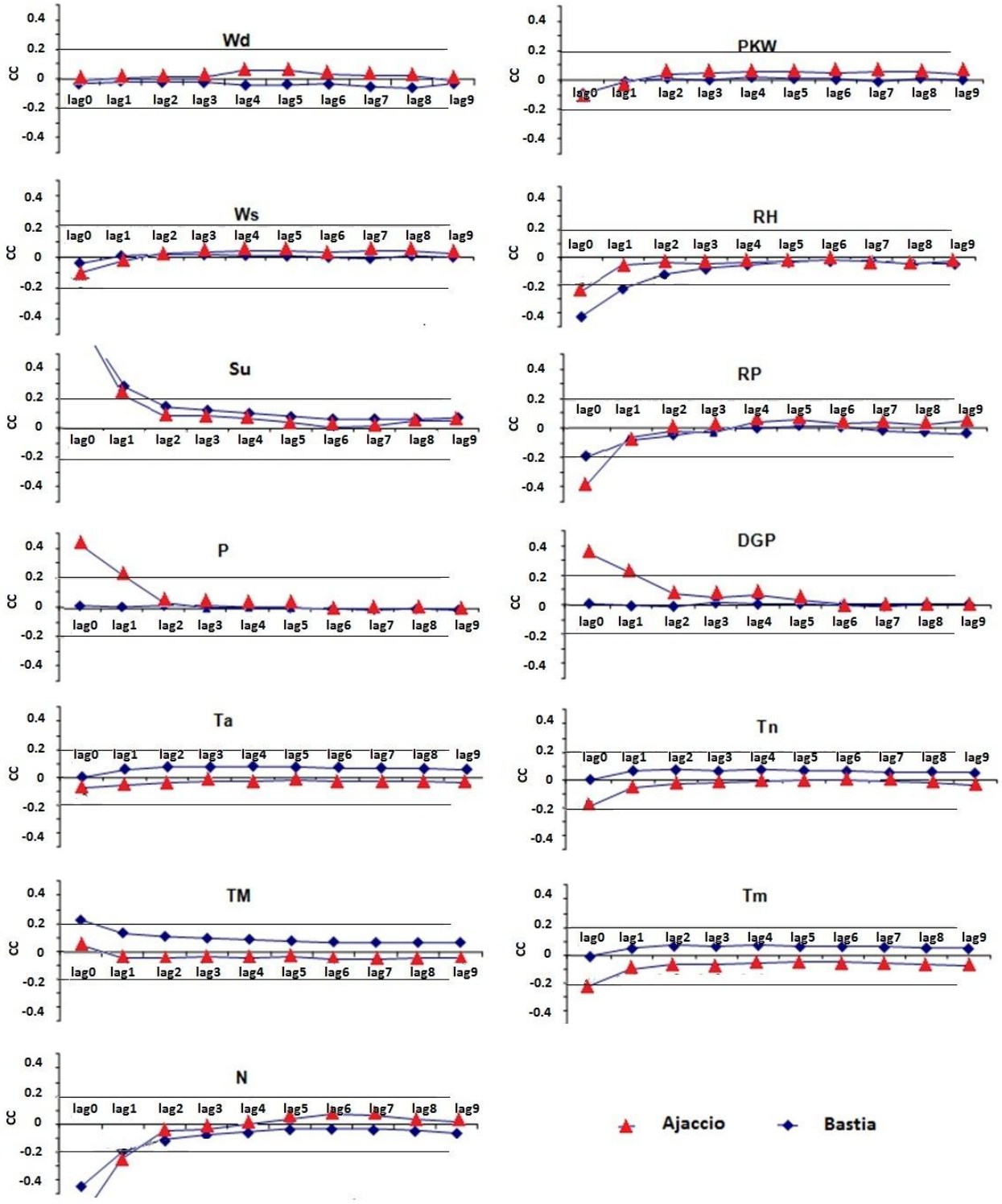


Figure 4-9 : corrélation de Pearson entre l'indice ciel clair et les variables exogènes pour Bastia et Ajaccio. *P* = pression, *DGP* = gradient de pression journalier, *N* = nébulosité, $T_{M,m,a,n}$ = la température maximale, minimale, moyenne et de nuit, *Ws* = vitesse du vent, *Wd* = direction du vent, *PKW* = vitesse de rafale, *RP* = hauteur de précipitations, *RH* = humidité et *Su* = insolation

Sur les Tableaux 4-7 et 4-8, est représenté l'impact des variables exogènes utilisées sur le site d'Ajaccio et de Bastia. Le nombre de neurones cachés est fixé arbitrairement à 5, ce nombre sera affiné lors de la prochaine étape d'optimisation.

Lag Endo	Lag Su	Lag RH	Lag N	Neurones cachés	*nRMSE*	IC95%
4	-	-	-	5	25,85%	0,46%
4	**1**	**1**	**1**	**5**	**25,60%**	**0,18%**
4	1	-	-	5	25,65%	0,17%
4	-	1	-	5	25,66%	0,09%
4	-	-	1	5	25,58%	0,13%

Tableau 4-7 : erreurs de prédiction liées à l'utilisation des variables exogènes sur la couche d'entrée. Site de Bastia, en gras les meilleurs résultats

Lag Endo	Lag Su	Lag P	Lag DGP	Lag N	Neurones cachés	*nRMSE*	IC95%
2	-	-	-	-	5	22,50%	0,16%
2	**1**	**1**	**1**	**1**	**5**	**21,62%**	**0,16%**
2	1	-	-	-	5	22,04%	0,08%
2	-	1	-	-	5	22,43%	0,14%
2	-	-	1	-	5	22,47%	0,14%
2	-	-	-	1	5	21,79%	0,11%

Tableau 4-8 : erreurs de prédiction liées à l'utilisation des variables exogènes sur la couche d'entrée. Site d'Ajaccio, en gras les meilleurs résultats

Sur le site de Bastia, on constate que l'apport des variables exogènes est minime. Le fait d'utiliser séparément les trois variables donne une erreur de prédiction similaire à l'utilisation d'une seule variable. Comme, il est impossible de les départager, dans la suite de notre optimisation, nous exploiterons ces trois grandeurs. Pour Ajaccio, le phénomène est relativement différent, les quatre grandeurs étudiées séparément ne permettent pas de prédire le rayonnement global avec la même efficacité. En effet, la pression et la pression différentielle sont moins intéressantes que les deux autres lorsqu'elles sont placées en entrée du *PMC*. De ce fait, pour ne pas trop complexifier le *PMC* et en accord avec le principe de

parcimonie, dans la suite, nous ne garderons que l'insolation et la nébulosité pour la prédiction du rayonnement global sur Ajaccio. La partie suivante concerne l'optimisation du nombre de neurones sur la couche cachée. Nous avons fait varier ce nombre conformément à ce que l'on décrit dans le chapitre 3, section 3-1, p 116. Les tableaux 4-9 et 4-10 donnent les résultats de cette optimisation.

Lag Endo	Lag Su	Lag RH	Lag N	Neurones cachés	*nRMSE*	IC95%
4	1	1	1	1	25,52%	0,10%
4	**1**	**1**	**1**	**2**	**25,43%**	**0,11%**
4	**1**	**1**	**1**	**3**	**25,43%**	**0,16%**
4	1	1	1	4	25,61%	0,05%
4	1	1	1	5	25,60%	0,18%
4	1	1	1	6	25,59%	0,42%
4	1	1	1	7	25,60%	0,27%
4	1	1	1	10	25,70%	0,30%

Tableau 4-9 : optimisation de la couche cachée pour le site de Bastia. En gras, les valeurs minimales de *nRMSE*

Lag Endo	Lag Su	Lag N	Neurones cachés	*nRMSE*	IC95%
2	1	1	1	21,61%	0,01%
2	1	1	2	21,75%	0,18%
2	**1**	**1**	**3**	**21,54%**	**0,05%**
2	1	1	4	21,67%	0,11%
2	1	1	5	21,61%	0,12%
2	1	1	6	21,83%	0,53%
2	1	1	7	21,87%	0,20%
2	1	1	10	21,78%	0,19%

Tableau 4-10 : optimisation de la couche cachée pour le site d'Ajaccio. En gras, la valeur minimale de *nRMSE*

Sur la base de ces travaux, pour les deux sites, les résultats sont identiques. Il semble que trois neurones cachés donnent les meilleurs résultats. Pour Bastia, il faut calculer jusqu'au 3^{ième} chiffre après la virgule pour dépasser l'égalité. Les caractéristiques du *PMC* que l'on vient d'énoncer peuvent s'écrire sous forme abrégée :

- ($Endo^{1\text{-}4}N^{1}Su^{1}RH^{1}$)x*3*x*1* pour Bastia et ses 4 neurones endogènes, 3 neurones exogènes et 3 neurones cachés ;

- ($Endo^{1\text{-}2}N^{1}Su^{1}$)x*3*x*1* pour Ajaccio et ses 2 neurones endogènes, 2 neurones exogènes et 3 neurones cachés.

Sur la Figure 4-10 on peut voir graphiquement l'impact lié à l'utilisation des données exogènes sur une représentation y = x de la correspondance entre rayonnement mesuré et calculé.

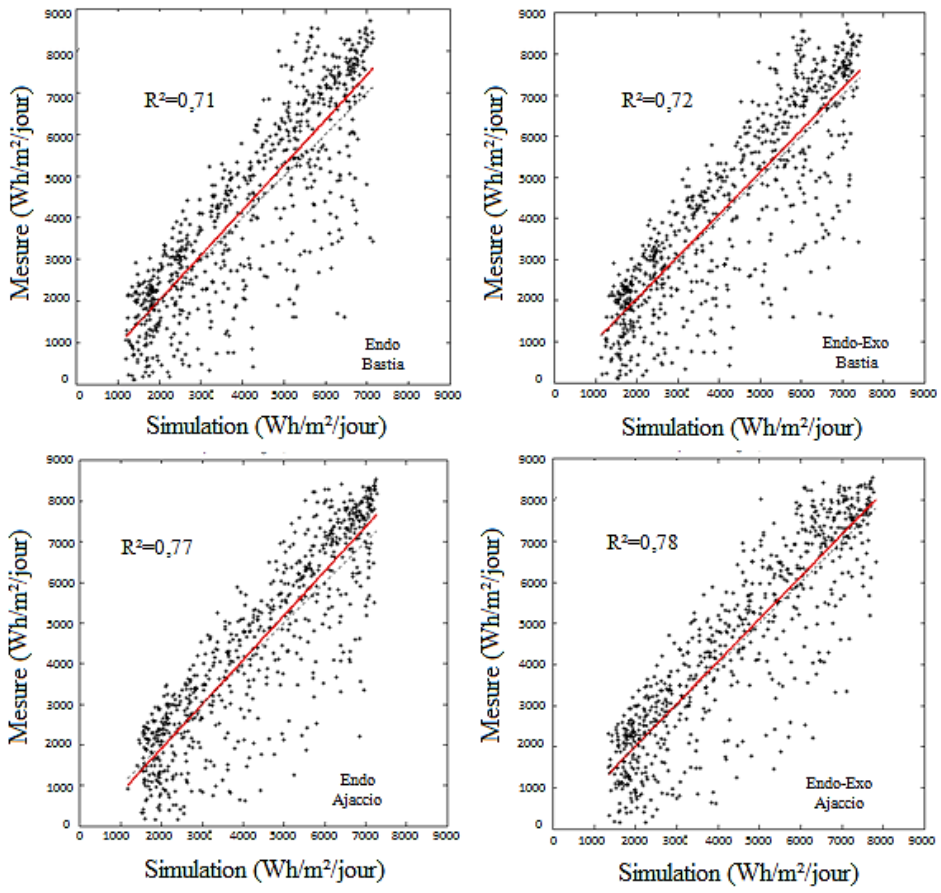


Figure 4-10 : comparaison entre l'utilisation des données endogènes et exogènes (à droite) et seulement endogènes (à gauche). Les courbes du haut sont relatives à Bastia et celles du bas à Ajaccio. La ligne rouge représente la meilleure approximation linéaire, et le trait pointillé, la droite y = x

Même si le coefficient de détermination est amélioré par l'utilisation des variables exogènes sur ces graphiques, il n'est pas évident d'en quantifier le réel impact : on constate une augmentation de 0,01 (1%) pour les deux sites. D'après les tableaux 4-9 et 4-10, on constate que l'apport de la méthodologie multivariée permet un gain (diminution du *nRMSE*) de 0,5% sur Bastia et d'environ 1 % sur Ajaccio. Sur la Figure 4-11, nous pouvons apprécier mois par mois l'apport des variables exogènes sur le *nRMSE*. Une valeur positive signifie une diminution du *nRMSE* avec l'utilisation des données exogènes. On constate que dans la majorité des cas, l'utilisation des variables exogènes est une véritable valeur ajoutée à la prédiction : 17 mois sur 24 pour Ajaccio et 15 mois sur 24 pour Bastia. On observe que la différence de *nRMSE* varie entre -3 % et +3 %, pour Ajaccio et entre +2 % et – 2 % pour Bastia. Même si cette différence est souvent positive, sa valeur absolue est importante lorsqu'elle devient négative.

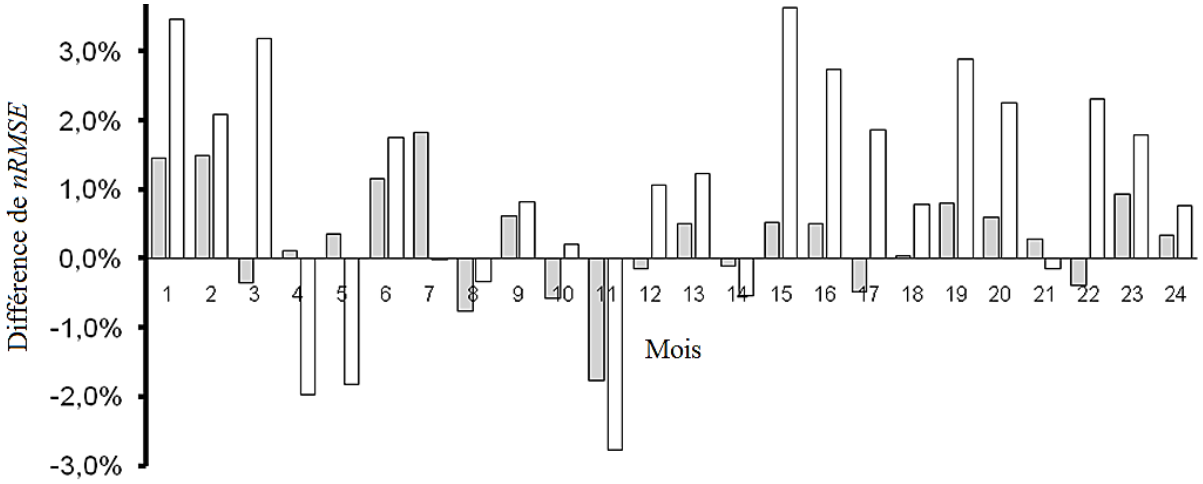


Figure 4-11 : différence mensuelle calculée sur les deux ans de prédiction, entre les *nRMSE* obtenus avec l'utilisation des variables exogènes et ceux sans ces variables. Histogrammes : gris = Bastia ; blanc = Ajaccio

Dans le premier paragraphe, nous avions testé l'erreur liée au cumul ou à l'intégration des mesures et des prédictions. Nous allons voir dans le cadre de l'apport des variables exogènes l'interprétation de ce paramètre. La Figure 4-12 décrit les résultats issus de cette étude. Cependant, comme dans le cas uniquement endogène et la comparaison des modèles *PMC* et *ARMA*, cette représentation graphique ne permet pas de différentier les deux approches. La prédiction sous-estime systématiquement la ressource.

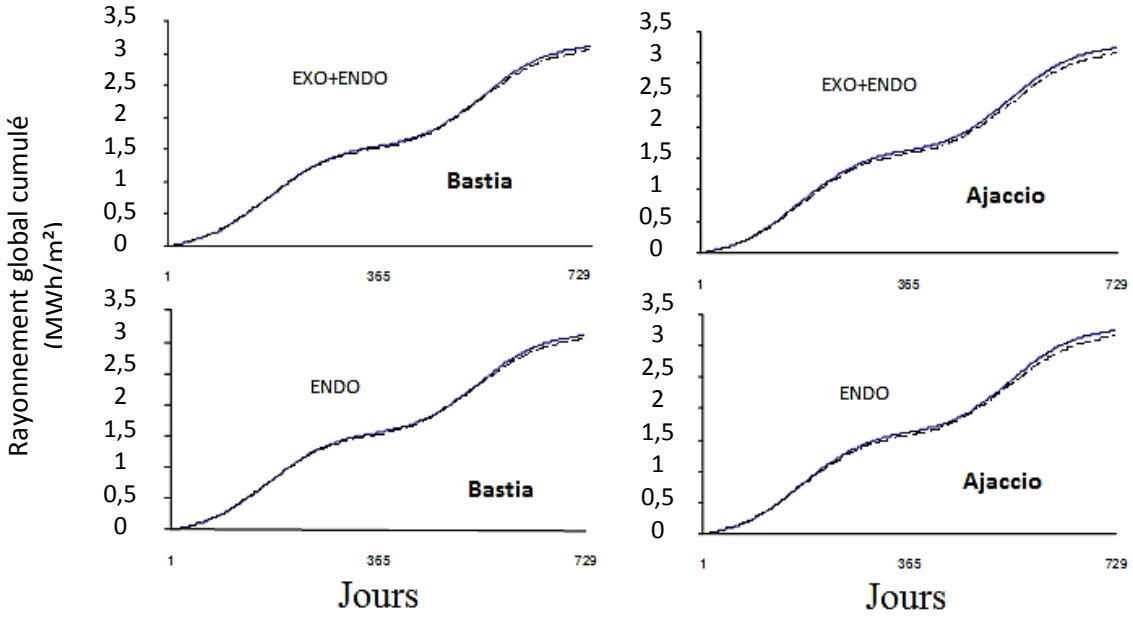


Figure 4-12 : rayonnement global cumulé avec l'utilisation des entrées exogènes. Trait pointillé : prédiction ; trait plein : mesure. (R^2=0,99 pour les quatre graphiques)

Il existe une alternative pour représenter la différence entre le cumul de la prédiction et celui de la mesure. Cette dernière se focalise sur l'écart qui existe entre les deux courbes. Elle consiste à utiliser la différence algébrique entre le cumul prédit et le cumul mesuré, et ce pour les deux ans de prédiction. Une valeur négative signifiant que la simulation sous-estime le rayonnement global. La Figure 4-13 matérialise ces différences de cumul entre la mesure et la prédiction réalisée avec ou sans données exogènes.

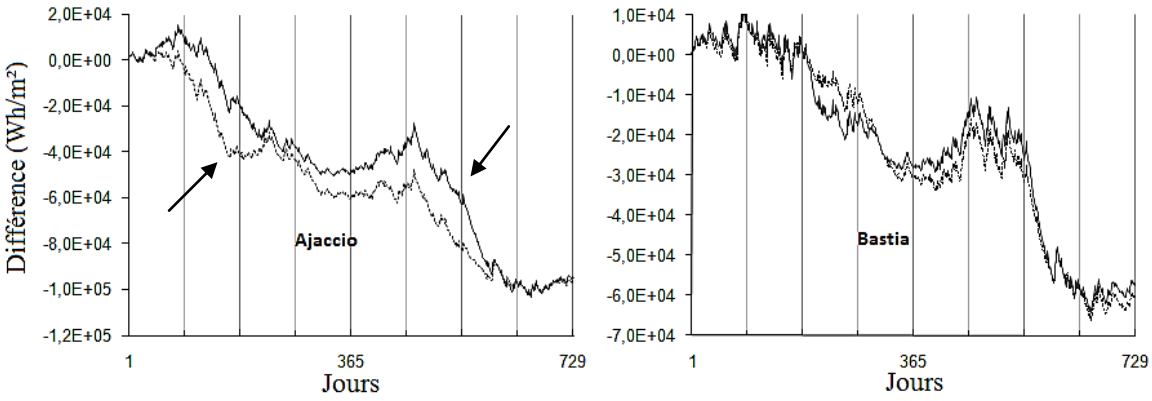


Figure 4-13 : différence entre la simulation (*PMC*) et la mesure pour Ajaccio et Bastia. Trait plein : *PMC* avec données endogènes ; traits pointillés : *PMC* avec variables endogènes et exogènes

On remarque sur Bastia que le cumul des prédictions faites avec l'utilisation des variables exogènes et des endogènes est assez proche durant les deux années. Cependant, dans le cas d'Ajaccio, même si à l'issue des deux années, l'écart est équivalent, on constate que l'écart est maximum en été (flèches sur la figure 4-13) et minimum en automne. Ce résultat laisse présager que l'utilisation des variables exogènes est intéressante les mois à forte occurrence nuageuse. Les autres mois essentiellement en été et au printemps, l'apport semble moins marqué. Rappelons que cette conclusion fait suite à une série de manipulation sur le cumul de la prédiction et non relative à l'erreur de prédiction standard *nRMSE*. De plus, à l'issue des deux années de prédiction, on peut quantifier l'écart engendré par la simulation sur Ajaccio par la règle de régression suivante :

$$\text{Erreur\_cumulée [Wh/m²]} = -146{,}8.\text{nombre\_de\_jour} \ (R^2=0{,}94)$$

Et Pour Bastia par la relation :

$$\text{Erreur\_cumulée [Wh/m²]} = -73{,}7.\text{nombre\_de\_jour} \ (R^2=0{,}81)$$

On voit que l'erreur cumulée sur Bastia est moitié plus faible que celle calculée sur Ajaccio. On peut estimer que l'erreur cumulée est de -146 Wh/m² (sous-estimation) par jour à Ajaccio et -73,7 Wh/m² par jour à Bastia. Cependant, il faut garder à l'esprit que même si l'erreur cumulée est plus faible sur Bastia, la prédiction sur ce site est plus compliquée et moins précise que celle faite sur le site d'Ajaccio (Tableaux 4-9 et 4-10).

En résumé à l'issue du paragraphe précédent, nous avions trouvé que les prédictions faites avec le *PMC*, une stationnarisation et des valeurs endogènes, étaient aussi performantes que celles faites avec le modèle *ARMA* et la même stationnarisation. Ici, Nous avons montré que l'utilisation des variables exogènes améliorait la prédiction des *PMC* de 1 % dans le cas d'Ajaccio (*nRMSE* = 21,54 % Vs 22,5 %) et de 0,4 % sur Bastia (*nRMSE* = 25,43 % Vs 25,85 %).

b. Prédiction de l'énergie *PV*

La dernière expérience concernant l'impact des variables exogènes, est relative au champ *PV* et l'expérience menée et décrite dans la section précédente (chapitre 4 section 2-2). L'unique différence avec la manipulation déjà menée, concerne l'utilisation de neurones d'entrée du *PMC* d'origine exogène. Les résultats sont donnés dans le tableau 4-11. La

prédiction concerne l'énergie *AC* et le test porte sur les quatre méthodologies suivantes : *PMC* avec entrées endogènes (PMC endo) ; *PMC* avec entrées endogènes et exogènes (PMC endo-exo); *ARMA* ; persistance. Pour mémoire, le performance ratio de l'installation est de 0,76 et la stationnarisation retenue concerne le CSI^*.

	PMC endo-exo	*PMC endo*	*ARMA*	*Persistance*
Jan-Fev				
nRMSE (%)	**37,5**	37,9	38,4	47,0
RMSE (Wh/jour)	**207,2**	209,7	212,4	260,2
MBE (Wh/jour)	27,6	4,3	-9,5	**-1,3**
MAE (Wh/jour)	**177,7**	184,1	189,2	195,1
Mars-Avril				
nRMSE (%)	**36,1**	37,9	37,7	47,1
RMSE (Wh/jour)	**188,2**	197,8	196,8	245,6
MBE (Wh/jour)	9,89	-3,6	**0,8**	1,9
MAE (Wh/jour)	**148,5**	162,3	158,5	179,1
Mai-Juin				
nRMSE (%)	**16,4**	**16,4**	**16,4**	18,8
RMSE (Wh/jour)	**68,4**	68,5	69,2	78,1
MBE (Wh/jour)	-10,2	-19,6	-24,8	**2,4**
MAE (Wh/jour)	**122,4**	130,8	131,2	140,2
Jan-Juin				
nRMSE (%)	**33,1**	34,1	34,2	42,3
RMSE (Wh/jour)	**164,4**	171,2	172,0	210,2
MBE (Wh/jour)	7,31	-7,0	-9,9	**1,34**
MAE (Wh/jour)	**122,4**	130,7	131,2	140,2

Tableau 4-11 : prédiction de l'énergie E_{ac} issue du mur *PV* incliné à 80° d'Ajaccio. Les quatre meilleurs estimateurs (*PMC* endo-exo, *PMC* endo, *ARMA* et la persistance) sont testés

On constate que l'utilisation des données exogènes, comme dans le cas de la prédiction du rayonnement global précédemment discuté, permet de diminuer le *nRMSE* de 1 % sur Ajaccio. De plus, on retrouve le résultat déjà énoncé dans le paragraphe précédent concernant le cumul de la prédiction : l'utilisation des données exogènes montrent un intérêt, essentiellement durant les mois à forte occurrence nuageuse. Au printemps, le *PMC* doté d'entrées uniquement endogènes ou la prédiction de référence basée sur la méthodologie *ARMA*, semblent largement suffisants. De plus, sur la globalité de la période du test, par comparaison avec le prédicteur naïf (persistance) ou avec notre référence (*ARMA*), le *PMC* avec utilisation de variables exogènes, améliore la prédiction *AC* de respectivement 9 % et 1

%. On retrouve le résultat précédemment cité, à savoir que le processus *ARMA* est équivalent à un *PMC* sans variable exogène. Ce dernier étant moins pertinent que le *PMC* couplant les deux types d'entrées. La Figure 4-14 montre l'écart entre la prédiction cumulée et la mesure de l'énergie électrique concernant l'utilisation du *PMC* avec les variables exogènes. On s'aperçoit que la prédiction surestime la production d'énergie. Cependant, l'utilisation du *PMC* avec des données exogènes peut relativement bien quantifier la ressource comme le montre le coefficient de détermination (R^2 = 0,99). Si l'on trace le même type de courbe pour *ARMA* et la persistance, on obtient respectivement R^2 = 0,99 et R^2 = 1.

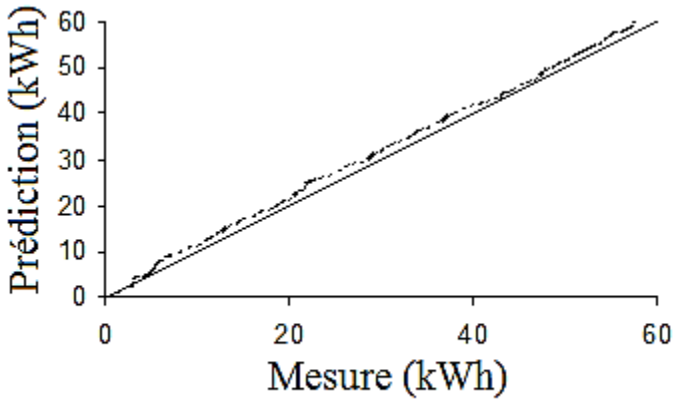


Figure 4-14 : relation entre la mesure cumulée et la prédiction avec le *PMC*' endo/exo (R^2=0,99) pour la période janvier-juin 2009 (le trait plein représente le cas idéal y = x)

Cette Figure 4-14 qui est de nature cumulative, est intéressante car elle montre l'erreur globale générée. Une bonne estimation doit avoir une faible valeur de *nRMSE* journalière mais aussi une faible erreur de la prédiction d'énergie cumulée. Durant les 6 mois, l'erreur globale absolue de la prédiction cumulée est inférieure à 4 kWh (écart absolu de 7%). Ce qui est relativement faible. Sur la Figure 4-15 qui présente l'allure des mesures et des prédictions, nous pouvons remarquer que les mesures sont très bruitées et que l'inclinaison à 80° des modules *PV* est très pénalisante ; les hautes valeurs d'énergie sont observées en hiver quand la couverture nuageuse est souvent très prononcée. De plus, l'erreur relative est importante en été dans ce cas-là, chose que nous n'avons pas observé dans le cas horizontal. Cependant, après 90 jours (3 mois à partir du mois d'avril), la prédiction est très proche de la mesure.

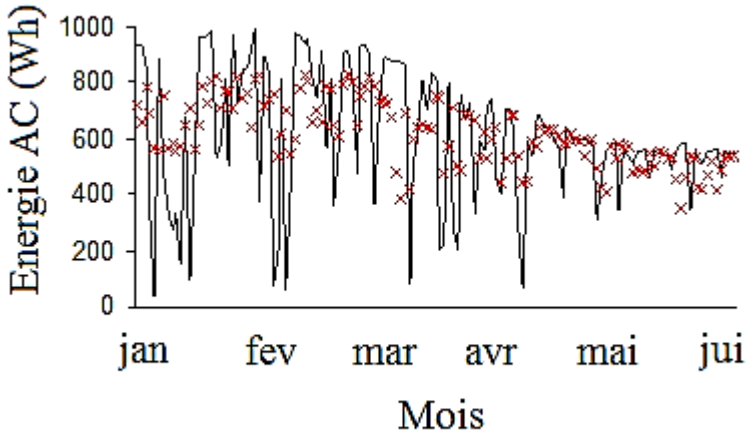


Figure 4-15 : représentation de la puissance électrique mesurée et simulée sur les modules *PV* d'Ajaccio en 2009. Les croix sont relatives aux prédictions et le trait plein aux mesures

L'étude que l'on vient de présenter ici corrobore l'analyse issue du paragraphe précédent liée à la prédiction du rayonnement global. En effet, le bilan de cette série de manipulations est le suivant :

- sans l'utilisation de variables exogènes, *PMC* et *ARMA* sont équivalents ;

- pour les mois où la nébulosité est importante, l'utilisation des variables exogènes permet d'améliorer la qualité de la prédiction ;

- il faut stationnariser les séries temporelles avec la méthodologie k^*, ou si possible CSI^* ;

- la persistance est un prédicteur naïf intéressant, qui donne de très bons résultats au printemps et en été ;

- la méthodologie de prédiction testée s'applique aussi bien au cas du rayonnement global qu'au cas de l'énergie *PV*.

La prochaine étude que l'on va présenter dans la section 2-4, vise à tester la méthodologie basée sur le *PMC* en situation assez particulière. Il s'agit de délocaliser géographiquement le recueil des mesures d'apprentissage et la zone où sont faites les prédictions. L'idée de cette manipulation est apparue lors de la manipulation de prédiction de la puissance *PV*. En effet, une des étapes consistait à entrainer le prédicteur sur du

rayonnement horizontal puis à effectuer un test sur du rayonnement incliné converti en énergie *PV*. Les deux zones dédiées à l'entraînement et à l'apprentissage étaient certes sur Ajaccio, mais tout de même distantes d'environ 10 km. Cependant, les résultats observés étaient en parfaite adéquation avec ceux obtenus dans le cas standard du rayonnement global horizontal. Les 10 km séparant les deux zones n'ont pas eu d'influence apparente sur l'efficacité du *PMC*. Le prochain chapitre doit permettre de vérifier si l'augmentation de cette distance perturbe les résultats de prédiction.

2.4. La non-cohérence spatiale entre apprentissage et test

Il est souvent fondamental d'étudier le rayonnement global sur un site où aucun relevé antérieur n'est disponible. Cet exercice est typiquement nécessaire pour travailler sur la problématique du foisonnement ou la faisabilité d'intégration d'une centrale solaire. Il nous est apparu intéressant de tester cette problématique afin de vérifier si le *PMC* était assez robuste pour supporter cette approche (délocalisation du test). Nous avons choisi le pas journalier pour sa « simplicité » vis-à-vis des autres horizons plus courts. Nous avons considéré uniquement le *PMC* avec l'indice de ciel clair et exclusivement des variables endogènes pour restreindre le nombre de manipulation. De plus, nous étions limités quant aux variables exogènes suivant les sites concernés (Bastia et Corte) par cette prédiction délocalisée. Nous avons ainsi testé trois techniques de prédiction de rayonnement global pour ces deux sites : la méthode A qui est liée à l'utilisation de la délocalisation de l'apprentissage, la méthode B qui est la méthode classique de prédiction avec un *PMC* comme détaillée plus en amont de cette section et la méthode C qui est le prédicteur naïf. Plus précisément, ces trois méthodes sont :

- apprentissage obtenu avec les données d'Ajaccio de 1971-1987 (A), et la prédiction délocalisée calculée sur Bastia (année 1996) et Corte (année 2007) ;

- apprentissage obtenu avec les données issues du site testé (B ; pour Corte de 2002-2006 et pour Bastia de 1991-1995), et la prédiction calculée sur Corte (année 2007) et sur Bastia (année 1996) ;

- le prédicteur naïf (C) basé sur la persistance qui ne nécessite pas d'apprentissage.

Dans les cas A et B les années de prédiction pour Bastia et Corte sont différentes. Le but de cette étude n'étant pas de comparer ces localisations entre elles, cette caractéristique n'est pas dommageable pour la suite de l'étude. Le Tableau 4-8 synthétise les résultats de prédiction. Outre la comparaison des modèles, le site de Bastia possède un comportement météorologique très difficile à modéliser.

Localisation de la prédiction	**Techniques de Prédiction**	***RMSE* (Wh/m²)** ***nRMSE* (%)± IC95**	**R^2**
Bastia (année 1996)	Ajaccio (A)	1383 29,19±0,13	**0,709**
	Bastia (B)	**1288** **27,51±0,2**	**0,709**
	Persistance (C)	1468 31,4	0,651
Corte (année 2007)	Ajaccio (A)	1213 25,88±0,17	**0,787**
	Corte (B)	**1112** **23,73±0,15**	**0,787**
	Persistance (C)	1325 28,3	0,712

Tableau 4-12 : comparaison des trois techniques de prédiction journalière

Conformément à ce que nous pouvions imaginer, la localisation de l'apprentissage modifie la qualité de la prédiction. Le cas B donne systématiquement les meilleurs résultats. Nous voyons que la délocalisation est en partie compensée par la taille de l'échantillon d'apprentissage (16 années pour Ajaccio contre 5 pour les deux autres villes). Sur un site où aucun historique n'est disponible, une des seules façons habituelles de procéder est l'utilisation de la persistance. Nous voyons que la méthodologie basée sur la délocalisation, améliore sensiblement les résultats (*RMSE*=1468 Vs 1383 Wh/m² pour Bastia et 1325 Vs 1213 Wh/m² pour Corte). En termes de *nRMSE*, le gain moyen est conséquent en utilisant l'apprentissage délocalisé (-4,2 % vis-à-vis de la persistance). Si l'on se réfère uniquement au coefficient de détermination, on constate que quelle que soit la localisation de l'apprentissage la qualité de la prédiction reste inchangée avec cet indicateur. La courbe 4-16 permet de visualiser l'allure de la simulation dans le cas de Bastia avec un apprentissage délocalisé. On s'aperçoit que le *PMC* arrive bien à approximer la mesure avec cependant une nette sous-estimation liée certainement au lissage entre les jours ensoleillés et nuageux.

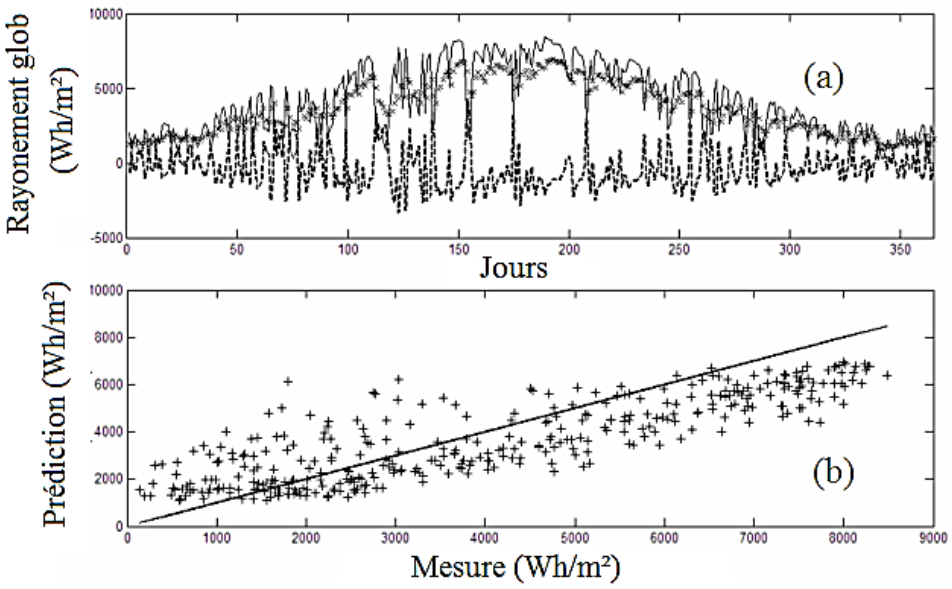


Figure 4-16 : impact de la délocalisation de l'apprentissage sur Bastia. (a), rayonnement global simulé et mesuré, les pointillés représentent l'erreur algébrique, le trait plein, les mesures et les croix les prédictions, (b) corrélation entre le rayonnement simulé et mesuré

Cette étude permet d'ouvrir la possibilité d'utiliser le *PMC* sur un site où aucune mesure n'est disponible. Le fondement des *ANN*, et plus particulièrement des *PMC* les rendent robustes. On connaissait cette robustesse vis-à-vis du bruit, de la non-linéarité ou encore de la non-stationnarité dans les données, on découvre grâce à cette étude qu'elle est aussi relative à la non cohérence spatiale entre apprentissage et test. Cet élément renforce l'idée que le *PMC* est un bon candidat pour l'estimation du rayonnement global ou de l'énergie *PV*. D'autres manipulations (comme la prédiction horaire) ont été effectuées sur cette thématique de délocalisation. Cependant, pour faciliter la lecture, dans la suite de ce manuscrit nous n'exposerons pas les résultats de ces expériences, le lecteur intéressé peut se référer à [Voyant et al. 2009].

2.5. Conclusion

Cette première étude sur l'horizon journalier a permis de bien comprendre comment utiliser les *PMC* et les autres prédicteurs étudiés. Nous avons montré que les prédicteurs testés de type Markov, Bayes et *k-NN* sont relativement équivalents en terme de prédiction. Le détail de cette comparaison est donné dans [Paoli et al. 2010]. Ces prédicteurs se sont avérés

beaucoup moins adaptés à la prédiction du rayonnement global que *ARMA* ou encore les *PMC*. Dans la suite, nous nous limiterons à l'étude de ces derniers. Concernant les estimateurs naïfs (ciel clair, moyenne et persistance), par la suite nous nous focaliserons sur un seul afin de ne pas alourdir les résultats. Seule la persistance sera utilisée, pour sa facilité d'utilisation et ses bons résultats notamment les jours de faible nébulosité (*nRMSE* = 19 % lors de la prédiction de l'énergie du champ *PV* aux mois de mai et juin). Nous avons également mis en évidence, que pour prédire correctement la ressource, il est nécessaire d'effectuer un prétraitement des séries temporelles afin de les stationnariser. Cette étape est primordiale et permet d'améliorer la qualité des modélisations *PMC* et *ARMA* (gain de presque 1 % lors de la prédiction de l'énergie issue du champ *PV* durant la période janvier-juin). La méthode de stationnarisation utilisée n'induit que peu de changement. Il y a un léger avantage quant à l'utilisation du *CSI* et des *CP* (nommé CSI^*). Le détail de cette série de manipulation est référencé dans [Paoli et al. 2010]. En analyse univariée, aucune différence significative n'a été démontrée entre *PMC* et *ARMA* tant que les deux modèles subissent le même prétraitement. Concernant l'analyse multivariée possible avec le *PMC*, les résultats sont assez probants et permettent un gain oscillant entre 0,4 % et 1 % (*nRMSE*) suivant la localisation étudiée. Il faudra en tenir compte dans les prochaines analyses. Tous les résultats relatifs à l'utilisation des variables exogènes sont synthétisés dans [Voyant et al. 2010] et [Voyant et al. 2011a]. Les manipulations sur les données énergétiques issues des modules *PV*, nous permettent de considérer que le bilan dressé dans le cas de la prédiction de rayonnement global horizontal, reste tout aussi valable dans le cas de cellules *PV* et de prédiction d'énergie *DC* (ou *AC*). Les données du mur *PV* ne possédant pas assez d'historique, la prédiction d'énergie, comme nous l'avons présenté ici, est équivalente à une prédiction de rayonnement global. Dans la suite, nous ne retraiterons plus le cas *PV* du mur, sauf pour l'horizon m+5. En Effet, une année de mesure représente dans le cas journalier 365 points, dans le cas horaire 3285 (365x9), et dans le cas 5 minutes 39420 (365x9x12) ; en première approximation, on peut considérer que plus l'horizon diminue et moins l'apprentissage nécessite de longues périodes de mesures. La localisation géographique influe fortement sur les résultats de prédiction, c'est ainsi que dans les manipulations présentées les *nRMSE* propre à Ajaccio et à Bastia différaient de quelques pourcents. A ce titre, dans la suite du manuscrit, nous essayerons, aussi souvent que possible, de dresser des résultats relatifs à différentes localisations. L'ultime élément qui a été développé dans cette section concerne la délocalisation de l'apprentissage. Nous avons montré que notre méthodologie de prédiction basée sur l'emploi d'un prétraitement et d'une modélisation connexionniste est relativement

robuste au point de permettre de compenser la perte d'information engendrée par la non-cohérence spatiale entre test et apprentissage. Concernant les critères de comparaison, le *MBE* et l'erreur sur le cumul sont assez similaires dans leur interprétation. De plus, comme le cumul et l'erreur de tendance n'ont pas été discriminants, dans la suite nous n'utiliserons plus ces critères.

Toutes les manipulations et les conclusions que nous venons d'exposer, vont être reprises et adaptées pour chaque horizon de prédiction. La prochaine section consistera à étudier l'horizon horaire de prédiction, qui de par la double périodicité des signaux (horaire et journalier) est a priori plus difficile à appréhender.

3. Le cas horaire : *h+1*

La possibilité de prédire le rayonnement global une heure à l'avance est un enjeu pour la gestion de la production électrique d'un réseau mixte composé de moyens de production conventionnels et intermittents (voir chapitre d'introduction, section 3-2, p 24). Cet horizon correspond notamment au temps nécessaire pour alimenter le réseau en énergie issue des moteurs thermiques (~50 % de la puissance totale installée en Corse). Dans cette section, nous allons appliquer les méthodes précédemment décrites pour le cas journalier. Nous verrons que, outre la nécessité de perfectionner encore d'avantage cette approche, le cas horaire se prête relativement bien à l'utilisation d'une méthodologie nouvelle basée sur une approche hybride des prédicteurs. Les différentes manipulations que nous verrons sont relatives à plusieurs sites. Le fait de proposer ce type d'analyse permet de vérifier si la méthodologie est « site-dépendant » ou si elle peut être appliquée à d'autres environnements. Dans la dernière partie nous utiliserons les résultats du modèle numérique de prévision météorologique (modèle *ALADIN* de Météo-France) en entrée de notre estimateur de rayonnement global pour en observer les bénéfices éventuels.

3.1. Prédiction du rayonnement global horizontal

Durant cette étude, nous avons souhaité disposer de dix années communes de mesures issues des cinq sites présentés dans le chapitre 1 (section 5-1, p 55). Ces derniers sont situés dans la zone méditerranéenne, huit années ont servi à l'apprentissage (1998-2005) et deux

années aux tests de prédiction (2006-2007). Outre les mesures de rayonnement global horizontal, nous avons utilisé trois jeux de mesures de variables dites exogènes représentant les grandeurs de pression, nébulosité et précipitation (chapitre 1 section 5-1, p 55). Nous les avons sélectionnés d'après les résultats concernant les variables exogènes de la section précédente. Le phénomène de non-stationnarité est, comme dans le cas précédent, présent au sein de la série de rayonnement global mesuré. Pour s'affranchir de ce dernier, nous avons utilisé l'indice de ciel clair, l'indice de clarté et les coefficients périodiques comme présentés dans le chapitre 3 section 3-3. La Figure 4-17 montre l'impact que peut avoir la stationnarisation sur l'autocorrélogramme de la série temporelle de rayonnement global. Nous rappelons que les heures de nuit ne sont pas prises en compte lors des simulations. Pour le cas horaire, nous considérons neuf mesures par jour.

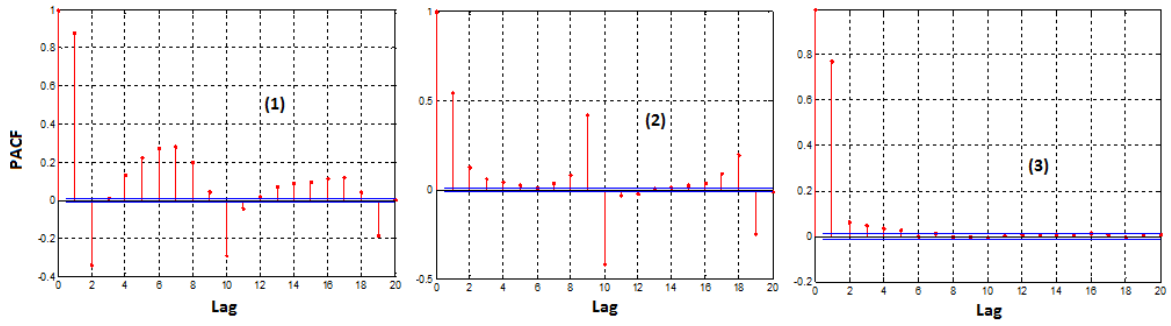


Figure 4-17 : autocorrélogramme de la série temporelle de rayonnement global (Ajaccio), (1) série originale, (2) *CSI*, et (3) CSI^*

On s'aperçoit que le rapport à la tendance par ciel clair (appelé *CSI*) tend à diminuer la périodicité journalière, mais ne s'en affranchit pas totalement (pic à t-9). L'utilisation des coefficients périodiques (noté *CSI**) améliore nettement le phénomène. Nous vérifierons, plus tard dans cette section, si cette amélioration de la stationnarité se répercute aussi sur la qualité de la prédiction.

Tout comme au paragraphe précédent, notre objectif sera de comparer notre méthodologie à base de *PMC* avec la méthode *ARMA* considérée comme référence. On peut voir sur le Tableau 4-13 un exemple d'optimisation de ce genre de modèle. Sur les cinq sites étudiés et concernant la prédiction de l'indice ciel clair, on remarque que les modèles les plus performants sont relativement simples et ne comportent pas de partie en moyenne mobile (q = 0).

Station	*ARMA*(p,q)	φ_1	φ_2
Ajaccio	*ARMA* (1,0)	0,5435	0
Bastia	*ARMA* (1,0)	0,5838	0
Montpellier	*ARMA* (1,0)	0,5178	0
Marseille	*ARMA* (2,0)	0,4176	0,1350
Nice	*ARMA* (1,0)	0,5248	0

Tableau 4-13 : modèles *ARMA* optimisés pour la prédiction du *CSI* (centré réduit)

Pour les modèles de prédiction, les huit années d'entrainement et les deux de test correspondent à 80 % et 20 % de l'ensemble des données. L'utilisation de la toolbox *RNA* de Matlab™ impose de décomposer la phase que nous appelons entraînement en trois sous-phases : training ; validation ; test. Les répartitions utilisées durant les manipulations pour ces trois sous-phases sont respectivement : 90 % ; 10 % ; 0 %. Le bilan total de la répartition entre apprentissage et test est : 72 % d'entrainement (au sens « training Matlab™ » concernant l'optimisation des poids), 8 % de validation (au sens « training Matlab™ » pour stopper l'apprentissage et éviter la divergence) et 20% de test. La sélection des variables d'entrée est réalisée avec la méthode des *PACF* et des coefficients de corrélation de Pearson (chapitre 3, section 3-2). Pour ces derniers, l'utilisation d'un test de type Student pour connaître la significativité des corrélations, renvoie à des valeurs trop faibles, annihilant toute sélection. Dans ce cas précis, cette méthode n'est pas appropriée. La limite doit être plus élevée afin de sélectionner un nombre restreint de variables exogènes ; plus le nombre de neurones d'entrée est élevé, plus le *PMC* se complexifie et moins les résultats sont cohérents (principe de parcimonie inhérent à l'utilisation des *PMC*). Dans cette logique et après une étude de sensibilité, nous avons délibérément choisi une limite de 50 % pour la nébulosité et 15 % pour les autres paramètres. Pour la nébulosité, la valeur de 50 % correspond en moyenne à la valeur du *PACF* au premier lag pour le *CSI*. Ainsi on considère que si la corrélation est supérieure à 50 % (~ autocorrélation au lag 1), on aura des informations supplémentaires, qui permettront de mieux prédire le rayonnement global. Concernant la pression et la nébulosité, avec une valeur supérieure à 15 %, on aurait trop peu de variables sélectionnables (voire aucune), avec une valeur inférieure il y en aurait trop. Ce critère de sélection (*R*) laisse une part de subjectivité non négligeable qui peut déranger, c'est pourquoi nous avons travaillé en parallèle sur une autre méthode, plus objective (présentée au chapitre 3, section 3-2c et

utilisée dans ce chapitre, section 3-3b). Le tableau 4-14 montre le résultat de l'optimisation pour les deux types de *PMC* à disposition (avec ou sans utilisation de données exogènes).

Station	Architecture endo-exogène	Architecture endogène
Ajaccio	$(Endo^{1\text{-}10},N^{1,2},P^{1},RP^{1})$x15x1	$(Endo^{1\text{-}10})$x15x1
Bastia	$(Endo^{1\text{-}10},N^{1,2},P^{1},RP^{1})$x15x1	$(Endo^{1\text{-}10})$x10x1
Montpellier	$(Endo^{1\text{-}10},N^{1,2},P^{1})$x10x1	$(Endo^{1\text{-}10})$x10x1
Marseille	$(Endo^{1\text{-}10},N^{1}, P^{1})$x15x1	$(Endo^{1\text{-}10})$x10x1
Nice	$(Endo^{1\text{-}10},N^{1}, RP^{1})$x15x1	$(Endo^{1\text{-}10})$x10x1

Tableau 4-14 : configurations *PMC* optimisées en utilisant des variables endogènes et exogènes avec la stationnarisation *CSI. N*omenclature : $(Endo^{Ne*},Exo^{Me*})xN_c x1$

Après optimisation de chaque ville pour la stationnarisation *CSI*, on s'aperçoit que le nombre de lags endogènes est toujours égal à dix et le nombre de neurones cachés toujours supérieur ou égal à dix. Concernant les variables exogènes, il y en a, en moyenne, trois par *PMC*. Ce qui correspond à environ 30 % du nombre de variable endogènes. Une autre étape importante de l'optimisation consiste à tester l'impact de l'utilisation de *CSI*, *k* et CSI^{*} sur la qualité de la prédiction. Il faut, tout d'abord, quantifier l'effet des prétraitements sur l'ensemble des autocorrélations, corrélations et donc sur la sélection des variables d'entrée. Le Tableau 4-15 montre l'influence de la stationnarisation sur l'architecture du *PMC*, le coefficient de variation et la qualité de prédiction. Nous pouvons remarquer que la méthodologie *CSI* minimise systématiquement l'erreur de prédiction, suivi de *k* et enfin *CSI**. En fait, dans le cas traité ici, la stationnarité ne semble pas affecter énormément les résultats issus des *PMC*, le gain est de 0,5 % en moyenne. De plus, dans cet exemple, on constate que l'interprétation du *CV* n'est pas évidente car il est non corrélé avec l'erreur de prédiction. On peut cependant penser qu'une très haute valeur induit une erreur de prédiction tout aussi élevée, mais l'opposé n'est pas forcément démontré dans cette étude.

Station Lat/Long/Alt	Stationnarité	*ANN* Architecture	*CV*	*nRMSE (%)*
Ajaccio 41,6°/8,5°/4m	aucune	$(Endo^{1-10},N^{1,2})x15x1$	0,61	17,1
	k	$(Endo^{1-10}, N^{1-3},P^{1})x15x1$	0,44	16,9
	CSI	$(Endo^{1-10}, N^{1,2},P^{1},RP^{1})x15x1$	0,46	**16,7**
	*CSI**	$(Endo^{1}, N^{1,2},P^{1},RP^{1})x5x1$	**0,38**	17,2
Bastia 42,9°/9,5°/10m	aucune	$(Endo^{1-10}, N^{1,2})x10x1$	0,64	20,1
	k	$(Endo^{1-10}, N^{1-3})x15x1$	0,46	19,9
	CSI	$(Endo^{1-10},N^{1,2},P^{1},RP^{1})x15x$	0,48	**19,6**
	*CSI**	$(Endo^{1}, N^{1,2})x3x1$	**0,44**	20,0
Montpellier 43,6°/3,9°/2m	aucune	$(Endo^{1-10},N^{1})x15x1$	0,63	17,0
	k	$(Endo^{1-10}, N^{1-3})x15x1$	0,45	16,5
	CSI	$(Endo^{1-10}, N^{1,2},P^{1})x10x1$	0,49	**16,4**
	*CSI**	$(Endo^{1}, N^{1,2},P^{1})x5x1$	**0,40**	16,8
Marseille 43,4°/5,2°/5m	aucune	$(Endo^{1-10},N^{1})x10x1$	0,60	15,5
	k	$(Endo^{1-10}, N^{1,2},P^{1})x15x1$	0,42	**14,9**
	CSI	$(Endo^{1-10},N^{1}, P^{1})x15x1$	0,45	**14,9**
	*CSI**	$(Endo^{1},N^{1},P^{1})x8x1$	**0,35**	15,1
Nice 43,6°/7,2°/2m	aucune	$(Endo^{1-10}, N^{1})x15x1$	0,63	16,9
	k	$(Endo^{1-10}, N^{1,2})x10x1$	0,44	16,9
	CSI	$(Endo^{1},N^{1}, RP^{1})x15x1$	0,46	**16,6**
	*CSI**	$(Endo^{1-10},N^{1}, RP^{1})x3x1$	**0,38**	17,2

Tableau 4-15 : impact des méthodes de stationnarisation *k, CSI et CSI** sur les prédictions

L'utilisation de *CSI** simplifie énormément l'architecture des *PMC* (un seul neurone endogène et un maximum de 8 neurones cachés), mais comme cela n'améliore pas l'erreur de prédiction nous n'utiliserons plus les coefficients périodiques. En effectuant la même étude dans le cas des prédictions *ARMA*, les modes de stationnarisation *CSI* et *CSI** donnent des résultats similaires. Dorénavant, nous n'utiliserons donc que le *CSI* avec l'utilisation de ces prédicteurs. Le Tableau 4-16 dresse la comparaison des *nRMSE* saisonniers avec les estimateurs de rayonnement global (persistance, ciel clair, moyenne, *ARMA* et *PMC*) pour

chacune des cinq villes. Pour les prédictions *PMC*, nous étudierons le cas avec uniquement les variables endogènes (*PMC* endo) et celui mêlant variables endogènes et exogènes (*PMC* endo-exo).

Station	Modèle	Annuel	Hiver	Printemps	Eté	Automne
Ajaccio	Persistance	35,3	55,9	32,2	32,3	36,9
	Ciel clair	41,8	66,9	39,4	37,9	43,2
	Moyenne	32,5	51,3	29,9	28,6	33,9
	ARMA	16,9	25,7	**15,9**	**14,4**	19,3
	PMC endo	17,1	**23,8**	16,3	15,8	18,3
	PMC endo-exo	**16,7**	**23,8**	16,3	15,8	**17,6**
Bastia	Persistance	39,6	50,2	41,9	34,5	37,7
	Ciel clair	45,9	62,4	48,7	38,9	42,5
	Moyenne	35,7	48,1	37,3	29,8	36,5
	ARMA	**17,7**	23,4	**18,2**	**15,7**	**18,8**
	PMC endo	19,4	22,5	21,5	17,2	19,9
	PMC endo-exo	19,6	**22,4**	21,4	17,6	20,1
Montpellier	Persistance	37,3	44,1	36,0	30,5	47,2
	Ciel clair	43,2	55,9	36,8	34,0	61,3
	Moyenne	34,9	41,9	32,6	27,9	47,8
	ARMA	**16,0**	20,1	**14,4**	**13,3**	20,5
	PMC endo	16,5	18,6	14,6	15,0	18,3
	PMC endo-exo	16,4	**18,3**	14,7	15,1	**17,8**
Marseille	Persistance	32,7	40,1	31,9	23,5	43,7
	Ciel clair	40,0	56,0	38,2	31,6	49,7
	Moyenne	29,9	39,9	28,9	21,3	39,7
	ARMA	15,4	21,4	14,8	**12,5**	19,7
	PMC endo	15,2	**18,3**	14,5	13,3	18,1
	PMC endo-exo	**14,9**	**18,3**	**14,4**	12,8	**17,0**
Nice	Persistance	35,2	41,6	39,4	29,2	41,8
	Ciel clair	40,8	51,5	41,7	33,2	48,3
	Moyenne	31,7	40,2	33,6	27,1	38,4
	ARMA	**16,2**	20,6	**16,2**	**14,4**	18,6
	PMC endo	16,8	18,5	18,3	16,7	15,9
	PMC endo-exo	16,6	**18,2**	17,7	16,3	**15,0**

Tableau 4-16 : comparaison de performance (*nRMSE* en %) des modèles persistance, ciel clair, simple moyenne, *ARMA*, *PMC* avec variables endogènes et *PMC* avec variables endogènes et exogènes. Les chiffres en gras représentent les *nRMSE* les plus faibles

En été, sur les sites que nous avons sélectionnés, l'intérêt d'utiliser les méthodes de type *PMC* endo et *PMC* endo-exo est minime. Ce phénomène est certainement lié à la faible probabilité d'occurrence de nuages durant cette période. Un processus linéaire comme *ARMA* semble alors le mieux adapté. Nous pouvons vraisemblablement conclure que l'utilisation du *PMC* avec variables endogènes et exogènes est la plus intéressante durant les périodes où la couverture nuageuse est intense (essentiellement en automne et en hiver). De plus, Bastia, Nice et Montpellier sont des villes où l'erreur annuelle est très faible avec l'utilisation d'une autorégression (respectivement *nRMSE* = 17,7 %, *nRMSE* = 16,2 % et *nRMSE* = 16 %).

Nous venons de discuter qualitativement du lien entre saison et type de prédicteur. Il semble intéressant d'approfondir cette étude, et de vérifier quantitativement s'il existe des saisons où certains prédicteurs sont plus performants que d'autres. Pour cela, nous avons imaginé une méthode qui consiste à classer l'efficacité des prédicteur selon les saisons. Le classement est basé sur une analyse monocritère de l'erreur saisonnière pour chaque prédicteur. Cette analyse est une aide à la décision permettant de trouver le modèle le plus performant pour chaque saison. Notre postulat consiste à considérer que le meilleur prédicteur est celui qui minimise le critère *nRMSE*. *A contrario*, le plus mauvais sera celui qui le maximise. Le système de notation est définit comme suit : 1 point pour le meilleur prédicteur, 2 points pour le second, …, 6 points pour le moins performant. Nous avons posé que l'erreur de prédiction produit les mêmes effets pour toutes les saisons (isopondération des saisons), cependant, le but de l'étude n'est pas de comparer les saisons entre elles. L'opération est répétée pour les cinq villes et pour chaque saison comme le montre le Tableau 4-17. Le meilleur prédicteur est celui qui minimise son score.

Modèles	Hiver	Printemps	Eté	Automne
PMC endo-exo	**1er (5 pts)**	2 ième (10 pts)	2ième (12 pts)	**1er (7 pts)**
PMC endo	2ième (8 pts)	3ième (12 pts)	2ième (12 pts)	2ième (10 pts)
ARMA	3ième (15 pts)	**1er (7 pts)**	**1er (5 pts)**	3ième (13 pts)
Moyenne	4ième (20 pts)	4ième (20 pts)	4ième (20 pts)	4ième (20 pts)
Persistance	5ième (25 pts)	5ième (25 pts)	5ième (25 pts)	5ième (25 pts)
Ciel clair	6ième (30 pts)	6ième (30 pts)	6ième (30 pts)	6ième (30 pts)

Tableau 4-17 : classification et score des prédicteurs

Le score peut osciller entre 5 points et 30 points. Le modèle ciel clair, la persistance et la moyenne sont, pour les cinq villes, les méthodes les moins efficaces. Durant les saisons très nuageuses (hiver et automne), le *PMC* couplé avec des variables endogènes et exogènes (PMC endo-exo) donne les meilleurs résultats. Concernant les deux autres saisons, il semble que *ARMA* soit le plus intéressant. Comme aucun prédicteur ne fait l'unanimité, il semble opportun de se tourner vers un type de prédicteur plus à même d'estimer la ressource horaire. Les résultats saisonniers concernant *ARMA* et *PMC* suggèrent de mixer les deux approches au sein d'une seule plus générique que l'on peut qualifier d'hybride. Ce type d'approche est décrite dans la littérature et donne de bons résultats (chapitre 3, section 4, p 129).

3.2. Approche hybride

Pour construire nos modèles hybrides, nous avons utilisé le tableau 4-17. Sur ce dernier, nous avions remarqué qu'il était opportun d'utiliser le *PMC* en automne et en hiver, et *ARMA* pour les deux autres saisons. Ce constat conduit directement au premier modèle testé, nommé modèle A. Outre cette approche, nous avons aussi choisi d'en tester deux autres. La première est identique à la précédente, mais l'apprentissage (*PMC* et *ARMA*) n'est réalisé que sur les mois où *PMC* ou *ARMA* sont les plus efficaces (été et printemps pour *ARMA* et le reste de l'année pour *PMC*), c'est le modèle B. Le dernier modèle n'exploite pas directement le Tableau 4-17. Il découle du constat que le *PMC* est performant quand il y a une forte occurrence nuageuse et *ARMA* dans le cas contraire. Plutôt que de fixer une règle de sélection de modèle basée sur une saisonnalité rigide, le dernier modèle repose sur l'étude d'une saisonnalité souple, non prédéterminée. Ainsi, si les prédictions *PMC* sont performantes, on gardera ce prédicteur pour les pas suivant, sinon on en change. Cette méthode est appelé le modèle C. Les modèles *ARMA* et *PMC* utilisés avec ces trois méthodes hybrides, correspondent, pour chaque ville et pour la méthode *CSI*, aux modèles optimisés séparément avec utilisation des variables exogènes (dans le cas *PMC*, voir Tableau 4-15). Le récapitulatif des trois modèles est le suivant :

- modèle A : on utilise les prédictions issus du *PMC* en automne et hiver, puis les prédictions *ARMA* pour les deux autres saisons, la sélection est représentée par :

$$si\ t\ \in \{prin\ ou\ été\}\ alors\ \ \hat{X}(t+1) = \hat{X}^{AR}(t+1)\ sinon\ \hat{X}(t+1) = \hat{X}^{PMC}(t+1) \quad (4\text{-}3)$$

- modèle B : identique au modèle précédent, mais l'apprentissage (*PMC* et *ARMA*) n'est réalisé que sur les mois où *PMC* ou *ARMA* sont les plus efficaces (été et printemps pour *ARMA* et le reste de l'année pour *PMC*).

$$si\ t\ \in \{prin\ ou\ été\}\ alors\ \hat{X}(t+1) = \hat{X}^{AR}_{prin/été}(t+1)\ sinon\ \hat{X}(t+1) = \hat{X}^{PMC}_{aut/hiv}(t+1) \quad (4\text{-}4)$$

- modèle C : il repose sur l'erreur de prédiction du pas d'avant, ainsi si *PMC* était meilleur au pas *t-1*, on utilisera *PMC* au pas *t*, ou *ARMA* dans le cas inverse.

$$si\ |\varepsilon^{AR}(t)|\ \leq\ |\varepsilon^{PMC}(t)|\ alors\ \hat{X}(t+1) = \hat{X}^{AR}(t+1)\ sinon\ \hat{X}(t+1) = \hat{X}^{PMC}(t+1) \quad (4\text{-}5)$$

Où $\hat{X}^{AR}(t)$ et $\hat{X}^{PMC}(t)$ sont les prédictions faites avec *AR* et *PMC* pour le rayonnement global au temps *t* et $\hat{X}^{AR}_{prin/été}(t)$ la prédiction réalisée avec *AR* pour le rayonnement global au temps *t* et avec un espace d'entrainement limité aux saisons d'été et de printemps. Concernant $\hat{X}^{PMC}_{aut/hiv}(t)$, ce paramètre représente la prédiction réalisée avec *PMC* pour le rayonnement global au temps *t* et avec un espace d'entrainement limité aux saisons d'hiver et d'automne. La Figure 4-18 synthétise l'approche hybride des trois modèles.

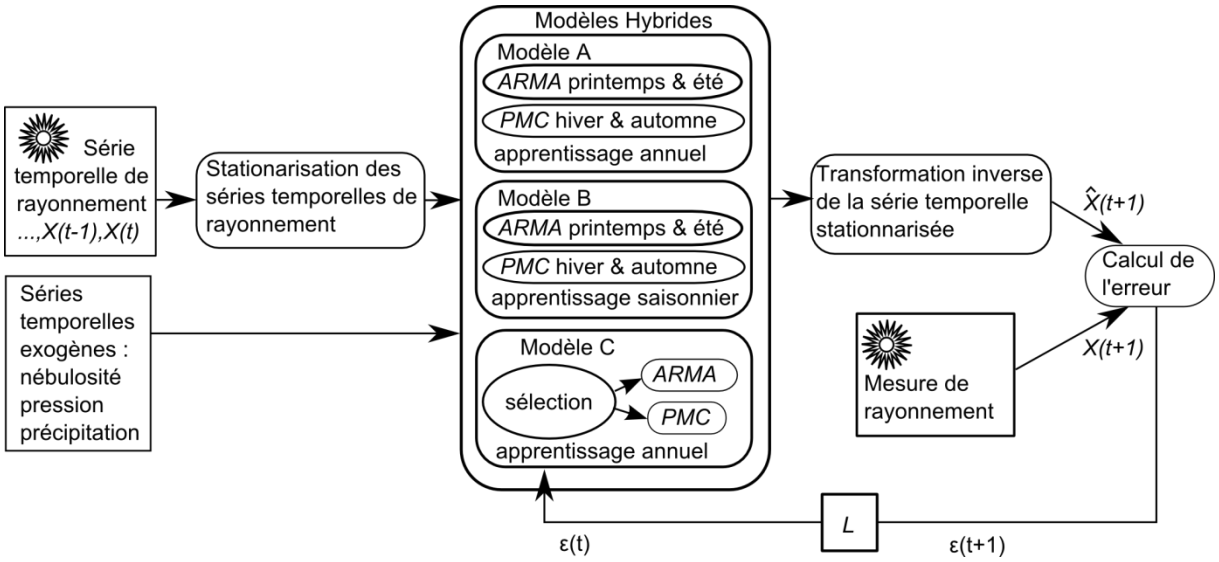


Figure 4-18 : schéma de la méthodologie hybride de prédiction

Sur le Tableau 4-18 sont représentés les résultats obtenus dans le cadre de l'hybridation des modèles. Les lignes nommées « *NH* (signifiant non hybrides) » montrent les meilleurs

résultats obtenus avec des prédicteurs classiques *PMC* ou *ARMA*. Ces données sont issues du Tableau 4-16.

Station	Modèle	Annuel	Hiver	Printemps	Eté	Automne
Ajaccio	A	15,9	23,8	15,9	14,4	17,6
	B	16,4	24,3	15,9	14,4	17,6
	C	**15,3**	**21,1**	**15,0**	**14,3**	**16,0**
	NH	*16,7*	*23,8*	*15,9*	*14,4*	*17,6*
Bastia	A	**17,7**	23,4	**18,2**	**15,7**	18,8
	B	17,9	22,4	**18,2**	**15,7**	20,0
	C	**17,7**	**21,3**	19,0	16,1	**17,9**
	NH	***17,7***	*22,4*	***18,2***	***15,7***	*18,8*
Montpellier	A	15,5	18,3	14,4	**13,3**	17,8
	B	15,7	19,1	14,5	13,4	18,4
	C	**15,1**	**17,2**	**13,5**	13,4	**17,1**
	NH	*16,0*	*18,3*	*14,4*	***13,3***	*17,8*
Marseille	A	14,7	18,3	14,8	12,5	17,0
	B	15,0	19,4	14,8	12,4	18,7
	C	**13,7**	**17,5**	**13,2**	**11,5**	**16,0**
	NH	*14,9*	*18,3*	*14,4*	*12,5*	*17,0*
Nice	A	15,3	18,2	**16,2**	**14,4**	15,0
	B	15,7	19,5	**16,2**	**14,4**	16,2
	C	**15,1**	**16,2**	16,3	14,9	**14,2**
	NH	*16,8*	*18,2*	***16,2***	***14,4***	*15,0*

Tableau 4-18: comparaison des performances (*nRMSE* en %) des trois modèles hybrides étudiés. Les chiffres en gras représentent les erreurs de prédiction minimales

L'erreur annuelle est nettement minimisée dès que l'on utilise le modèle C (sélection *PMC-ARMA*), il semble que ce modèle soit le plus performant des trois étudiés. Il faut aussi remarquer que, hormis en été où elle est équivalente, la qualité de la prédiction est très significativement accrue et ce, quel que soit le modèle hybride utilisé. Concernant le cas B (apprentissage saisonnier), il y a un biais dans l'analyse car l'espace d'apprentissage est divisé par deux, ce qui entraîne un déficit de données durant l'apprentissage très certainement responsable des valeurs de *nRMSE* observées. Pour le modèle C, qui est le plus pertinent, en moyenne 1/3 des prédictions sont réalisées avec *ARMA* et 2/3 avec *PMC*. Cette méthode permet un gain moyen de *nRMSE* supérieur à 1 % vis-à-vis des méthodes non-hybrides. La Figure 4-19 montre ce gain moyen (calculé sur les cinq villes) de *nRMSE* obtenu par

l'utilisation des modèles hybrides (C) comparé aux meilleurs *PMC* (barres grises) et aux meilleurs *ARMA* (barres blanches barrées) de chaque station. Le gain est positif, quand l'hybridation est meilleure que les méthodes classiques.

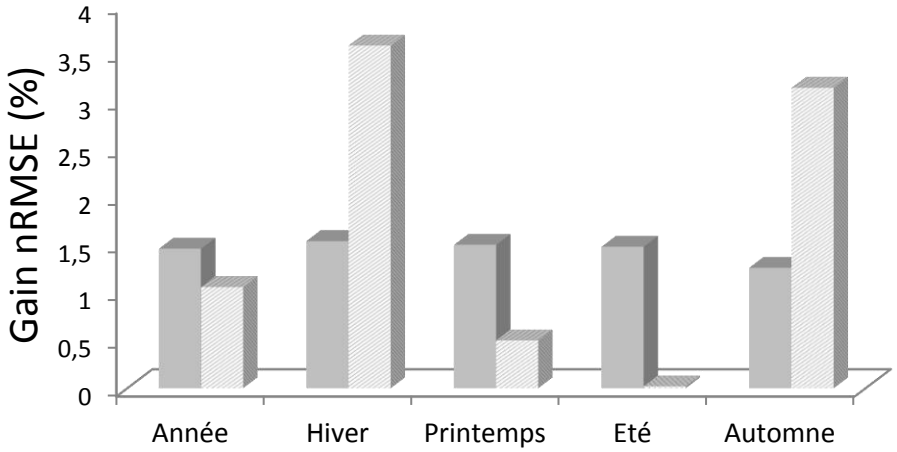
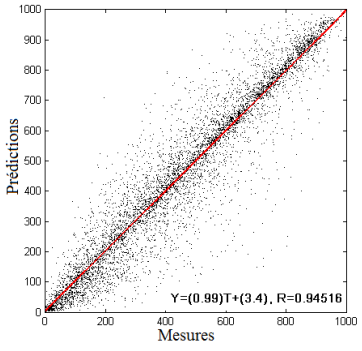
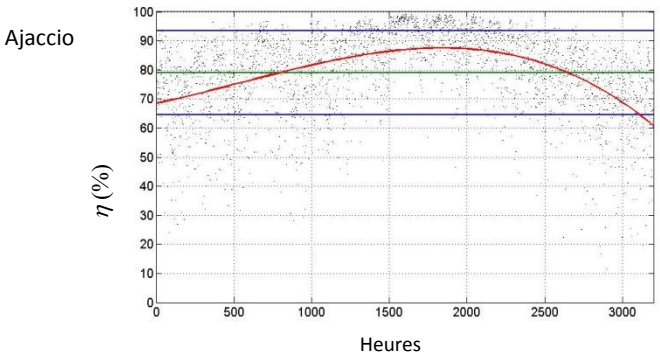


Figure 4-19 : gain moyen lié à l'utilisation du modèle hybride C comparé aux modèles *PMC* (barres grises) et *ARMA* (barres blanches barrées)

Le gain maximum est observé en hiver (3,8 % de mieux que le modèle *ARMA*) et le minimum est en été, quand la méthode hybride est aussi intéressante que la méthode *ARMA* (0,02 % de gain).

Sur la Figure 4-20, la partie de gauche, montre quant à elle, la comparaison entre la mesure et la prédiction réalisée dans le cas du modèle hybride *C* (modèle avec prise en compte de l'erreur de prédiction passée) pour le rayonnement global horaire horizontal (graphique de type y = x). Pour tous les sites, il est assez clair que le modèle hybride C approxime correctement la grandeur recherchée. Le coefficient de détermination est compris entre 0,92 et 0,95 pour les cinq sites étudiés.





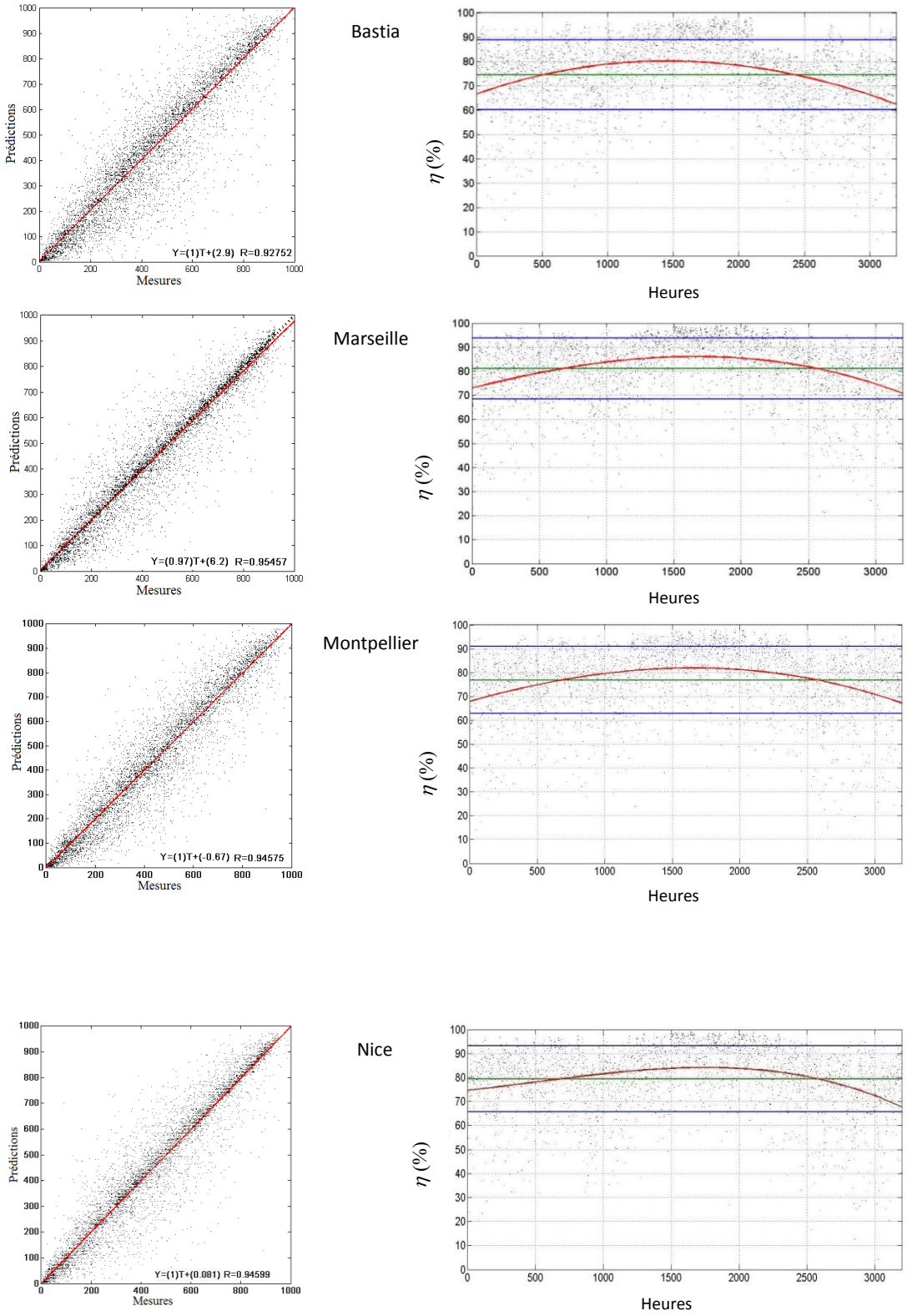


Figure 4-20 : comparaison entre mesures et prédictions obtenues avec le modèle hybride C (*PMC/ARMA*) (à gauche) et indices de similarité relatifs (à droite ; ligne verte = moyenne annuelle ; lignes bleues = moyenne ± écart type ; ligne rouge = tendances polynomiales)

Sur cette même figure, la partie de droite est dédiée à l'indice de similarité ou de fiabilité des mesures développé dans le chapitre 2 (section 4-1, p 91). Chaque prédiction réalisée avec la méthodologie hybride C est corrélée à cet indice $\eta(\tau)$ *:* 100 % signifie que le résultat est sûr et 0 % que l'erreur attendue est au moins aussi importante que la mesure elle-même. Le calcul est initié durant la phase d'entrainement du *PMC*. Durant la prédiction, le simulateur donne pour chaque heure deux paramètres : la prévision à h+1 et l'intervalle de confiance pour cette prévision. Les 3285 heures sur l'axe des abscisses des courbes de droite, correspondent à une année standard de prédiction ($1^{ière}$ heure = 1^{ier} janvier ; $1600^{ième}$ heure= 1^{ier} juillet et $3285^{ième}$ heure = 31 décembre). Ces paramètres sont identiques d'une année sur l'autre. La ligne verte représente la valeur moyenne, la ligne rouge la tendance, et les lignes bleues la moyenne ± l'écart type. On voit que c'est en été que la similarité maximale est atteinte (partie centrale des courbes) avec parfois 80 % de correspondance, et ce pour toutes les localisations. En hiver, la similarité peut-être inférieure à 70 %. La moyenne annuelle est d'environ 78 %, cependant certaines prédictions sont accompagnées d'une similarité très faible, essentiellement en hiver avec des valeurs inférieures à 30 %.

Sur la Figure 4-21, nous avons représenté la comparaison entre la prédiction réalisée avec la méthodologie hybride C (croix rouges) et la mesure (ligne noire) pour une période de 11 jours comprise entre juin et juillet 2006 sur le site d'Ajaccio. L'intervalle de confiance (comme défini dans le chapitre 3, section 4-1, p 91) est calculé à partir de l'indice de similarité. Cet intervalle accompagne chaque prédiction, il représente la fourchette dans laquelle la mesure devrait se situer.

La première partie de la courbe (les quatre premières journées) est typiquement le profil de journées ensoleillées sans nuage (ciel clair). On observe que le modèle de prédiction est très fiable. Quelques journées sont pourvues de très faibles valeurs de rayonnement global (valeur inférieure à 400 Wh/m² indiquées par une flèche sur la Figure), mais la prédiction avec la méthodologie hybride semble tout de même efficace. En règle générale, au niveau de la partie centrale et de la dernière partie de la courbe, les profils coïncident avec des journées plus nuageuses. Les allures des profils journaliers sont moins régulières, cependant l'approche hybride semble tout de même très probante.

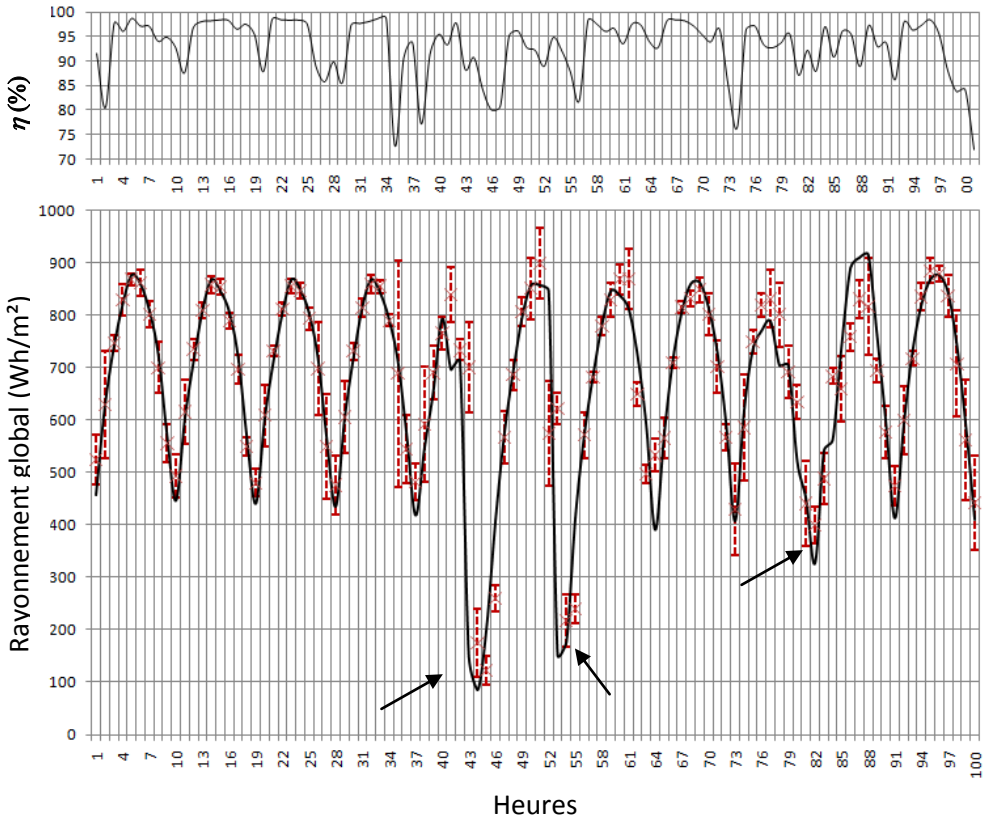


Figure 4-21 : comparaison entre la prédiction (croix rouges avec intervalle de confiance) et mesures (ligne noire). La fiabilité correspondante à chaque mesure est donnée sur la courbe du haut (11 jours pris entre juin et juillet 2006, Ajaccio)

Il faut noter que généralement la prédiction ± *IC* inclut correctement la mesure, mais certaines fois l'erreur est très importante (exemple des 41^{ième} et 43^{ième} heures). C'est la conséquence de la méthodologie de calcul de *l'IC* qui est basée sur une erreur de prédiction moyenne (durant l'apprentissage). Les occurrences « extraordinaires » (tant en fréquence qu'en intensité) ne sont pas bien prises en compte dans le modèle. Pour essayer de pallier à cet inconvénient, nous avons imaginé utiliser en complément de notre approche hybride (*ARMA* et *PMC* endo-exo sur *CSI*), un modèle classique de prévision. Nous avons opté pour le modèle utilisé par Météo-France pour ses prévisions météorologiques.

3.3. Utilisation des données issues du modèle ALADIN

L'utilisation de variables exogènes comme définies dans le paragraphe précédent, contribue à améliorer la qualité des prédictions. Cependant il subsiste encore des périodes où le modèle hybride utilisé n'est pas assez réactif, et ne parvient pas à prédire les valeurs de

rayonnement global très faibles ou très élevées. Un des moyens qui permettrait peut-être de s'affranchir de cette non-réactivité, serait l'utilisation de variables exogènes prédites par un autre système. Avec ce processus on aurait, pour prédire le rayonnement au temps t+1, en entrée de *PMC*, non-pas les grandeurs météorologiques du temps *t*, mais celles estimées pour le temps t+1. La prévision de ces variables doit être élaborée avec un modèle fiable et de référence, c'est pourquoi nous avons opté pour le modèle *ALADIN* exploité et développé par un consortium qui porte le même nom, et auquel Météo-France appartient (Chapitre 1, section 5-2, p 58). Les données exploitées sont le rayonnement global horizontal sur les cinq sites précédemment étudiés au cours d'une période de plus de six ans (Octobre 2002 – Décembre 2008). En complément, nous avons utilisé les sorties du modèle *ALADIN* (pression, nébulosité, précipitation et température) sur la même période. Les quatre premières années ont servi à l'apprentissage et les deux dernières à la phase de test.

a. Etude de stationnarisation

Avant de commencer à énoncer les résultats relatifs au modèle hybride pourvu des prévisions *ALADIN*, nous allons développer un résultat qui semblait contradictoire lors des manipulations précédentes. En effet, nous avions vu précédemment que la stationnarité n'était pas forcément liée aux meilleurs résultats de prédiction. Ce constat laisse présager que le coefficient de variation (*CV*) n'est peut-être pas un bon critère de jugement de la stationnarité. Dans cette partie, nous allons essayer de mieux comprendre et décrire la stationnarisation. Nous allons utiliser le test de Fisher décrit dans le chapitre 3 section 3-3 (p 126) et qui permet d'apprécier la stationnarité d'une série temporelle. Le tableau 4-19 recense les valeurs des critères de stationnarité (*CSI* et le CSI^*) pour les cinq villes étudiées.

		Ajaccio	Bastia	Montpellier	Marseille	Nice
Série Brute	*CV*	-3,45	-2,48	-3,36	-3,62	-3,21
	$F_{c(annuel)}$	12,99	12,44	11,12	15,07	13,72
	$F_{c(quotidien)}$	6,81	**1,01**	7,02	6,53	7,27
CSI	*CV*	-0,40	-0,4	**-0,37**	**-0,38**	**-0,42**
	$F_{c(annuel)}$	3,25	3,05	3,39	4,08	3,44
	$F_{c(quotidien)}$	2,06	2,07	2,07	2,07	2,08
CSI*	*CV*	**-0,09**	**-0,08**	-1,42	-0,76	-1,2
	$F_{c(annuel)}$	**0,83**	**0,87**	**0,72**	**0,70**	**0,81**
	$F_{c(quotidien)}$	**2,05**	2,04	**1,41**	**1,80**	**1,52**

Tableau 4-19 : efficacité de la stationnarité induite par *CSI* et *CSI**. En gras sont représentées les valeurs des estimateurs de stationnarité les plus faibles pour chaque méthodologie et chaque ville

Les valeurs sont données après une normalisation du rayonnement global comprise entre -0,9 et 0,9. Cette normalisation correspond à l'opération réalisée juste avant l'apprentissage des *PMC*, visant à rendre compatible les données avec les fonctions de transferts de la couche cachée. En accord avec la table de Fisher nous obtenons un seuil de $F_{\infty,\infty}^{0,05} = 1$ dans le cas annuel et $F_{9,\infty}^{0,05} = 1{,}88$ dans le cas journalier. Le critère *CV* est certainement le moins intéressant car il ne permet pas de différencier les périodicités annuelles et journalières. Cependant il reste très facile à mettre en œuvre et rend compte d'une certaine variabilité au sein de la série. Ce paramètre est systématiquement supérieur (en valeur absolue) à un dans le cas de la série brute, et inférieur dans le cas *CSI* et *CSI**. Les deux autres paramètres montrent que *CSI** est le plus efficace. Il n'y a qu'un cas où F_c n'est pas minimum avec l'étude *CSI** (Bastia avec 2,04). De plus, avec cette méthodologie, huit des dix paramètres obtenus (deux par ville) sont inférieurs au seuil du Fisher, permettant de statuer sur la stationnarité. Notons que dans le cas de Bastia, le test de Fisher n'est certainement pas exploitable à cause du bruit contenu dans la série, corroborant le fait que sur ce site, le rayonnement global est très difficile à prédire. Sur la Figure 4-22, nous pouvons voir l'impact, sur l'allure de la série pour cette ville, du prétraitement *CSI* et *CSI**. Pour ces deux derniers, les nouvelles séries sont centrées et normalisées. Les résultats sont équivalents pour les quatre autres localisations mais ne sont pas présentés ici.

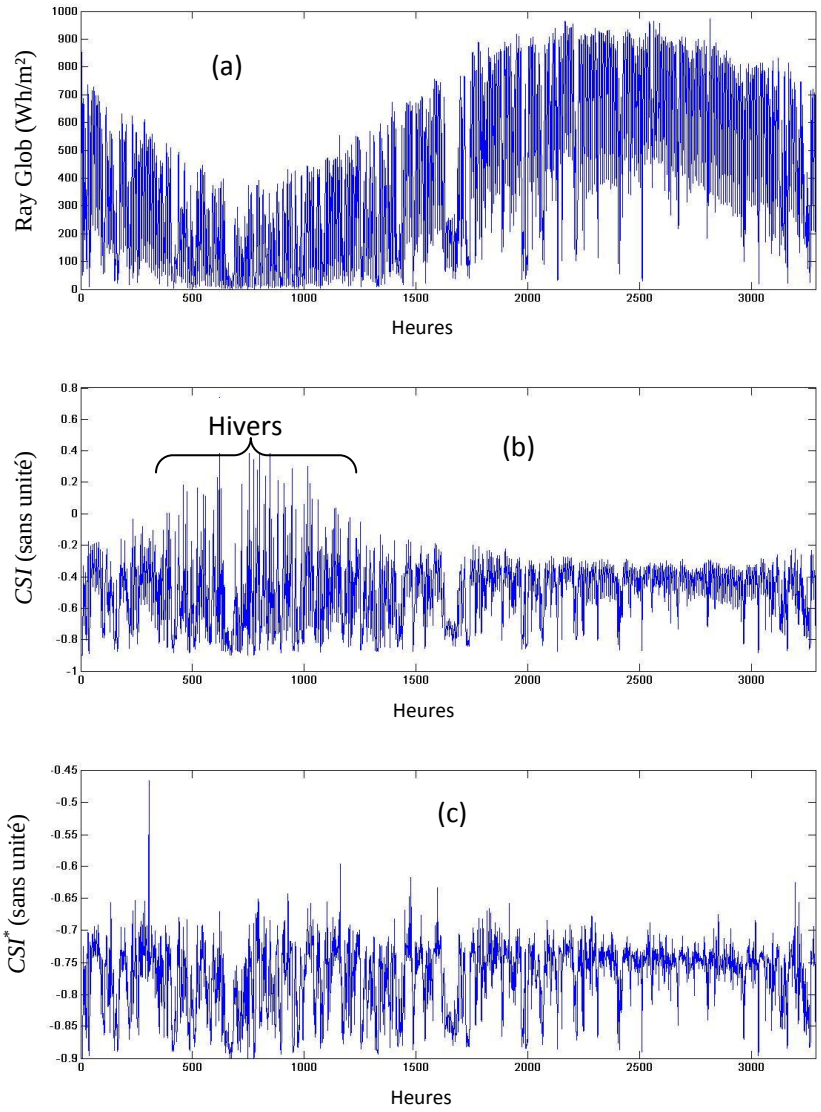


Figure 4-22 : effet de la stationnarisation sur la série temporelle de rayonnement global de Bastia (365 jours allant d'octobre 2002 à octobre 2003). (a) : série originale ; (b) : indice de ciel clair ; (c) : indice de ciel clair corrigé des variations saisonnières

Alors que le prétraitement CSI^* semble le plus intéressant, le *CSI* semble graphiquement donner le premier moment de la série constant (moyenne fixe sur l'année ~-0,4). Cependant, le second moment de la série reste périodique (variance beaucoup plus importante en hiver qu'en été). Nous pouvons voir que la courbe relative au processus CSI^* semble la plus stationnaire ou désaisonnalisée. Les critères développés dans cette section permettent de classifier les trois modes de stationnarisation en fonction des critères F_c. Nous vérifierons ultérieurement si cette hiérarchisation est respectée lors du test de prédiction sur les deux années 2007-2008.

b. Utilisations des prévisions *ALADIN*

Afin d'optimiser les différents modèles utilisés lors de ces manipulations (à base de *PMC*, *ALADIN*, *CSI* et *CSI**), nous avons eu recours à trois méthodologies de sélections de variables d'entrée. Deux de ces méthodes ont déjà été évoquées en de multiples occasions (autocorrélation et corrélation de Pearson), nous présenterons aussi les résultats obtenus grâce à l'utilisation de la régression linéaire et au test de Student (chapitre 3 section 3-2, p 120). Il s'agit de la méthode de sélection la plus aboutie des trois présentées. La prochaine série d'expériences vise à tester cinq modèles mêlant différents modes de stationnarisation, de variables exogènes, et de sélections de variables. Les cinq modèles testés sont :

- *ARMA* + *CSI** (modèle I) : le meilleur *ARMA* avec le prétraitement *CSI** ;

- *PMC* + *CSI* (modèle II) : le meilleur *PMC* n'utilisant que des variables endogènes et le *CSI*. L'optimisation est issue de l'interprétation des *PACF* (chapitre 3, section 3-2a) ;

- *PMC* + *CSI** (modèle III) : le meilleur *PMC* n'utilisant que des variables endogènes et le *CSI**. L'optimisation est réalisée avec l'interprétation des *PACF* ;

- *PMC* + *ALADIN* + *CSI* (modèle IV) : le meilleur *PMC* utilisant les variables endogènes et le modèle *ALADIN* avec le traitement *CSI*. L'optimisation est réalisée avec la méthodologie de la pré-couche d'entrée et la régression linéaire multiple (chapitre 3, section 3-2c);

- *PMC* + *ALADIN* + *CSI** (modèle V) : identique au modèle IV mais avec le prétraitement *CSI**.

Le tableau 4-20 synthétise les résultats de ces cinq modèles testés sur les cinq sites expérimentaux. Si l'on analyse l'erreur calculée annuellement, la méthodologie V est la plus efficace d'un point de vue prédiction. Mais si l'on se focalise sur le calcul de l'erreur saisonnière, nous trouvons que dans quatre cas (Ajaccio, Bastia, Nice durant le printemps et Montpellier pendant l'été) *ARMA* (modèle I) est le meilleur prédicteur. Concernant les autres cas, *ARMA* et *PMC* sans prétraitement semblent équivalents ; ils sont cependant moins performants que le *PMC* avec stationnarisation. Le mixage entre le traitement *CSI** (méthode III) et l'utilisation des données *ALADIN* (méthode IV) améliore la qualité de la prédiction (méthode V).

	Modèles	Annuel	Hiver	Print	Ete	Automne
Ajaccio	**Persistance**	25,1	34,7	25,2	21,4	33,9
	I, *ARMA*(1,0)	19,4	29,4	**17,7**	14,3	26,8
	II, *PMC + CSI* (Endo$^{1\text{-}10}$)x15x1	20,3	27,2	20,4	13,7	24,1
	III, *PMC + CSI** (Endo$^{1\text{-}4}$)x15x1	18,6	25,3	18,4	12,2	24,2
	IV, *PMC + ALADIN +CSI* (Endo^{1,2,5,6,10}PR^{1,2}N^{1,2}T^{1,2})x15x1	19,0	26,8	19,1	12,3	23,0
	V, *PMC + ALADIN + CSI** (Endo^{1,5,9,10}PR^{1,2}N^{1}P^{1,2}T^{1,2}) x15x1	**17,8**	**24,9**	18,1	**11,7**	**22,0**
Bastia	**Persistance**	27,1	35,0	27,1	22,6	34,4
	I, *ARMA*(1,0)	21,1	26,7	**20,3**	15,8	26,9
	II, *PMC + CSI* (Endo$^{1\text{-}10}$)x10x1	22,8	27,3	23,4	16,1	25,7
	III, *PMC + CSI** (Endo$^{1\text{-}4}$)x15x1	20,8	24,9	21,4	14,9	24,9
	IV, *PMC + ALADIN +CSI* (Endo$^{1\text{-}3,8\text{-}10}$PR^{1}N^{1,2}P^{1,2}T^{1,2})	21,3	25,8	21,7	15,1	24,0
	V, *PMC + ALADIN + CSI** (Endo$^{1\text{-}3,5}$PR^{1,2}N^{1,2}P^{1,2}T^{1,2})	**19,9**	**24,4**	20,5	**14,2**	**23,3**
Montpellier	**Persistance**	26,9	32,6	25,9	24,6	33,2
	I, *ARMA*(1,0)	20,1	23,5	18,7	**15,5**	21,9
	II, *PMC + CSI* (Endo$^{1\text{-}10}$)x10x1	20,8	22,4	20,2	17,9	19,3
	III, *PMC + CSI** (Endo$^{1\text{-}3}$)x15x1	19,3	20,4	18,8	16,0	19,8
	IV, *PMC + ALADIN +CSI* (Endo$^{1\text{-}3,6\text{-}10}$Pr^{1,2}N^{1,2}P^{1,2}T^{1,2})	19,3	20,3	18,6	16,8	**18,1**
	V, *PMC + ALADIN + CSI** (Endo^{1,5,10}Pr^{1,2}N^{1,2}P^{1,2}T^{1,2})	**18,6**	**20,1**	**17,9**	**15,5**	19,2
Marseille	**Persistance**	25,3	32,9	25,3	20,0	32,3
	I, *ARMA*(2,0)	18,9	23,9	19,0	11,8	21,4
	II, *PMC + CSI* (Endo$^{1\text{-}10}$)x10x1	19,0	22,5	20,7	11,3	18,8
	III, *PMC + CSI** (Endo$^{1\text{-}4}$)x15x1	16,9	20,6	17,8	10,5	17,1
	IV, *PMC + ALADIN +CSI* (Endo$^{1,2,6\text{-}10}$PR^{1,2}N^{1,2}P^{1,2}T^{1})	17,4	20,4	18,5	10,4	**16,4**
	V, *PMC + ALADIN + CSI** (Endo^{1,3,7}PR^{1,2}N^{1,2}P^{1,2}T^{1,2})	**16,3**	**19,6**	**16,6**	**10,3**	**16,4**
Nice	**Persistance**	26,4	32,1	24,5	21,1	37,1
	I, *ARMA*(1,0)	20,7	23,5	**17,6**	12,4	37,5
	II, *PMC + CSI* (Endo$^{1\text{-}10}$)x10x1	20,9	21,7	18,8	11,7	32,3
	III, *PMC + CSI** **(Endo^{1,2,3})x15x1**	20,1	20,5	19,1	11,4	30,9
	IV, *PMC + ALADIN +CSI* (Endo$^{1,3,7:10}$PR^{1}N^{1}P^{1,2}T^{1,2})	20,1	20,4	18,1	11,4	32,3
	V, *PMC + ALADIN + CSI** (Endo$^{1:5}$PR^{1,2}N^{1,2}P^{1,2}T^{1,2})	**19,4**	**19,9**	18,6	**11,0**	**30,1**

Tableau 4-20 : comparaison (*nRMSE* en %) des cinq méthodes de prédiction pour les cinq villes étudiées. Les chiffres en gras représentent les meilleurs résultats

Comme remarqué au cours des expériences antérieures, l'été est la période la plus favorable à la prédiction : la ressource solaire est abondante et les nuages plus rares. Dans les conditions expérimentales fixées ici, les coefficients périodiques sont un outil de correction saisonnière très intéressant. En fait, la qualité de la prédiction est directement dépendante du degré de stationnarisation des séries temporelles.

c. Approche Hybride : *PMC, ARMA, ALADIN, CSI**

Si l'on se réfère au cas exogène (section 3-1 de ce chapitre), il paraît important d'utiliser un modèle hybride de prédiction, mixant le modèle linéaire *ARMA* et le modèle *PMC*. Le modèle que l'on a considéré correspond à celui donnant les meilleurs résultats dans le cas exogène, à savoir le modèle hybride *PMC* et *ARMA* (modèle C) présenté dans le paragraphe précédent et décrit par l'Equation 4-6.

$$si\ |\varepsilon^{AR}(t)| \leq |\varepsilon^{PMC}(t)|\ alors\ \hat{X}(t+1) = \hat{X}^{AR}(t+1)\ sinon\ \hat{X}(t+1) = \hat{X}^{PMC}(t+1) \quad (4\text{-}6)$$

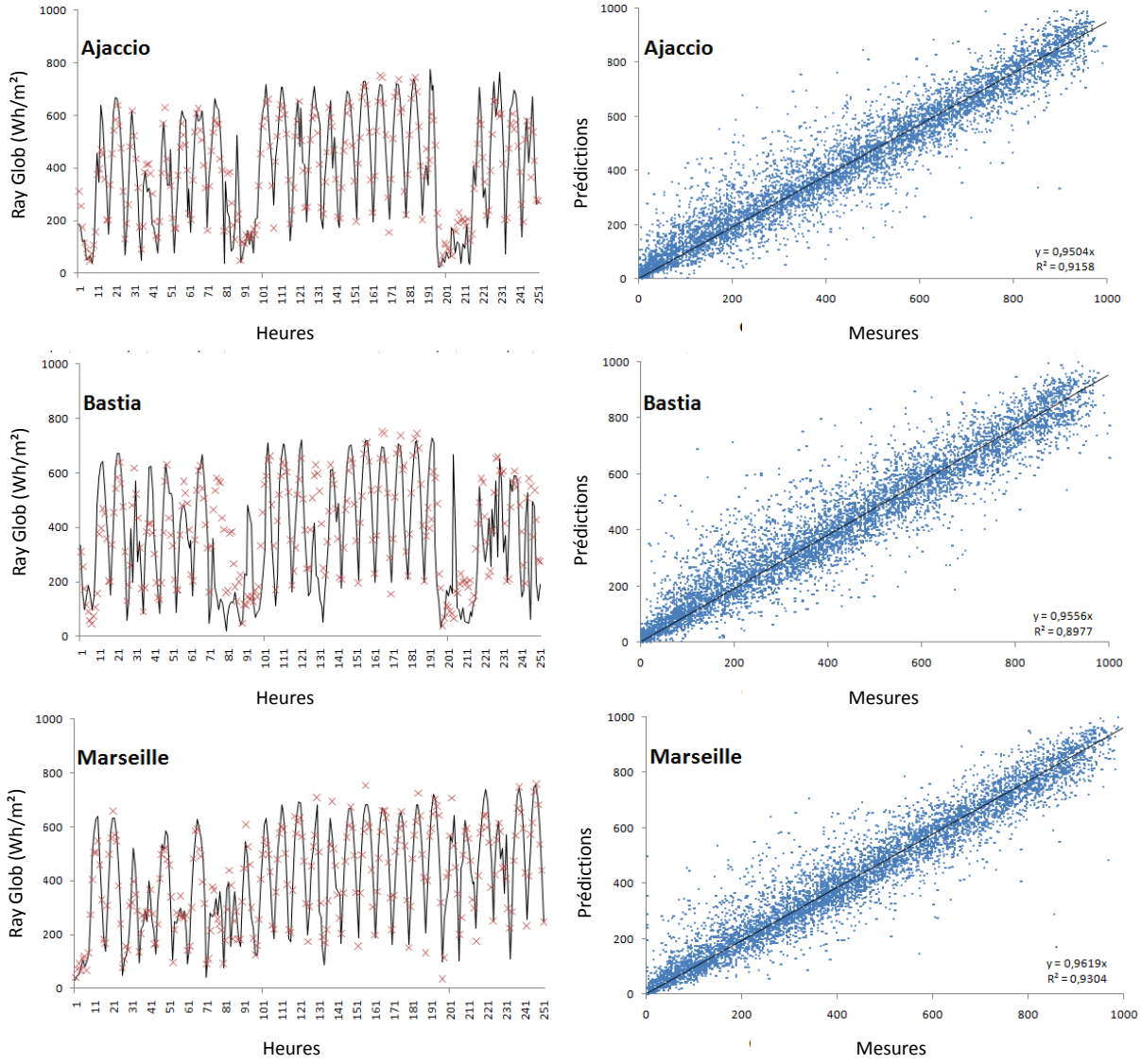
Les modèles *ARMA* et *PMC* couplés correspondent aux meilleurs prédicteurs de l'étude précédente sur l'utilisation des modèles ALADIN, à savoir une combinaison entre le modèle I et le modèle V. Le Tableau 4-21 montre les résultats de cette étude :

	Ratio ARMA/ANN	Annuel	Hiver	Printemps	Eté	Automne
Ajaccio	2592/3978	14,9(25,1)	19,4(34,7)	15,5(25,2)	11,0(21,4)	17,0(33,9)
Bastia	2557/4013	16,5(27,1)	19,5(35,0)	17,5(27,1)	13,2(22,6)	17,9(34,4)
Montpellier	2348/4222	14,7(26,9)	15,7(32,6)	15,2(25,9)	13,4(24,6)	15,5(33,2)
Marseille	2124/4446	13,4(25,3)	16,6(32,9)	14,8(25,3)	9,3(20,0)	13,8(32,3)
Nice	2301/4269	15,3(26,4)	16,6(32,1)	15,3(24,5)	10,3(21,1)	26,2(37,1)

Tableau 4-21 : erreur de prédiction (*nRMSE* en %) calculée annuellement et selon les saisons pour le modèle hybride *ARMA*, *PMC*, *ALADIN*, *CSI**. Entre parenthèse sont présentés les résultats issus de la persistance

Ce modèle hybride est très intéressant de par ses bons résultats : le seuil des 10 % est même franchi en été sur Marseille. De plus, comme dans le cas précédent, la répartition entre *PMC* et *ARMA* se fait en 2/3 ; 1/3. Toutes les saisons et villes bénéficient de cette hybridation du modèle. L'été est la saison où la méthodologie hybride apporte le moins. Certainement parce que c'est une période où les méthodes *PMC* et *ARMA* seules sont déjà très efficaces. Au

total, l'utilisation de la méthode hybride permet de diminuer l'erreur de 11 % en moyenne par rapport à la prédiction par persistance dont le *nRMSE* est de 26,2 % (moyenne sur les 5 villes). Sur la Figure 4-23, nous pouvons voir, pour chaque ville, la comparaison graphique entre la mesure et la simulation (méthode hybride décrite ci-dessus). Nous pouvons observer la bonne corrélation entre ces deux quantités. Excepté pour Bastia, les périodes nuageuses (comme la période allant de la 71^{ième} à la 81^{ième} heure) semblent bien modélisées. Pour cette localisation le modèle ne semble pas capable d'anticiper la nébulosité. Il n'est pas assez réactif et n'approxime pas très bien la réalité, impliquant des améliorations à envisager.



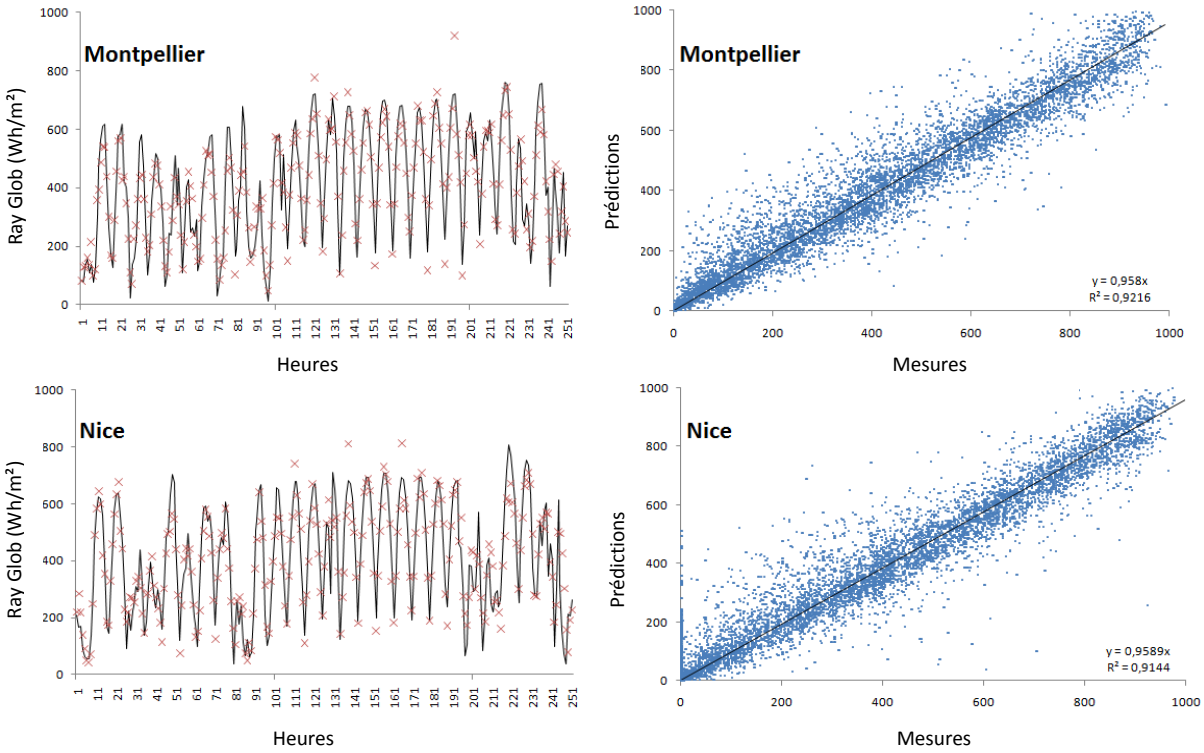


Figure 4-23 : comparaison entre le rayonnement global mesuré et simulé avec la méthode hybride *ARMA, ANN, ALADIN, CSI**. Partie gauche : les 250 heures correspondent à la période mi-Février à mi-Mars ; partie droite : toutes les prédictions sont utilisées

Pour compléter la Figure 4-23, il est souhaitable d'établir, comme dans le cas de l'étude du paragraphe précédent (section 3-2) les indices de similarité, ou les intervalles de confiance qui en découlent, pour les cinq villes étudiées. Cette approche permet de quantifier dans quelle mesure il est possible de faire confiance aux modèles utilisés. La Figure 4-24 montre, dans le cas particulier des 50 premiers heures du mois de janvier 2007, la représentation de cet intervalle de confiance dans le cadre d'une prédiction obtenue avec :

- *PMC* et *CSI** (A) doté de données exclusivement endogènes ;
- *PMC* et *CSI** (B) couplés à *ALADIN* ;
- hybridation de *PMC et ARMA* avec *CSI** couplés à *ALADIN* (C).

La période considérée est l'une des plus compliquées à prédire. Il est intéressant de voir que le simple *PMC* avec données endogènes décrit relativement bien les phénomènes mesurés, cependant, de mauvaises prédictions sont visibles (15^{ième} et 41^{ième} heure) et ne sont pas compensées par l'intervalle de confiance. Sur la courbe du milieu relative à *PMC, CSI**, *ALADIN*, l'erreur est plus régulière, les points extrêmes (de par leur non-conformité) semblent

inexistants sur la période couverte. La troisième courbe est visuellement la plus intéressante, même si l'intervalle de confiance n'encadre pas certaines valeurs mesurées (de la 29ième à la 33ième heure). En fait, avec l'approche hybride (*PMC*, *ARMA*, *CSI**, *ALADIN*) durant les journées nuageuses, c'est le *PMC* qui est le plus souvent utilisé, rendant les deux dernières courbes relativement proches pour le mois considéré, avec tout de même un léger gain.

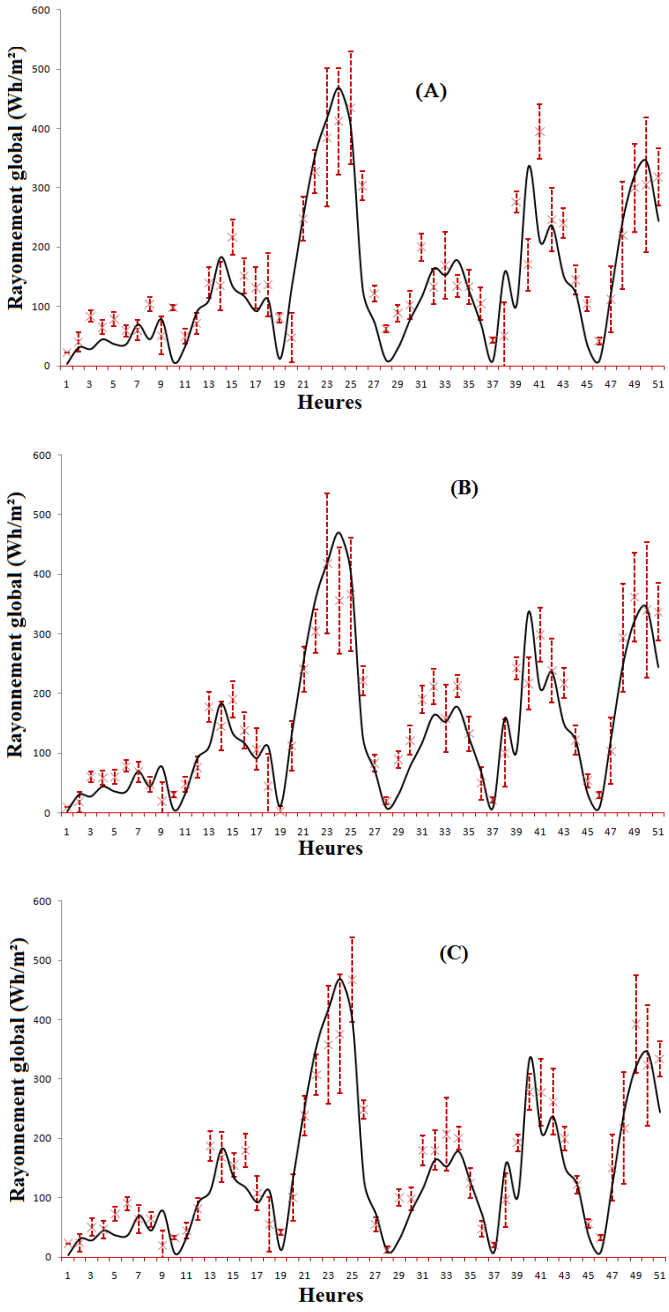


Figure 4-24 : prédictions issues des trois modèles (A= *PMC endo* + *CSI** ; B= *PMC* + *CSI** + *ALADIN* ; C= *PMC* + *ARMA* + *CSI** + *ALADIN*) accompagnées de leurs intervalles de confiance (croix et pointillés rouges). Exemple pris sur Ajaccio durant les 50 premières heures de janvier 2007, les traits pleins correspondent aux mesures

Sur la Figure 4-25, tous les résultats de cette section concernant les différentes approches et les erreurs moyennes obtenues à chaque étape sont synthétisés. On s'aperçoit que l'utilisation des variables exogènes de prévision issues du modèle *ALADIN* améliore la qualité de la prédiction. Cela confirme aussi que le facteur permettant de bien quantifier le rayonnement global, est certainement l'utilisation d'une méthodologie hybride. Il faut cependant être vigilant, car ces conclusions sont valables dans le cadre de cette expérimentation. Si les conditions initiales changent, il n'est pas garanti que l'on obtienne les mêmes conclusions.

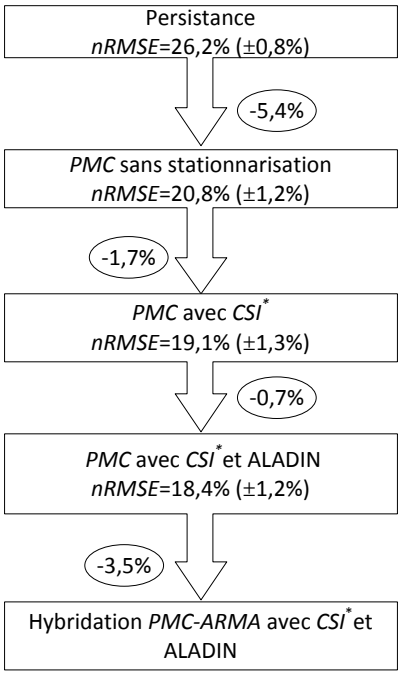


Figure 4-25 : récapitulatif des erreurs de prédiction en moyenne sur les cinq villes

En résumé, le fait de stationnariser le rayonnement global permet de diminuer l'erreur de 1,7 % en moyenne pour les cinq localisations étudiées. L'utilisation du modèle *ALADIN* a montré un vrai potentiel et un véritable intérêt. Cependant, c'est aussi l'étape qui mérite le plus d'investigation afin de mieux comprendre la complexité du modèle et de bien choisir les données à utiliser parmi les nombreuses possibles. Enfin, la dernière étape décrite, est relative à l'hybridation des résultats *PMC* et *ARMA*. Cette étape permet d'augmenter considérablement la qualité de la prédiction (gain de 3,5 %). Au final, si l'on compare notre méthodologie globale avec une prédiction simple telle la persistance, on obtient une réduction de l'erreur de prédiction de plus de 11 %.

3.4. Conclusion

Dans cette section, nous venons de présenter les modèles et les résultats obtenus dans le cas de la prédiction du rayonnement global horaire. La méthodologie de prédiction la plus sophistiquée, est certes assez lourde à mettre en place, mais donne des résultats largement supérieurs à ceux issus des autres techniques testées. On retiendra de cet horizon de prédiction, la nécessité de stationnariser par indice de ciel clair et d'utiliser des coefficients périodiques pour s'affranchir des variations saisonnières. Des outils de validation des processus de dessaisonalisation ont été mis au point. Leurs études ont révélé qu'une très bonne stationnarisation n'était pas forcément synonyme d'amélioration des résultats prédictifs. Néanmoins, le fait de ne pas stationnariser induit des résultats systématiquement moins performants. De plus, l'utilisation de variables exogènes est une valeur ajoutée à la modélisation. Les prévisions de données météorologiques *ALADIN* sont les grandeurs qui offrent le plus de garantie de précision de prédiction. Tous les résultats relatifs à ces manipulations sont en instance de publication, ils sont disponibles dans [Voyant et al. 2011c]. Cependant l'utilisation des mesures uniquement météorologiques donne aussi de bons résultats, bien que moins performants. Les règles de sélection de ces variables font aussi partie de l'étude. L'utilisation de la régression linéaire et du test de Student associé conduit à des architectures performantes capables de bien modéliser le rayonnement global. Enfin, le cumul de toutes les améliorations que l'on vient de proposer amplifient la qualité de la prédiction. En somme, l'utilisation du CSI^*, du modèle *ALADIN* sur la pré-couche du *PMC*, de la règle de transition vers la couche d'entrée basée sur la régression, et de l'hybridation des résultats issus du *PMC* avec ceux issus de *ARMA*, représente la meilleure configuration testée. Le détail de ces manipulations est disponible dans [Voyant et al. 2011b].

Dans le prochain paragraphe, nous allons mélanger les deux horizons « horaire et journalier » pour étudier le cas h+24, qui représente la prédiction par pas horaire de la journée du lendemain.

4. Le cas horaire : *h+24*

Cette nouvelle étude réalisée, concerne la prédiction à l'horizon h+24, c'est-à-dire la prédiction pour le lendemain du rayonnement solaire global horizontal heure par heure

[Chaouachi et al 2009]. L'horizon de prédiction h+24 a été peu traité dans la littérature contrairement aux horizons horaires, journaliers et mensuels. Certaines équipes de recherche se sont récemment penchées sur le cas de la prédiction h+24 à l'aide d'un réseau de neurones *PMC*. L'approche adoptée par Mellit et Pavan [2010] consiste à utiliser en entrée de l'outil de prédiction (un réseau de neurones *PMC*) les valeurs moyennes journalières du rayonnement solaire (variables endogènes moyennées) et de la température (variable exogène moyennée), ainsi que le jour du mois considéré. Pour satisfaire la prédiction de cet horizon, nous avons envisagé des approches basées sur l'utilisation des *PMC*, en nous inspirant des conclusions présentées précédemment dans ce rapport et de ce que propose la littérature. La prochaine section va permettre de détailler ces approches ainsi que les comparateurs naïfs et de référence utilisés. Seul le site d'Ajaccio a servi de support aux manipulations car au moment de l'étude, les mesures relatives aux autres localités n'étaient pas disponibles. De plus, dans un premier temps, nous nous focaliserons sur le cas endogène, puis nous introduirons les paramètres de type exogène.

4.1. Choix d'architecture

Comme dans les cas précédents, cette étude ne considère que les heures ensoleillées de la journée. Rappelons que la prédiction du rayonnement global tôt le matin, la nuit, et en fin d'après-midi n'a pas de réelle justification. De plus, les heures de faible ensoleillement ont été volontairement exclues de l'étude pour éviter d'être confronté au problème de masque solaire. Les heures de prédiction retenues, s'étalent donc de 8h00 à 16h00 en *TSV*. Cet horizon de prédiction se démarque des deux autres déjà étudiés. Si les deux premiers modèles de prédictions étaient relativement similaires, ici, les méthodologies doivent être totalement révisées et adaptées. Les deux architectures retenues et étudiées sont :

- **un réseau de *PMC*** : chaque heure de prédiction passe par l'utilisation d'un prédicteur unique et dédié. Nous travaillons donc sur neufs séries temporelles, une pour chaque heure considérée. En fait, le vecteur d'entrée du prédicteur est constitué des mesures relatives à une seule heure d'ensoleillement et de ses observations passées comme le suggère la Figure 4-26.

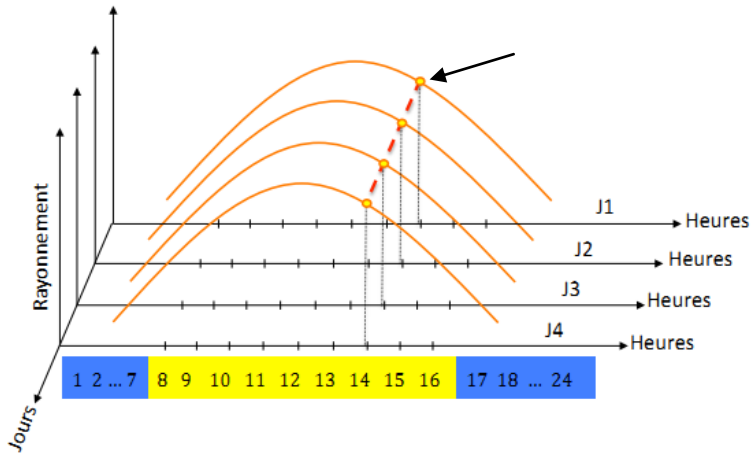


Figure 4-26 : construction des neuf séries journalières couvrant la plage 8h00-16h00. Les ronds jaunes représentent les éléments de la série temporelle à 14h00 concernant les jours j1, j2, j3 et j4

Sur cette figure, on voit que pour réaliser la prédiction du rayonnement solaire à 14h00 au pas j+1, le vecteur d'entrée est défini par l'observation à 14h00 aux pas j, j-1…j-*Ne*. Le processus global est constitué de neuf modèles de prédiction, à partir desquels on peut reconstituer la journée d'ensoleillement prédite (neuf réseaux de *Ne* vecteurs d'entrée). Les séries temporelles du rayonnement solaire sont alors constituées de la succession des mesures journalières pour une seule tranche horaire. Si on prend le cas de la figure précédente, la série temporelle extraite est celle du rayonnement solaire global horizontal à 14h00 chaque jour pendant 10 ans. Les séries sur lesquelles on travaille sont donc des séries temporelles journalières, et le type de prédiction qu'on leur applique est de type j+1. Les profils des journées d'ensoleillement prédites sont reconstitués à partir des neuf prédictions comme le montre la Figure 4-27A.

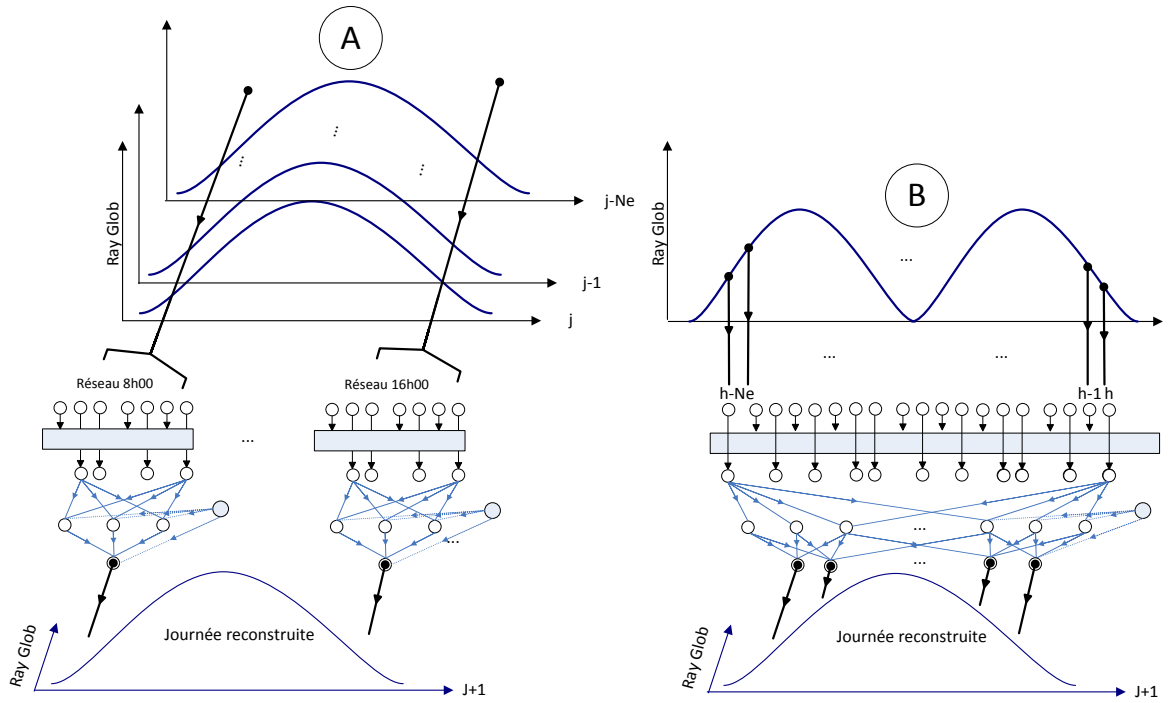


Figure 4-27 : reconstruction de la journée du lendemain avec neuf prédicteurs indépendants (A) et avec un *PMC* à neuf sorties (B)

- **un *PMC* à multiples sorties**. Cette approche correspond à la méthodologie conventionnelle proposée dans les sections précédentes. Toutes les mesures historiques « récentes » sont positionnées en entrée d'un *PMC* unique. Les mesures sont positionnées chronologiquement dans le vecteur d'entrée (Figure 4-27B).

Pour tester l'intérêt de ces deux architectures, nous avons choisi de confronter les résultats avec ceux obtenus par méthodes de « persistance » (prédictions = mesures de la veille corrigées) et *ARMA*. Cette dernière est basée sur la méthode de neuf *ARMA* dotés chacun d'une sortie comme détaillée précédemment. On notera que toutes ces méthodologies sont compatibles avec l'utilisation de l'indice de clarté, de l'indice ciel clair et des coefficients périodiques.

4.2. Optimisation

La prédiction 24 heures à l'avance est plus complexe que les deux horizons que nous avons traités précédemment (j+1 et h+1). Les architectures étant plus sophistiquées, nous

avons souhaité dans un premier temps, nous focaliser sur leur comparaison, estimation et validation. Ainsi nous n'avons pas utilisé une approche hybride mixant *ARMA* et *PMC* ni des paramètres exogènes de type *ALADIN*. L'optimisation repose ici sur le choix du nombre de neurones d'entrée et de neurones cachés. Les données ayant permis de réaliser l'optimisation sont issues de la station d'Ajaccio et concernent les années 1999-2006, les années 2007 et 2008 ont servi de support de comparaison des différents prédicteurs. Les paramètres fixés sont : une seule couche cachée avec la fonction d'activation tangente hyperbolique et l'algorithme *LM*. La méthodologie de sélection des variables endogènes de la couche d'entrée et du nombre de neurones présents sur la couche cachée sont basés sur une approche exhaustive, où tous les cas de figures ont été testés, pour 1-20 neurones cachés et 1-20 neurones d'entrée. De plus, les trois modes : série brute (pas de stationnarisation ; *X*), indice de clarté (*k*) et indice ciel clair (*CSI*) ont pu être testés. Nous n'avons pas voulu utiliser une approche de sélection de variables évoluée comme présentée dans la section 3 de ce chapitre : en effet, ici la problématique est plus complexe, et le nombre de paramètres à optimiser est nettement supérieur. Notons aussi que durant cette étude, les coefficients périodiques n'ont pas été utilisés, là aussi par soucis de simplicité. Le Tableau 4-22 montre le résultat d'optimisation obtenu pour le réseau de neuf *PMC*.

Réseau	*X*	*k*	*CSI*
8h00	$Endo^{1\text{-}3}$x4x1	$Endo^{1\text{-}3}$x4x1	$Endo^{1\text{-}11}$x8x1
9h00	$Endo^{1\text{-}10}$x4x1	$Endo^{1\text{-}8}$x3x1	$Endo^{1\text{-}8}$x3x1
10h00	$Endo^{1\text{-}11}$x4x1	$Endo^{1\text{-}20}$x2x1	$Endo^{1\text{-}20}$x2x1
11h00	$Endo^{1\text{-}3}$x8x1	$Endo^{1\text{-}2}$x4x1	$Endo^{1\text{-}3}$x3x1
12h00	$Endo^{1\text{-}20}$x1x1	$Endo^{1\text{-}20}$x1x1	$Endo^{1\text{-}20}$x1x1
13h00	$Endo^{1\text{-}12}$x7x1	$Endo^{1\text{-}19}$x1x1	$Endo^{1\text{-}1}$x14x1
14h00	$Endo^{1\text{-}18}$x2x1	$Endo^{1\text{-}20}$x1x1	$Endo^{1\text{-}3}$x19x1
15h00	$Endo^{1\text{-}11}$x5x1	$Endo^{1\text{-}12}$x1x1	$Endo^{1\text{-}4}$x10x1
16h00	$Endo^{1\text{-}13}$x3x1	$Endo^{1\text{-}1}$x1x1	$Endo^{1\text{-}6}$x5x1

Tableau 4-22 : optimisation des neuf *PMC* indépendants

Concernant l'approche basée sur le *PMC* unique pourvu de neuf sorties, c'est le Tableau 4-23 qui décrit la configuration optimale. Les 27 neurones d'entrée observables sur les trois réseaux correspondent à trois journées de neufs heures (=3x9).

	X	*k*	*CSI*
PMC	$Endo^{1\text{-}27}$x*1*x*9*	$Endo^{1\text{-}27}$x*1*x*9*	$Endo^{1\text{-}27}$x*2*x*9*

Tableau 4-23 : optimisation du *PMC* doté de neuf sorties

Quand on regarde les deux derniers tableaux, aucune règle de sélection ne semble se dégager. Le nombre de neurones endogènes sur la couche d'entrée et de neurones cachés fluctue énormément. On remarquera surtout que l'utilisation de l'indice ciel clair semble accroître le nombre de neurones cachés.

Concernant les *ARMA*, nous nous sommes focalisés uniquement sur la partie autorégressive *AR*, car le fait d'intégrer des termes de régression sur le résidu n'améliore pas les résultats calculés. Pour l'optimisation, nous avons tout d'abord utilisé une méthode basée sur l'analyse des autocorrélogrammes, déjà présentée précédemment dans le chapitre 3 section 3-2, p 116. A partir des lags choisis, nous avons ensuite estimé les modèles *AR* correspondants pour chaque heure considérée. Un test de Student sur les *p* coefficients de régression des *AR*(p) notés φ_i (pour $i \in [1, p]$) révèle que seul le lag 1 de chaque réseau horaire est significativement différent de zéro (Figure 4-28) si l'on se place dans le cas de l'indice ciel clair ou de l'indice de clarté. Dans l'optique de simplifier la suite des manipulations, nous avons opté pour ne garder que le lag un des neuf réseaux *AR*, soit neuf réseaux *AR*(1).

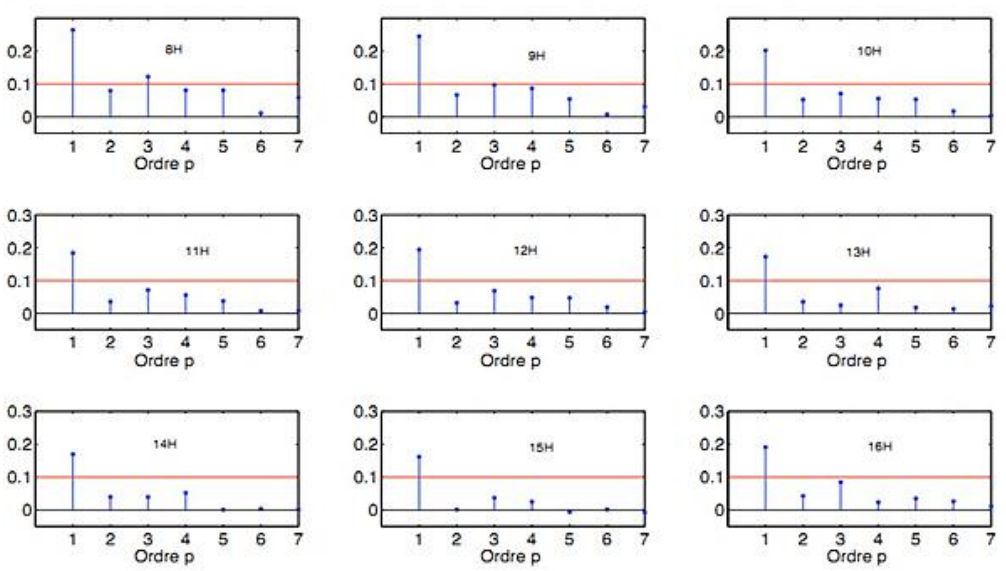


Figure 4-28 : valeurs des coefficients d'autorégression des neuf processus *AR(p)* avec la stationnarisation *CSI*. La ligne rouge indique le seuil de significativité

4.3. Résultats de prédiction

L'optimisation des différents modèles proposés dans la section précédente a permis de sélectionner parmi toutes les architectures disponibles les plus pertinentes pour le problème posé ici. Le test de prédiction est réalisé sur le même site que l'apprentissage (Ajaccio) et sur deux années complètes (2007-2008). Le tableau 4-24 dresse les erreurs relatives aux modèles étudiés avec l'approche des neuf réseaux indépendants. Les *nRMSE* sont calculés sur la base de la moyenne journalière du rayonnement global.

Modèle		Annuel	Hiver	Printemps	Eté	Automne
Persistance		*35,1*	*54,8*	35,2	28,0	40,4
ARMA	*k*	*29,1*	*44,6*	29,2	24,0	33,2
	CSI	*28,6*	*44,2*	28,6	23,1	**32,8**
PMC	*k*	*28,5*	*44,6*	28,8	22,9	**32,8**
	CSI	***28,2***	***44,1***	**28,6**	**22,4**	33,2

Tableau 4-24 : *nRMSE* (%) des prédictions réalisées avec la méthode des 9 réseaux et *AR* indépendants. En gras les meilleurs résultats

On constate que les approches sophistiquées comme *ARMA* ou *PMC* surclassent largement le modèle naïf, et ce, surtout en hiver. De telles performances dans le cas du modèle de persistance s'expliquent par le choix de l'horizon de prédiction. Ce modèle est dédié, de par sa définition, à une prédiction sur un horizon de temps court, de l'ordre de la minute ou du moins infra-horaire. En règle générale on observe que les *PMC* sont plus performants que les modèles *ARMA*. Mais le gain reste tout de même minime, sauf en été où le *PMC* améliore la prédiction de plus de 0,7 % vis-à-vis de son concurrent linéaire. On notera aussi, que les meilleures prédictions sont obtenues avec l'utilisation de l'indice ciel clair. Pour compléter cette étude, le Tableau 4-25 présente les résultats obtenus avec la méthodologie du *PMC* unique avec neuf sorties.

Type		Annuel	Hiver	Printemps	Eté	Automne
Persistance		*35,1*	*54,8*	35,2	28,0	40,4
PMC	*k*	*27,9*	*44,2*	**27,9**	22,2	32,7
	CSI	***27,8***	***42,8***	28,4	**22,0**	**31,3**

Tableau 4-25 : *nRMSE* (%) des prédictions réalisées avec la méthode du réseau unique avec neuf sorties. En gras, les meilleurs résultats

Il y a une véritable différence liée au mode de stationnarisation. En effet, il n'y a qu'une saison (printemps) où *CSI* n'est pas le mode donnant le meilleur résultat. De plus, cette méthodologie de prédiction visant à utiliser un seul *PMC* avec neuf sorties améliore considérablement la qualité de la prédiction. Si l'on se réfère au cas *CSI*, le gain généré est de 0,4 % sur les deux années, et si l'on se restreint à ne regarder que les mois hivernaux, on a alors un gain de plus de 1,5 %. Il semble donc que le *PMC* tire profit de l'utilisation de toutes les mesures du jour précédent et pas seulement de celles liées à l'heure considérée. Ce mécanisme qui permet d'utiliser plus d'une mesure du jour précédent évite certainement toutes incertitudes liées au bruit de la série temporelle qui pourrait polluer la prédiction. Ainsi, par exemple si la veille à 8h00 il y avait une forte nébulosité qui a disparu par la suite :

- dans le cas du réseau à neuf sorties, cette occurrence sera effacée par les autres mesures. Pour la prédiction à 8h00 le lendemain cette information sera minime. Le bruit est lissé sur les neuf heures prélevées ;

- dans le cas des neufs réseaux, ce sera essentiellement cette occurrence nuageuse ponctuelle qui conditionnera la prédiction du jour suivant à 8h00.

La Figure 4-29a illustre la comparaison entre les deux meilleurs prédicteurs de chaque famille (neufs *ARMA* à une sortie et un *PMC* à neuf sorties avec l'utilisation du *CSI*). Dans ce graphique, on ne présente que l'allure des prédictions à 12h00 réalisées chaque jour.

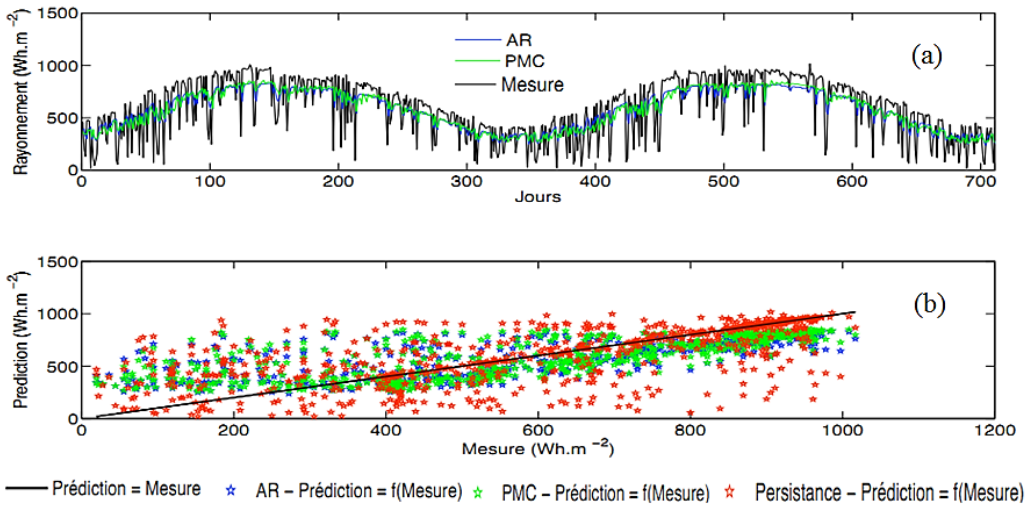


Figure 4-29 : comparaison des prédictions faites avec le meilleur *PMC* et le meilleur *ARMA*. En (a) le profil des séries prédites et en (b) la comparaison avec le graphique de type y = x

On peut ainsi voir graphiquement deux années de simulation sans être gêné par un nombre important de mesures ou une périodicité journalière masquant certaines parties du graphique. La persistance n'y est pas représentée contrairement à la Figure 4-29b où on peut voir cette fois la comparaison des prédictions faites avec *ARMA*, *PMC* et la persistance. Le fait d'utiliser le *nRMSE* tend à faire disparaître les gros écarts (voir chapitre 2, section 4-1, p 87), au profit des plus petits mais qui sont plus nombreux. C'est pour cette raison que tant pour *AR* que pour *PMC*, on a l'impression que la simulation décrit une moyenne harmonieuse, ne prenant « pas le risque » de s'écarter de sa valeur centrale. Ainsi pour les faibles valeurs de rayonnement global, les modèles surestiment le rayonnement, et pour les fortes, ils le sous-estiment. En hiver, le modèle de persistance donne une prédiction dont les éléments sont très dispersés, elle devient inexploitable alors que *AR* et *PMC* gardent globalement le même profil. En été, les trois prédicteurs sont similaires, mais ce résultat est local car il y a un fort ensoleillement sur Ajaccio et un taux d'aléas climatiques très faibles. Pour essayer d'améliorer les résultats obtenus sur les autres saisons, nous avons vu dans le cas journalier et horaire, qu'il est possible d'utiliser des variables exogènes en entrée des *PMC*. Nous allons voir dans la suite de cette étude l'impact que vont avoir ces mesures sur la qualité de la prédiction.

4.4. Utilisation de variables exogènes

Jusqu'alors, les prédicteurs utilisent pour unique source de données la série temporelle d'observations qu'ils cherchent à prédire. Rappelons que les modèles *ARMA* et persistance ne peuvent être construits qu'à partir de ces variables endogènes. Le cas des réseaux de neurones est diffèrent. Les variables exogènes peuvent être intégrées au vecteur d'entrée comme nous l'avons vu dans les horizons précédemment étudiés. L'horizon *h+24* ne déroge pas à la règle. La construction de ces nouveaux modèles utilise les réseaux optimisés des cas endogènes étudiés antérieurement. La sélection des variables exogènes se fait avec la méthode des corrélations de Pearson déjà présentées dans la section 2-3 de ce chapitre. Les variables exogènes testées sont :

- la pression, le gradient de pression, la nébulosité et les précipitations pour les neuf heures d'ensoleillement ;

- les valeurs moyennes journalières de la pression, du gradient de pression, de la nébulosité et des précipitations ;

- la différence de pression entre la première et la dernière heure d'ensoleillement.

Parmi toutes ces variables exogènes, l'étude des corrélations montre que les précipitations, le gradient de pression et la différence de pression sur la journée ne sont pas suffisamment corrélés au rayonnement global pour être intégrées au *PMC*. Au final, le vecteur des entrées exogènes est constitué de la pression des heures 8 à 16 au lag 1 (neuf composantes), la nébulosité des heures 8 à 16 au lag 2, soit (dix-huit composantes) et la nébulosité moyenne journalière au lag 2 (deux composantes). La même étude que celle menée précédemment dans le cas endogène est décrite sur le Tableau 4-26 (seule la stationnarisation *CSI* est utilisée).

Type		Annuel	Hiver	Printemps	Eté	Automne
Persistance		*35,1*	*54,8*	35,2	28,0	40,4
ARMA		*28,6*	*44,2*	28,6	23,1	32,8
PMC endo	*A*	*28,2*	*44,1*	28,6	22,4	33,2
	B	*27,8*	*42,8*	**27,4**	22,0	**31,3**
PMC exo	*A*	*28,0*	***42,4***	28,8	22,4	31,9
	B	***27,3***	***42,4***	27,8	**21,7**	**31,3**

Tableau 4-26 : apport des variables exogènes sur la qualité de prédiction pour la méthode des neuf réseaux indépendants (*A*) et celle du réseau unique avec neuf sorties (*B*). En gras les meilleurs résultats

L'apport des variables exogènes sur l'erreur annuelle, ne semble réellement intéressant que pour la méthode B. Cependant, le gain maximum concerne la méthode A. En hiver, on gagne 1,7 %, en automne 1,3 %, mais par contre on perd sur les deux saisons printemps et été. Sur la Figure 4-30, est représenté le détail de l'erreur horaire concernant le meilleur *PMC* (réseau à neuf sorties, avec indice ciel clair et variables exogènes), la meilleure méthodologie *ARMA* (neuf *AR* avec chacun une sortie et l'indice de ciel clair) et la persistance.

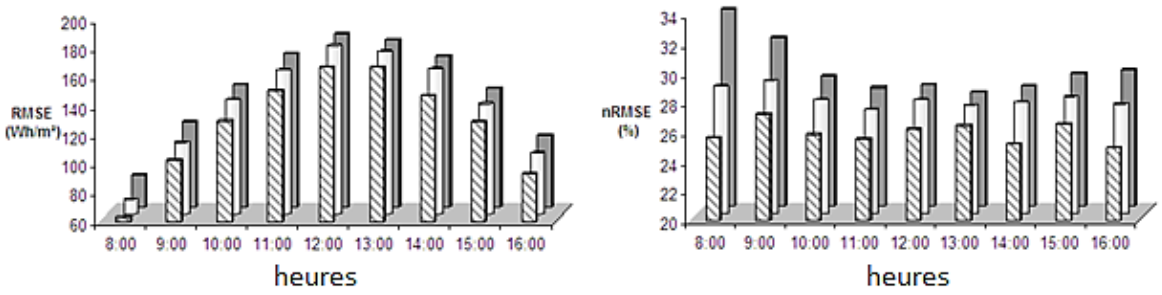


Figure 4-30 : distribution de l'erreur horaire (*RMSE* à gauche et *nRMSE* à droite) de prédiction concernant le *PMC* (premier plan hachuré), *ARMA* (second plan en blanc) et la persistance (arrière-plan en gris)

On constate une allure totalement différente concernant les erreurs *RMSE* et *nRMSE*. Ce dernier est obtenu avec une normalisation par la moyenne quadratique de la série horaire concernée et non de la série totale. Le maximum de l'erreur absolue se situe à 12h00 pour les trois prédicteurs. Par contre, pour l'erreur relative, le maximum est uniformément réparti pour *PMC* et *AR* et sur les heures de faible luminosité pour la persistance. Aucun des trois prédicteurs ne dépasse les 200 Wh/m² d'erreur, mais seul le *PMC* possède tous ses résultats inférieurs à 28 % de *nRMSE*.

4.5. Conclusion

Le fait d'aller chercher des horizons de prédiction profonds (h+24), modifie quelque peu l'approche à envisager. Ainsi, nous avons compris que ce type de prédiction, est particulièrement difficile à appréhender. Rechercher la finesse sur une prédiction 24 heures à l'avance dépend de trop de paramètres pour espérer obtenir le mêmes niveau de résultat que pour les horizons h+1 ou j+1. A l'issue de manipulations, nous pouvons conclure, qu'il est souhaitable de stationnariser les données. Pour ce faire il convient d'utiliser l'indice ciel clair, même si l'indice de clarté donne des résultats quasi-similaires. L'approche des neufs réseaux indépendants, bien que séduisante sur le papier n'est pas la plus performante. Il convient de rester sur une approche classique ne faisant intervenir qu'un seul *PMC* possédant neuf sorties. En travaillant sur une variété de réseau plus propice à l'utilisation de délais (*TDNN* ; time delay neural network), on pourrait certainement encore améliorer les résultats en intégrant plus facilement les différentes journées à considérer. L'utilisation des variables exogènes est

quant à elle délicate. Le ratio entre performance et complexité nous fait préconiser, en l'état actuel de nos connaissances, de ne pas utiliser de variables exogènes.

Au final, on peut estimer que l'approche simple où l'on implémente au fil de l'eau les données de rayonnement global en entrée du *PMC*, et en utilisant une approche simple de type indice de clarté pourrait être utilisée. Elle permettrait à un gestionnaire de réseau de connaître le soir à 17h00 le profil horaire de la journée du lendemain, avec une imprécision certes, mais qui ne serait guère améliorée par une approche beaucoup plus sophistiquée.

5. Le cas minute : *m+5*

Cet horizon de prédiction est, de par sa nature, totalement différent de ce que l'on a étudié jusqu'ici. L'originalité de ce cas de figure, vient de la fréquence d'échantillonnage des mesures qui est inférieure à la dynamique d'occurrence nuageuse. Ainsi, en 5 min le ciel a une grande probabilité de conserver sa couverture nuageuse. Les données à disposition sont relatives au mur *PV* du laboratoire de Vignola. Elles couvrent la période allant du 1/03/2009 au 31/08/2010 avec un pas d'échantillonnage de « cinq minutes ». Les séries temporelles sont composées de 158 112 échantillons. La tri-façade permet de dégager six zones distinctes, combinant une technologie de Silicium amorphe et une de Silicium monocristallin avec trois azimuts différents : 0°, 45°SE et 45°SO et une inclinaison de 80° par rapport à la surface du sol (voir chapitre 1, section 5-3). Sur chacune des façades, nous avons à disposition des données de rayonnement global dans le plan des champs *PV*, de courant *DC* et de courant *AC*. Nous allons voir dans la suite de cette section les résultats obtenus pour la prédiction du rayonnement global, puis pour les puissances *PV*. Nous traiterons particulièrement de l'intérêt de stationnariser les données d'entrée pour cet horizon très court.

5.1. Prédiction du rayonnement global

Avant de présenter l'étude sur les résultats concernant la production d'énergie électrique *PV,* il convient de commencer par analyser la prédiction du rayonnement global sur une surface inclinée, en fonction des trois orientations azimutales disponibles : 0°, 45°*SE* et 45°*SO*. Pour cette étude, nous avons utilisé les treize premiers mois pour l'apprentissage et les quatre derniers (de mai à août) pour le test de prédiction. La première étape vise à étudier la

stationnarité de la série en utilisant l'indice ciel clair (noté *CSI*). Nous n'utiliserons pas l'indice de clarté. En effet, l'inclinaison et l'orientation azimutale du rayonnement hors atmosphère sont délicates car celui-ci ne possède pas (ou peu) de composantes diffuses. De plus, à 80° et/ou pour des orientations *SE* ou *SO*, il est très difficile de définir clairement un modèle de ciel clair qui arrive à approximer le signal (limite du couplage Solis - Klutcher). Comme les panneaux et détecteurs sont fixés aux murs du bâtiment, les masques solaires ne sont pas négligeables et restent très difficiles à modéliser. De plus, du fait de la hauteur entre les détecteurs et le sol et de la localisation du bâtiment (surplombant la mer), les diffusés du sol et de l'atmosphère sont de natures différentes entre le début et la fin de la journée. En début de journée, le rayonnement global subit les effets des zones montagneuses et en fin de journée ce sera essentiellement la zone maritime et la réflexion sur l'eau qui conditionnera le rayonnement diffus. La modélisation de type ciel clair ne concerne que des régions homogènes (albédo du sol et composition de l'atmosphère constant tout au long de la journée). Pour comprendre le problème nous pouvons voir sur la Figure 4-31 une comparaison entre modélisation *SOLIS* incliné et la mesure du rayonnement global. De visu, le modèle ciel clair semble cohérent avec les mesures. On remarquera un problème lié à l'enregistrement des mesures au début du 2^{ième} jour et entre le 3^{ième} et le 4^{ième} jour indiqué par des flèches sur la Figure 4-31.

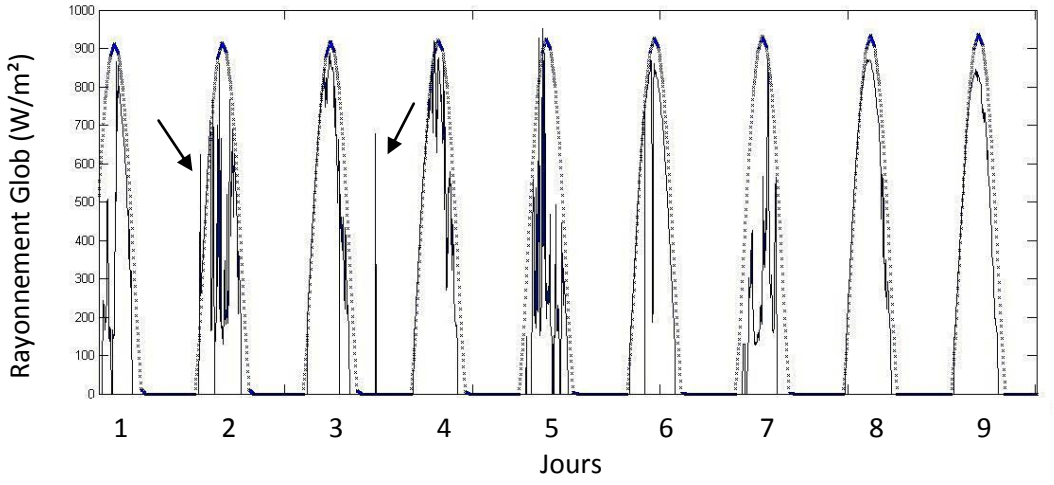


Figure 4-31 : modélisation ciel clair obtenue par le modèle *SOLIS* (pointillé) confrontée à la mesure du rayonnement global (trait plein). Mesures issues du champ *PV* de Vignola en août 2009, avec une inclinaison de 80° et une orientation azimutale Sud

La Figure 4-32 présente les résultats obtenus pour *CSI* concernant la même période que pour la figure précédente. On remarque que l'indice est égal à 1 en début de journée, mais chute rapidement en fin de journée. Même sans passage nuageux il ne reste pas constant. Le problème d'hétérogénéité développé dans la page précédente et l'inclinaison extrême des détecteurs (80°) expliquent certainement ce phénomène. Sur ces échantillons de fréquence d'échantillonnage élevée, la modélisation, comme utilisée jusque-là, montre des limites de fiabilité. Les deux problèmes de mesure précédemment évoqués sont en parties responsables du pic de *CSI* en début de seconde journée. Concernant le second pic présent sur la Figure 4-31, il est annulé ici car durant la nuit le *CSI* est considéré constant et nul.

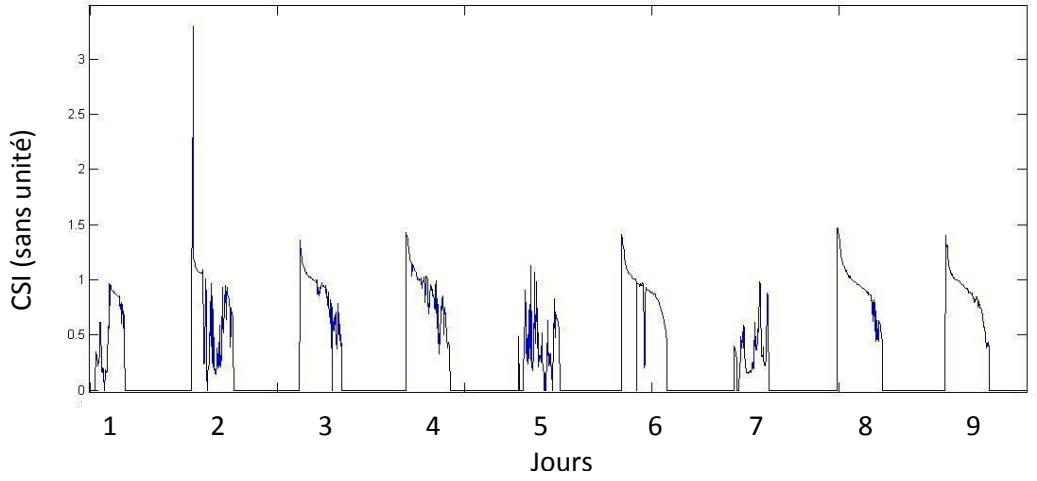


Figure 4-32 : indice de ciel clair estimé sur le champ *PV* de Vignola. Mesures effectuées en août 2009, avec une inclinaison de 80° et une orientation azimutale Sud

Bien que ces figures soient relatives à la face Sud de la façade tri-pans, les résultats sont équivalents sur les autres façades. Les modélisations de type « ciel clair » comme utilisées lors de ces manipulations ne semblent pas exploitables. Après avoir présenté les limites des modèles analytiques déjà utilisés pour d'autres horizons, il faut maintenant quantifier leurs impacts et ainsi vérifier l'intérêt de la stationnarisation sur la prédiction. Pour effectuer les tests de prédiction, il faut construire des modèles qui nécessitent une sélection de variables endogènes. Nous avons opté pour une étude des autocorrélogrammes car très simple et rapide à mettre en place. La Figure 4-33 suggère de se concentrer uniquement sur les quatre premiers lags endogènes, les autres n'étant pas réellement significatifs (le seuil de significativité est très faible, il vaut ~0,1). Ce résultat nous permet d'anticiper sur le fait que la persistance sera un bon prédicteur dans le cadre de cet horizon. En effet, la corrélation entre le rayonnement à l'instant t et celui à l'instant t-5 minutes (lag 1) est largement supérieur à ce que l'on peut

avoir avec les mesures plus anciennes. En utilisant les équations de Yule-Walker pour estimer un modèle *AR(1)* (chapitre 2, section 3-7), on obtient $X_{t+1} = X_t$ (autocorrélation au premier ordre égal à 1 et les autres nulles). Dans le cas de l'horizon m+5, la persistance correspond à un *AR* particulier. C'est un prédicteur de choix pour l'horizon m+5. La persistance (~*AR*(1)) devient notre prédicteur de référence, et de ce fait, nous n'avons pas pour cette étude de prédicteur naïf. En utilisant un *PMC* avec les lags supplémentaires 2, 3 et 4 en entrée, on augmente cependant théoriquement la puissance du prédicteur vis-à-vis d'une simple persistance. Le *PMC* permet en fait, de réaliser une persistance améliorée, avec prise en compte de la tendance des dernières minutes.

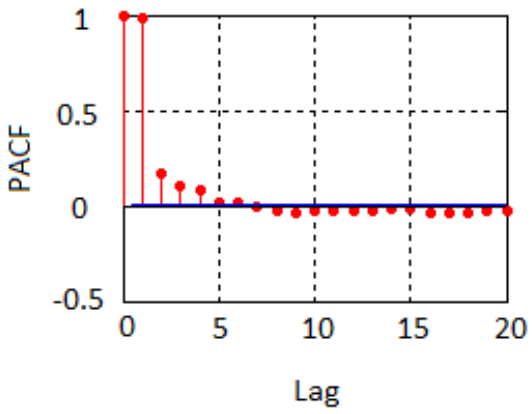


Figure 4-33 : *PACF* du rayonnement global sans stationnarisation. Façade sud, mur de Vignola, Ajaccio du 1/03/2009 au 31/08/2010

Concernant les autres orientations et les autres modes de stationnarisation (*k* et *CSI*), les résultats liés aux autocorrélations sont équivalents. L'optimisation du *PMC* avec une seule couche cachée, la fonction de transfert tangente hyperbolique et l'algorithme d'apprentissage *LM*, conduit à utiliser pour toutes les simulations, soit un $Endo^{1\text{-}4}$x*4*x*1*, soit un $Endo^{1\text{-}6}$x*6*x*1*. Le tableau 4-27 décrit les résultats des prédictions concernant l'orientation Sud-Ouest, Sud et Sud-Est. On y trouve les *nRMSE*, *MSE* et *MAE* relatifs aux différents types de stationnarisation. La persistance donne systématiquement les meilleurs résultats pour le *MAE* et la stationnarisation ne permet jamais d'améliorer les erreurs de prédiction. Si l'on se réfère aux définitions et interprétations des estimateurs de l'erreur *MSE* et *MAE* (chapitre 2, section 4-1), on peut dire que par rapport à la persistance, le *PMC* sans stationnarisation minimise d'avantage les gros écarts. On notera que l'utilisation de l'indice de clarté (*k*) permet une seule fois (*nRMSE* de juin) d'obtenir le meilleur résultat alors que le *CSI* ne le permet jamais.

Orientation / Type		MAE W/m²	MSE W/m²	nRMSE Total (%)	nRMSE mai (%)	nRMSE juin (%)	nRMSE juillet (%)	nRMSE août (%)
SO	*PMC* - Brute	13,3	**1423**	**21,4**	**31,4**	20,7	**14,2**	**19,5**
	PMC + k	14,4	1575	22,5	32,3	**20,1**	15,4	19,6
	PMC + CSI	17,2	1530	22,2	31,9	21,1	16,3	20,0
	Persistance	**12,3**	1471	21,8	32,3	20,9	14,4	19,6
S	*PMC* - Brute	11,2	**959**	**20,2**	**28,0**	**22,6**	**13,5**	**16,5**
	PMC + k	13,6	1107	21,7	29,6	23,7	14,8	18,4
	PMC + CSI	16,0	1235	21,9	29,7	25,5	17,4	19,5
	Persistance	**10,3**	1019	20,8	28,8	23,2	13,8	17,1
SE	*PMC* - Brute	13,2	**1413**	**23,2**	**31,8**	**26,5**	**14,6**	**20,6**
	PMC + k	12,8	1520	24,2	32,6	27,6	15,1	21,7
	PMC + CSI	16,2	1699	25,6	33,3	28,1	17,8	23,8
	Persistance	**11,8**	1559	24,5	33,3	27,9	14,8	22,0

Tableau 4-27 : résultats des prédictions de rayonnement global pour l'orientation *SO*, *S* et *SE* du mur de Vignola.

Cette étude ne permet pas de conclure que l'utilisation de la modélisation ciel clair ou du rayonnement hors atmosphère soient tout aussi efficaces que dans le cas horaire ou journalier. Si l'on se réfère à la Figure 4-34 et au profil journalier de la prédiction d'une journée typique du mois de juillet, on constate que l'erreur globale de prédiction est liée en partie aux points aberrants qui sur ou sous-estiment complètement la valeur réelle du rayonnement.

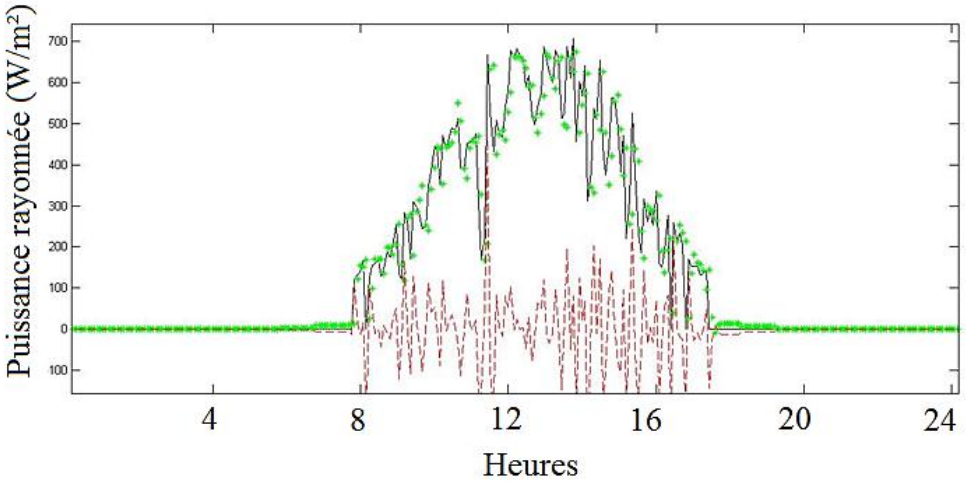


Figure 4-34 : prédiction faite avec le *PMC* du rayonnement global pour l'orientation Sud en juillet 2010 et avec la stationnarisation *CSI* (points verts) et erreur de prédiction (pointillés rouges)

Par la suite, lors des estimations des puissances *PV*, nous n'utiliserons que les séries brutes sans utilisation de *k* ou du *CSI* lors des modélisations avec *PMC*.

5.2. Prédiction de la puissance PV

Un des avantages liés à l'utilisation des *PMC* réside dans la souplesse de sélection des entrées. En effet, il est possible de positionner sur la couche d'entrée uniquement des données endogènes, uniquement des données exogènes, ou un mélange des deux. Nous allons voir dans la suite, le type de variables qui est le plus intéressant pour la prédiction de la puissance *PV* dans le cas m+5.

a. Sélection des entrées du *PMC*

Pour estimer les puissances *PV* issues des trois façades du laboratoire, nous avons le choix quant à la sélection de la nature des entrées. Le tableau 4-28 montre (uniquement dans le cas *a-Si*), les entrées du réseau qui permettent de bien modéliser les différentes puissances électriques *PV* pré ou post onduleur disponibles. Les réseaux de neurones utilisés sont de type *PMC* uniquement exogène ou uniquement endogène, nous n'avons pas souhaité mélanger les deux pour ne pas alourdir la méthodologie. Une entrée Glob *SO* et une sortie P_{ac} *SO* correspondent à un *PMC* construit avec du rayonnement global orienté *SO* en entrée et une prédiction de puissance *PV* post-onduleur.

		Sorties du réseau					
		P_{ac} *SO*	P_{ac} *S*	P_{ac} *SE*	P_{dc} *SO*	P_{dc} *S*	P_{dc} *SE*
Entrées du réseau	*Glob SO*	33,6	68,0	89,8	32,0	74,4	80,6
	Glob S	63,9	26,3	58,3	55,5	26,4	58,6
	Glob SE	73,4	63,8	45,9	70,9	62,3	46,0
	P_{ac} *SO*	**22,6**	61,9	77,3	**20,1**	59,8	77,5
	P_{ac} *S*	60,4	**20,5**	60,8	60,0	**18,7**	56,9
	P_{ac} *SE*	69,9	50,1	**23,4**	70,1	55,5	**21,3**
	P_{dc} *SO*	**21,9**	81,2	81,0	**20,1**	78,8	90,0
	P_{dc} *S*	82,8	**20,3**	69,7	72,5	**18,4**	60,0
	P_{dc} *SE*	89,2	65,5	**23,1**	84,7	57,2	**21,1**
	Persistance	*24,3*	*22,5*	*25,2*	*20,4*	*19,0*	*21,8*

Tableau 4-28 : interactions croisées et erreur de prédiction (*nRMSE* en %) entre les différentes entrées et sorties des *PMC* et comparaison avec la persistance

L'optimisation des différents réseaux conduit à utiliser presque tout le temps quatre neurones d'entrée (soit quatre lags) et quatre neurones cachés. La différence entre les panneaux *a-Si* et *mono-Si* est uniquement liée à une différence de rendement. Aussi pour l'étude menée ici, les interprétations sont identiques. On remarque qu'il est possible d'utiliser des variables exogènes, mais pas n'importe lesquelles ; uniquement celles relatives à la même orientation, et relatives à la puissance électrique. Il n'est pas envisageable de mixer le rayonnement global et une sortie de type puissance *PV*, les erreurs seraient trop importantes. Une des explications envisageables vient du fait que les masques solaires liés aux détecteurs de rayonnement global et ceux du mur *PV* sont différents, ce qui complique largement l'utilisation croisée du rayonnement global et de la puissance électrique. Dans les conditions optimales, les résultats obtenus sont systématiquement meilleurs que ceux que l'on aurait obtenus avec une approche plus simpliste basée sur la persistance. Si l'on regarde plus en détails, pour prédire du *AC* comme du *DC*, on constate qu'il est conseillé d'utiliser en entrée du *PMC* uniquement des grandeurs pré-onduleur *DC*. L'onduleur doit certainement induire une filtration de la puissance *PV* synonyme de perte d'information. Sur la Figure 4-35, on peut voir l'allure de la prédiction sur un cas particulier d'orientation de façade et ainsi la comparer à celle de la mesure.

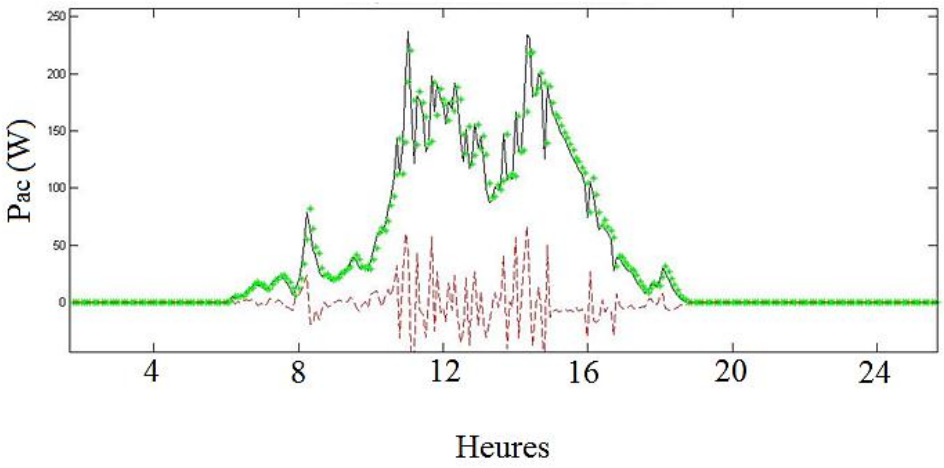


Figure 4-35 : prédiction en juin 2010 de la puissance *AC* post-onduleur (P_{AC}) pour l'orientation Sud (points verts) et erreur de prédiction associée (pointillés rouges). Les grandeurs P_{AC} sont mises en entrée du *PMC*

On voit sur cette figure que la prédiction issue du *PMC*, correspond dans une certaine mesure à une persistance améliorée, la courbe verte semble être l'équivalent de la noire translatée d'un lag 1 (retard de 5 min). En début de section, on a vu que l'utilisation d'outils

de stationnarisation conventionnels n'était pas évidente pour cette échelle temporelle. Cependant, des outils plus simples, accessibles avec les *RNA* pourraient permettre d'améliorer les résultats de prédiction. En effet, il est possible de laisser le *PMC* choisir tout seul sa propre stationnarisation, en utilisant en entrée des indices temporels, qui vont lui permettre d'établir une régression sur le temps du phénomène périodique. L'étude de ce cas particulier fait l'objet du prochain paragraphe.

b. Stationnarisation par indices temporels

Contrairement à l'expérience précédente, nous n'étudierons plus les interactions croisées, et nous nous limiterons aux entrées de réseaux compatibles avec les éléments de sortie. Les deux indices temporels utilisés sont liés à l'heure dans la journée et au jour dans l'année (chapitre 3, section 3-3, p 128). La fonction de transfert de la couche cachée qui donne les meilleurs résultats est la fonction Gaussienne. Ce résultat est sans doute lié à la similarité de cette courbe avec celle du rayonnement global. Seules les données relatives aux modules *mono-Si* sont considérées (voir Tableau 4-29).

		Sorties du réseau					
		P_{ac} *SO*	P_{dc} *SO*	P_{ac} *S*	P_{dc} *S*	P_{ac} *SE*	P_{dc} *SE*
Entrées du réseau	P_{ac} *SO*	30,0-29,5	28,2-28,0				
	P_{dc} *SO*	30,7-30,5	29,3-28,4				
	P_{ac} *S*			19,5-19,3	19,3-18,9		
	P_{dc} *S*			19,9-19,3	19,5-18,8		
	P_{ac} *SE*					30,9-30,1	29,5-28,9
	P_{dc} *SE*					31,5-30,7	29,8-29,0
	Persistance	*31,5*	*29,4*	*20,5*	*20,3*	*31,7*	*30,6*

Tableau 4-29 : erreur de prédiction (*nRMSE* en %) concernant l'utilisation du *PMC* simple et l'utilisation du *PMC* avec indice temporel (*PMC* simple- *PMC* avec indice temporel)

L'utilisation des indices temporels génère une valeur ajoutée à la qualité de la prédiction, les résultats sont systématiquement améliorés par cet outil. Cet apport est mis en valeur sur le Tableau 4-30. Une valeur positive du gain indique une amélioration de la prédiction liée aux indices temporels.

		Sorties du réseau					
		P_{ac} *SO*	P_{dc} *SO*	P_{ac} *S*	P_{dc} *S*	P_{ac} *SE*	P_{dc} *SE*
Entrées du réseau	P_{ac} *SO*	0,50	0,20				
	P_{dc} *SO*	0,20	0,90				
	P_{ac} *S*			0,20	0,40		
	P_{dc} *S*			0,60	0,70		
	P_{ac} *SE*					0,80	0,60
	P_{dc} *SE*					0,80	0,80

Tableau 4-30 : gain sur l'erreur de prédiction (*nRMSE* en %) entre l'utilisation du *PMC* simple et l'utilisation du *PMC* avec indice temporel

Le gain moyen est de plus de 0,5 %, garantissant un réel avantage quant à l'utilisation de ce mode de stationnarisation. Les façades *SO* et *SE* sont très difficiles à modéliser car les masques solaires ne sont pas négligeables et polluent la détection en début et en fin de journée (voir Figure 4-36). Ainsi, à certaines heures, n'est présent que le rayonnement diffus, et non le rayonnement direct.

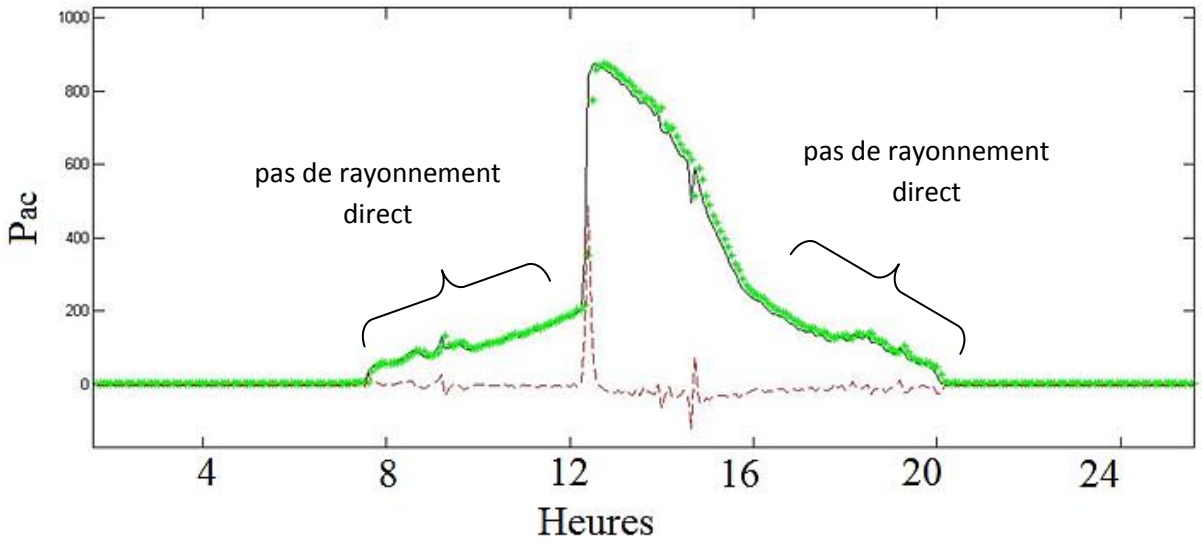


Figure 4-36 : prédiction de la puissance *AC* post-onduleur (P_{AC}) pour l'orientation *SO* (points verts) et erreur de prédiction (pointillés rouges). Utilisation du *PMC* avec les indices temporels. Mur de Vignola, août 2010

Dans les manipulations que l'on vient de présenter, les sorties des *PMC* étaient exploitées directement. Dans la prochaine partie, nous allons essayer d'utiliser un modèle de connaissance pour essayer de s'affranchir de la nécessité de disposer d'un historique de mesures relatives aux puissances *PV* qui fait souvent défaut.

5.3. Perspective et développement possible

Souvent les données de puissances électriques *PV* ne sont pas disponibles. Pour les prédire, il faut donc utiliser la prédiction du rayonnement global, puis la convertir en prédiction de puissance électrique en utilisant par exemple le rendement des modules. Dans notre cas, le réseau de neurone est utilisé pour prédire le rayonnement global puis une connaissance empirique basée sur le rendement du module considéré est utilisée pour la conversion électrique. Il suffit de posséder un détecteur de rayonnement global avec un historique de mesures conséquent, pour réussir à prédire la puissance d'un module *PV*. Pour cette étude, les indices temporels et la fonction de transfert gaussienne sont utilisés. Concernant la conversion rayonnement global vers puissance électrique, il est possible d'utiliser des modélisations sophistiquées (chapitre 1, section 4-3, p 53), mais souvent des approximations linéaires de type $P_{AC/DC}=a$.Ray_Glob+b sont suffisantes, a et b étant des paramètres de régression à estimer. Ces coefficients sont disponibles sur le Tableau 4-31, concernant les deux natures de détecteurs (*a-Si* et *mono-Si*) et les trois orientations.

	a-Si						*mono Si*					
	P_{ac} SE	P_{dc} SE	P_{ac} S	P_{dc} S	P_{ac} SO	P_{dc} SO	P_{ac} SE	P_{dc} SE	P_{ac} S	P_{dc} S	P_{ac} SO	P_{dc} SO
a	0,39	0,43	0,35	0,39	0,40	0,45	0,60	0,65	1,25	1,32	0,27	0,63
b	4,1	5,6	3,4	4,8	2,8	4,08	52,8	61,5	16,8	22,2	15,8	48,0

Tableau 4- 31 : coefficients de régression linéaire pour la puissance électrique des six zones du mur *PV* de Vignola

On remarque que le coefficient a qui est proportionnel au rendement des modules, est supérieur dans le cas mono-Si, conformément à ce que l'on avait proposé dans le chapitre 1 section 4-3. La Figure 4-37a montre le lien qu'il peut y avoir entre la puissance électrique *DC* et le rayonnement global pour l'orientation Sud. Le même type de courbe peut être tracé pour P_{AC} ainsi que pour les autres orientations.

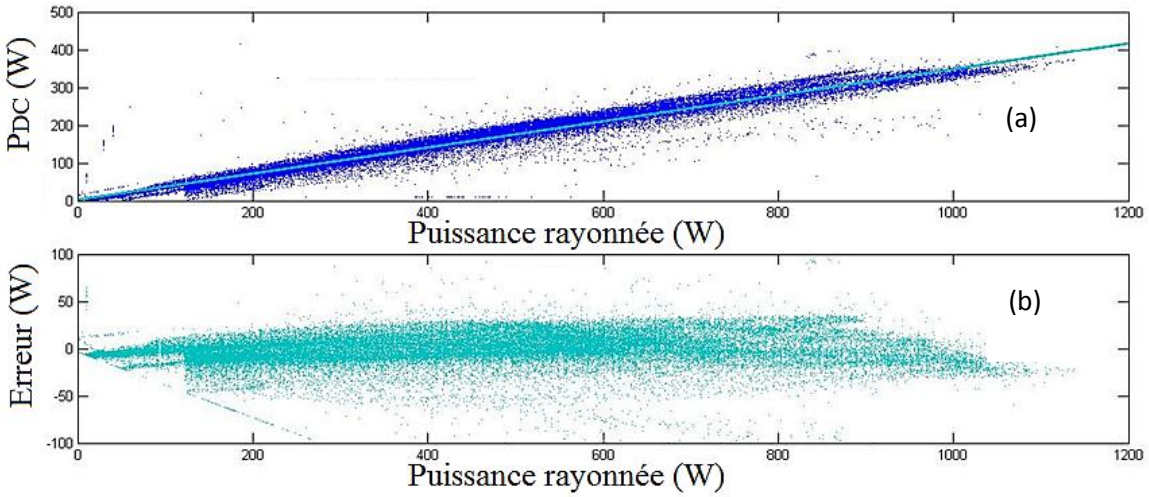


Figure 4-37 : (a) relation entre le rayonnement global et la puissance électrique *DC* sur la façade plein sud. (b) erreur engendrée par l'approximation linéaire liant les puissances rayonnées et P_{DC}. Mur de Vignola (1/03/2009 - 31/08/2010)

Sur la Figure 4-37b, on constate que les erreurs liées à la modélisation linéaire (aussi appelées résidus) sont majoritairement comprises entre -50W et +50W, ce qui induit une estimation relativement correcte concernant la conversion rayonnement-puissance *PV*. Le Tableau 4-32 décrit les résultats liés au modèle boîte grise construit : prédiction de la puissance solaire, puis conversion en puissance *PV*.

		Sorties du PMC														
					a-Si						*Mono-Si*					
		Glo SE	Glo S	Glo SO	P_{ac} SE	P_{dc} SE	P_{ac} S	P_{dc} S	P_{ac} SO	P_{dc} SO	P_{ac} SE	P_{dc} SE	P_{ac} S	P_{dc} S	P_{ac} SO	P_{dc} SO
Entrées du réseau	*Glo SE*	23,1			53,3	51,8					151	143				
	Glo S		19,5				26,4	25,7					47,0	47,2		
	Glo SO			*20,7*					*36,6*	*145*					*160*	*143*
	persist	24,5	20,8	21,8	25,2	21,9	22,6	19,0	24,4	20,5	31,7	30,6	20,5	20,3	31,5	29,4

Tableau 4-32 : erreur de prédiction (*nRMSE* en %) de la puissance *PV* obtenue avec un *PMC* et une connaissance experte pour la conversion du rayonnement global en puissance électrique

Ce tableau décrit l'erreur de prédiction engendrée par l'utilisation combinée du *PMC* et de la connaissance a priori d'une règle de conversion du rayonnement global en puissance électrique. Le *mono-Si* semble beaucoup plus compliqué à prédire, certainement car les trois

panneaux (formé chacun de 9 modules) le composant, se situent relativement proches du sol, engendrant un masque de la part du bâtiment qui n'est pas négligeable. De plus, le modèle linéaire de conversion utilisé montre ces limites et ne permet pas une bonne prédiction. La méthodologie semble applicable uniquement pour la façade plein sud et avec la technologie *a-Si*. L'erreur de prédiction est « presque équivalente » à une prédiction de type persistance. Cette méthodologie encore instable peut être utile pour prédire la puissance virtuelle générée sur un site dépourvu de module *PV*. La modélisation étant grossière, pour affiner le modèle, il faudrait entrer beaucoup plus dans le détail de la prise en compte des masques solaires, et utiliser des modèles de conversion de puissance rayonnée-puissance *PV* plus aboutis. Dans ce cas-là, les prédictions seraient certainement améliorées, mais la méthodologie de prédiction se complexifierait énormément.

5.4. Conclusion

Nous avons vu dans cette section que les méthodes de stationnarisation utilisées jusqu'ici n'étaient pas bénéfiques à l'horizon m+5 (*nRMSE* augmenté de près de 1 %). Il est plus pertinent de se servir de la série brute et non de l'indice ciel clair ou de clarté. Par contre, le fait de recourir aux indices temporels permet de mieux prendre en considération les différentes saisonnalités. Il s'agit d'un processus de stationnarisation que l'on peut qualifier d'indirect. On peut noter également que la méthodologie basée sur les *PMC* améliore les résultats obtenus (*nRMSE* amélioré de plus de 1 %) avec une approche plus simple et basée sur la persistance. Nous avons vu aussi qu'il est préférable de mettre en entrée du *PMC* des grandeurs « pré-onduleur » plutôt que « post-onduleur », cela améliore les résultats. Concernant le test de prédiction de la puissance *PV* à partir de l'estimation du rayonnement global des améliorations sont à envisager, car les manipulations présentées ici ne représentent que les premières perspectives envisageables pour ce travail. Il conviendra certainement, à l'avenir, d'utiliser des modes de conversion beaucoup plus aboutis. L'approche utilisée dans ce manuscrit semble convenir surtout pour les zones où le masque solaire ne perturbe pas trop les mesures.

6. Conclusion des manipulations

A l'issu de ces manipulations, un élément peut surprendre : suivant l'horizon considéré, les résultats peuvent être totalement différents. Pour cette raison, il faut bien faire attention aux méthodes appliquées et aux résultats attendus. Durant cette thèse, notre approche a évolué : au début, nous souhaitions optimiser les prédicteurs pour diminuer le plus possible l'erreur commise. Ce type d'approche est scientifiquement intéressant, mais cette spécialisation des prédicteurs va à l'encontre de la généralisation des résultats. Dans une seconde étape, nous avons alors choisi de nous attarder non plus particulièrement à l'amélioration de la technique, mais à ne pas travailler sur une méthode trop complexe qui, de ce fait, ne serait pas utilisable par d'autres équipes. Intellectuellement, il est toujours intéressant de diminuer de 0,1 % le *nRMSE* de prédiction, cependant d'un point de vue énergétique, il est probable que les personnes qui ont besoin de ces prédictions soient peu intéressées par le gain occasionné. Ce qui est recherché avant tout est une méthode simple à mettre en place, rentable et exploitable sur plusieurs sites : la sélection des données et des paramètres des modèles choisis doit être parcimonieuse.

Nous allons maintenant faire la synthèse de ces manipulations en essayant de dégager le rapport efficacité/complexité de chaque horizon. Nous comparerons ensuite ces résultats avec ceux issus des différentes études bibliographiques.

6.1. Synthèse des meilleurs prédicteurs en fonction des horizons

Au travers de cette thèse, nous avons dégagé certaines méthodologies relatives à la prédiction à horizon du rayonnement global et de la puissance *PV*. On peut conclure que ces deux types de prédictions sont relativement équivalents dans la méthodologie à mettre en œuvre. Pour prédire la puissance *PV*, il y a deux manières de procéder : soit on utilise des séries temporelles de puissance *PV* pour entraîner et prédire la grandeur endogène, soit on prédit le rayonnement global, que l'on convertit en puissance par le biais d'un modèle de connaissance approprié. Au final, la démarche utilisant le (ou les) *PMC* est identique. Dans le

Tableau 4-33 sont repris et synthétisés ce que nous préconisons pour les différents horizons de prédiction.

Horizon	Stationnarisation	Variables exogènes	Prédicteur requis	Difficulté	Erreur[15] (*nRMSE*)
j+1	*CSI*	Mesures : *Su,N,RH*	*PMC*	++	23,4%
h+1	*CSI**	Estimations ALADIN : *N, P, RP*	Hybride : *PMC+ARMA*	+++	14,9%
h+24	*k*	-	*PMC* multi-sorties	+	27,3%
m+5	*Indice temporel*	-	*PMC*	+	20,2%

Tableau 4-33 : synthèse des résultats présentés dans ce manuscrit

Dans le cas j+1, nous avons effectué des manipulations de prédiction de rayonnement global et d'énergie *PV* qui ont permis de mettre en place une méthodologie efficace de prédiction à base de *PMC*. De plus, nous avons remarqué qu'en l'absence de stationnarisation, il était équivalent d'utiliser un *PMC* ou une approche *ARMA*. Ces deux prédicteurs montrent des résultats globalement similaires. Pour ces deux approches, le fait de stationnariser apporte une valeur ajoutée à la qualité des prédictions. Nous avons montré que l'utilisation de l'indice de clarté était bénéfique. Cependant, pour une plus grande fiabilité, il est souhaitable d'utiliser l'indice de ciel clair couplé avec une correction des variations saisonnières à base de coefficients périodiques (gain sur le *nRMSE* proche de 1 % par rapport au cas non stationnarisé). La dernière amélioration que nous avons exposée concerne l'utilisation des variables exogènes. Néanmoins, ces dernières sont surtout efficaces durant les périodes à fortes occurrences nuageuses, sinon leur intérêt est minime (gain concernant le *nRMSE* de ~1 % en mars et en avril et proche de 0 % en mai et en juin). Lors de l'étude j+1, nous avons aussi constaté que les *PMC* étaient capables d'utiliser des données d'apprentissages issues de zones géographiques différentes des zones de prédiction. Cette série de manipulation tend à confirmer le fait que les *PMC*, et plus particulièrement la méthodologie que nous avons utilisée, sont favorables à la prédiction du rayonnement global ou de la puissance *PV*.

Concernant le cas h+1, nous l'avons considéré comme l'horizon dont l'impact est le plus important pour un gestionnaire. C'est aussi celui qui bénéficie le plus de l'apport des prédictions sophistiquées mises en œuvre. En effet, il n'y a que dans le cas horaire que nous estimons légitime d'utiliser un modèle complexe. Pour les autres horizons, la complexification des méthodologies induit des améliorations trop minimes pour les justifier. Le fait de stationnariser les séries temporelles permet de diminuer l'erreur de prédiction

[15] Moyenne sur tous les sites étudiés.

(*nRMSE* diminué de 1,7 %), et l'utilisation de variables exogènes (modèle prédictif *ALADIN*) a démontré une amélioration des résultats durant les tests de prédictions (*nRMSE* diminué de 0,7 %). Cet horizon a surtout permis de comprendre l'intérêt de combiner les différents prédicteurs entre eux. Cette hybridation entre *ARMA* et *PMC* est la dernière option d'amélioration testée, c'est aussi celle qui donne les meilleurs résultats et donc la plus probante (*nRMSE* diminué de 3,5 %). Pour cet horizon, si l'on compare notre méthodologie globale (la stationnarisation, les variables exogènes et l'hybridation des résultats) avec une prédiction simple telle la persistance, on obtient une réduction de l'erreur de prédiction de plus de 11 %. La saison où le gain est le plus important est l'automne (*nRMSE* diminué de 16,1 %), suivi de l'hiver (*nRMSE* diminué de 15,8 %), de l'été (*nRMSE* diminué de 10,4 %) et du printemps (*nRMSE* diminué de 10 %).

Pour l'horizon h+24, les conclusions sont légèrement différentes. En effet, bien que la stationnarisation soit intéressante comme dans les deux cas précédents, nous n'avons pas trouvé de réel intérêt à utiliser les variables exogènes. De plus, les réseaux de *PMC* comme proposés dans cette étude, ne conduisent pas à de bons résultats. Il convient d'utiliser un *PMC* pourvu d'autant de sorties que de prédictions souhaitées (pour neuf heures, neuf neurones de sortie).

L'horizon m+5 remet également en cause ce qui a été démontré pour les deux premiers horizons. En effet, cette fois-ci, c'est le mode de stationnarisation des séries temporelles qui n'est plus pertinent. Il est plus adéquat de manipuler les indices temporels que les indices de types *CSI* et *k*. Ces indices permettent de diminuer le *nRMSE* de 0,7 %. Les résultats de prédiction de rayonnement global et de puissance *PV* sont améliorés par rapport à une approche classique, cependant le gain occasionné est minime (*nRMSE* diminué d'environ 1,5 %). En première approximation, il n'est certainement pas nécessaire d'utiliser un modèle abouti formé de *PMC* et d'indice temporel, une simple persistance peut suffire. Dans le prochain paragraphe, nous allons positionner les conclusions que l'on vient de résumer dans le contexte général bibliographique des prédictions de rayonnement global et de puissance *PV*. Nous verrons ainsi si nos approches modifient ou corroborent les conclusions de ces publications.

6.2. Synthèse par rapport à l'état de l'art

Dans tous les articles bibliographiques relatifs à l'estimation du rayonnement global ou de l'énergie *PV*, on constate que les erreurs liées aux prédictions (mensuelles, journalières, horaires et minutes) diffèrent d'un site et d'un auteur à l'autre. Les méthodologies de prédictions sont, en règle générale, si différentes qu'elles sont difficiles à comparer. De plus, les estimations des erreurs sont hétérogènes : erreur de prédiction sur certains jours échantillonnés ou sur une large période, test effectué sur le cumul des prédictions, utilisation de paramètres d'erreur non usuels, etc. Pour s'affranchir de toutes ces spécificités qui entravent la comparaison, nous avons axé notre travail sur une méthodologie générale de l'estimation de l'erreur :

- test de prédiction sur une large période et non sur des journées « bien choisies » ;
- utilisation du *RMSE* pour pénaliser les gros écarts ;
- normalisation du *RMSE* pour pouvoir confronter les comparaisons sur plusieurs sites ;
- pas de cumul des prédictions (sauf études spécifiques) qui a pour effet de moyenner l'erreur et de la diminuer ;
- répartition des erreurs en fonction des saisons car la consommation électrique n'est pas la même tout au long de l'année ;
- tests réalisés sur plusieurs sites (dès que l'étude s'y prête), afin de s'affranchir des phénomènes climatologiques régionaux ;
- utilisation d'une prédiction test de référence et utilisation d'un prédicteur naïf afin de réellement statuer sur la qualité de notre méthodologie.

Cette méthodologie de l'estimation de l'erreur n'est pas celle utilisée dans la plupart des articles, il n'y a pas de réel consensus. A ce titre, il est difficile de comparer nos résultats avec ceux existants dans la littérature. Ce que nous présentons ici est une comparaison des différents prédicteurs élaborés et testés durant cette thèse afin d'en proposer une hiérarchisation.

Pour les horizons j+1 et h+1, nos résultats sont en partie conformes à la littérature. En effet, les *PMC* sont adaptés et permettent d'effectuer des prédictions de rayonnement global avec une erreur acceptable [Mubiru et Banda 2008] et sont aussi applicables aux zones montagneuses [Bosch et al. 2008]. Concernant la hiérarchisation des modèle *PMC* et *ARMA*, nous nous démarquons des résultats bibliographiques traditionnels [Reddy et Ranjan 2003 ; Mellit et al. 2005 ; Elminir et al. 2007]. En Effet, sans étude approfondie sur la stationnarisation des séries temporelles utilisées, nous ne pensons pas qu'il soit aisé de différencier *ARMA* et *PMC*. Cependant nous rejoignons Berhangh et al. [2010] sur le fait que l'utilisation des variables exogènes améliore les résultats des *PMC*. Comme dans la littérature, nous nous sommes aperçus que les approches pertinentes dans le cas de la prédiction du rayonnement l'étaient tout aussi dans le cas de la prédiction de la puissance *PV* [Mellit et al. 2006 ; Benghanem et Mellit 2010]. De plus, nous avons confirmé, dans le cadre de nos expérimentations, que les *PMC* pouvaient être utilisés quand la zone d'apprentissage et celle de test n'étaient pas identiques [Sozen et al. 2004]. Nous avons choisi la persistance comme prédicteur naïf. Bien qu'elle ne soit pas systématiquement utilisée dans la littérature, nous pensons que la persistance permet de juger correctement de la validité des techniques complexes. De plus, l'utilisation du modèle ciel clair et la correction des variations saisonnières à base de coefficients périodiques n'ont pas souvent été étudiés avec la prédiction du rayonnement global. Aux vues des résultats présentés ici, leur investigation semble prometteuse. Nous avions détaillé, dans le chapitre 3 (section 4) les principaux auteurs ayant étudié et démontré l'apport de l'hybridation des méthodes. Bien que notre approche soit radicalement différente des leurs, nos conclusions restent identiques. Enfin, concernant les horizons h+24 et m+5, il existe encore trop peu d'études utilisant les *PMC*. Cependant tout comme Mellit et Pavan [2010] et Chaabene et Ben Ammar [2008] nous pensons et avons montré que les *PMC* étaient adaptés à ces situations. De plus, notre approche avec l'utilisation des indices temporels semble efficace et mérite des approfondissements. En résumé, nos résultats sont relativement similaires à ceux que l'on trouve dans la littérature. Sous certains aspects ils sont complémentaires et, parfois, agrémentent la technique de prédiction d'outils novateurs. Lors du prochain chapitre qui constituera la conclusion générale, nous allons synthétiser la liste de tous ces outils en fonction de leur ordre d'apparition dans le manuscrit et des horizons considérés.

7. Bibliographie

Benghanem M et Mellit A. Radial Basis Function Network-based prediction of global solar radiation data: Application for sizing of a stand-alone photovoltaic system at Al-Madinah, Saudi Arabia. Energy, n°35(9), p 3751-3762, 2010.

Behrang MA, Assareh E, Ghanbarzadeh A, Noghrehabadi AR. The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data. Solar Energy, n°84(8), p 1468-1480, 2010.

Bosch J, Lopez G, Batlles F. Daily solar irradiation estimation over a mountainous area using artificial neural networks. Renewable Energy, n°7;33(7), p 1622-1628, 2008.

Chaabene M, Ben Ammar M. Neuro-fuzzy dynamic model with Kalman filter to forecast irradiance and temperature for solar energy systems. Renewable Energy, n°33(7), p 1435-1443, 2008.

Chaouachi A, Kamel RM, Ichikawa R, Hayashi H et al. Ensemble-Based solar power generation Short-Term forecasting. Neural network, n°54, 2009.

Elminir HK, Azzam YA, Younes FI. Prediction of hourly and daily diffuse fraction using neural network, as compared to linear regression models. Energy, n°32(8), p 1513-1523, 2007.

Mellit A, Benghanem M, Arab AH, Guessoum A. An adaptive artificial neural network model for sizing stand-alone photovoltaic systems: application for isolated sites in Algeria. Renewable Energy, n°30(10), p 1501-1524, 2005.

Mellit A, Benghanem M, Kalogirou SA. An adaptive wavelet-network model for forecasting daily total solar-radiation. Applied Energy, n°83(7), p 705-722, 2006.

Mellit A, Pavan AM. A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy. Solar Energy, n°. 84(5), p 807-821, 2010.

Mubiru J et Banda E. Estimation of monthly average daily global solar irradiation using artificial neural networks. Solar Energy, n° 2;82(2), p 181-187, 2008.

Paoli C, Voyant C, Muselli M, Nivet ML. Forecasting of preprocessed daily solar radiation time series using neural networks, Solar Energy, n°84(12), p 2146-2160, 2010.

Reddy KS et Ranjan M. Solar resource estimation using artificial neural networks and comparison with other correlation models. Energy Conversion and Management, n°44(15), p 2519-2530, 2003.

Sözen A, Arcaklioglu E, Özalp M. Estimation of solar potential in Turkey by artificial neural networks using meteorological and geographical data. Energy Conversion and Management, vol. 45, no. 18-19, p. 3033-3052, 2004.

Voyant C, Paoli C, Muselli M, Nivet ML. Optimization of an Artificial Neural Network (ANN) Dedicated to the Daily Global Radiation and PV Plant Production Forecasting Using Exogenous Data. 25th European Photovoltaic Solar Energy Conference and Exhibition (25th EU PVSEC). Valence, 2010.

Voyant C, Muselli M, Paoli C, Nivet ML, P Poggi. Predictability of PV power grid performance on insular sites without weather stations: use of artificial neural networks. 24th European Photovoltaic Solar Energy Conference, Hambourg, 2009.

Voyant C, Muselli M, Paoli C, Nivet ML. Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation. Energy, n°36, p 348-359, 2011.

Voyant C, Muselli M, Paoli C, Nivet ML. Hybrid method dedicated to the hourly global horizontal radiation forecasting for mediterranean locations. Energy conversion and management, in press, 2011.

Voyant C, Muselli M, Paoli C, Nivet ML. Use of ALADIN forecast data, artificial neural network and ARMA model to predict hourly global radiation time series. Solar energy, in press, 2011.

CONCLUSION

SOMMAIRE

Comme nous l'avons vu en introduction de ce manuscrit, l'intégration d'énergies renouvelables intermittentes à un réseau électrique est très délicate, et particulièrement dans le cas insulaire. Une possibilité en vue de cette intégration réside dans la caractérisation des différentes ressources. Cette étude s'inscrit donc dans la recherche de prédicteurs de la ressource solaire et *PV* afin de quantifier l'énergie disponible pour en faciliter la gestion au sein d'un réseau. Nous avons vu dans le chapitre précédent que cette démarche essentielle est tributaire de différentes modalités d'approche comme, notamment, l'horizon considéré, la stationnarité des données, la sélection des variables à utiliser, le calcul de l'erreur, etc. Nous avons essayé de tester différentes techniques de prédiction, sur quatre horizons différents, susceptibles d'intéresser un gestionnaire de réseau (j+1, h+24, h+1 et m+5).

La prochaine section propose de revenir sur les différentes méthodologies et manipulations présentées dans ce rapport avec l'objectif de les hiérarchiser en fonction de leur performance et de l'horizon de prédiction. Enfin la seconde section s'attachera à proposer une série de perspectives et d'améliorations qui pourront faire suite à ce travail doctoral.

1. Bilan des manipulations

Pour utiliser correctement les différents prédicteurs de rayonnement global, il a fallu dégager des connaissances précises en la matière. Le formalisme mathématique empirique énoncé dans le premier chapitre est utilisé dans de nombreux travaux de recherche. Il nous a permis d'élaborer différents modèles de rayonnement (ciel clair, fraction de diffusé et inclinaison) à partir d'un travail de synthèse sur les différents modèles déjà existants. Concernant les modèles utilisés (*SOLIS*, *CLIMED2* et Klutcher), ce ne sont pas forcément les plus communément choisis, ni les plus efficaces. Cependant, ils sont apparus comme les plus performants pour traiter nos données et s'adapter aux sites choisis.

Dans le deuxième chapitre nous énonçons certaines définitions relatives aux séries temporelles, leurs analyses, et leurs prédictions. Comme nous l'avons vu, le rayonnement global horizontal, peut se concevoir comme une succession de mesures et ainsi être interprété comme une série temporelle « ordinaire » pour laquelle tous les outils d'analyse (*ARMA*, *RNA*, Markov, etc.) sont potentiellement dédiés. Une grande part du travail de thèse est

relative à la compréhension et à l'utilisation de ces modèles dans le cas du rayonnement global.

Dans le troisième chapitre, forts des connaissances liées au rayonnement global et aux séries temporelles, nous avons établi une approche de prédiction des séries temporelles de rayonnement global avec *PMC*. Nous avons axé notre travail sur trois grands thèmes de recherche : l'optimisation des paramètres du *PMC*, la sélection des variables d'entrée et la stationnarisation des grandeurs utilisées. Peu d'études synthétisant ces trois thématiques, il nous a paru primordial de concevoir une méthodologie qui ne se limite pas à l'utilisation d'un *PMC,* mais plutôt une méthodologie de prédiction dont le *PMC* est un élément parmi d'autres.

Cette méthodologie a permis de prédire le rayonnement global et la puissance *PV* sur différents horizons de prédiction. Les résultats que nous avons obtenus sont restitués dans le dernier chapitre. La première étude est relative à l'horizon journalier (j+1). Grace à elle, nous avons pu mieux comprendre comment utiliser les *PMC* et les autres prédicteurs étudiés. Nous avons montré que les prédicteurs de type Markov, Bayes, *k-NN* sont quasi équivalents en terme de prédiction. Ils sont tout de même moins appropriés (dans le cadre de cette étude) à la prédiction de rayonnement global que ne le sont les modèles *ARMA* ou *PMC*. Durant cette étude, nous avons compris que pour prédire correctement la ressource, il est nécessaire d'effectuer un prétraitement de stationnarisation. Cette étape est primordiale et permet de potentialiser la modélisation *RNA* ou *ARMA*. L'implémentation de données exogènes météorologiques dans les *PMC* semble aussi accroître la qualité de la prédiction. De plus, nous avons montré que les conclusions issues de la prédiction de rayonnement global horizontal restent tout aussi valables dans le cas de cellules *PV* que de prédiction d'énergie *DC* ou *AC*.

Concernant la prédiction du rayonnement horaire (h+1), la méthodologie de prédiction présentée est certes assez contraignante, mais elle donne des résultats largement supérieurs aux autres techniques testées. On retiendra de cet horizon de prédiction la nécessité de stationnariser par indice ciel clair et par coefficients périodiques, les outils à disposition pour quantifier la désaisonnalisation (test de Fisher et *CV*), l'utilisation du modèle prédictif *ALADIN* en entrée de *PMC*, et surtout l'hybridation des modèles qui garantit une amélioration des résultats.

Le fait d'étudier des horizons de prédictions profonds (h+24), modifie quelque peu l'approche à envisager. Ce type de prédiction est très difficile à appréhender : rechercher la finesse de la prédiction sur une heure vingt-quatre heures à l'avance est soumis à trop de paramètres pour espérer obtenir les mêmes résultats que pour les horizons h+1 ou j+1. Nous avons dégagé qu'il était souhaitable de stationnariser les données. Pour ce faire, il est intéressant d'utiliser l'indice de clarté. Il convient aussi de rester sur une approche classique ne faisant intervenir qu'un seul *PMC* mais possédant plusieurs sorties. L'utilisation des variables exogènes est quant à elle délicate. Le ratio entre performance et complexité nous fait préconiser, en l'état actuel de nos connaissances, de ne pas y recourir. On peut estimer que l'approche simple où l'on implémente au fil de l'eau les données de rayonnement global sur un ou deux jours en entrée du *PMC,* avec l'utilisation d'indice de clarté, pourrait être utilisée assez aisément. Cela permettrait à un gestionnaire de réseau de connaître la veille le profil horaire de la journée du lendemain. Le recourt à une approche plus sophistiquée n'améliorerait pas suffisamment les résultats pour justifier la complexité de sa mise en œuvre.

Concernant l'horizon m+5, nous avons pu observer que la stationnarisation des données d'entrée est une étape délicate. Il est plus pertinent d'utiliser la série brute et non le *CSI* ou le *k*. Le fait d'utiliser les indices temporels (e.g. régression sur variables indicatrices) permet de mieux tenir compte des différentes saisonnalités. La méthodologie utilisant les *PMC* améliore quelque peu les résultats obtenus par rapport à une approche plus simple, basée sur la persistance. Concernant le test de prédiction de la puissance *PV* à partir de la prédiction du rayonnement global, des améliorations sont à envisager, car les manipulations présentées ici, ne représentent qu'une première perspective.

A l'issue de toutes ces manipulations, force est de constater que suivant l'horizon considéré, les résultats peuvent totalement diverger. C'est pourquoi il est nécessaire d'être vigilent quant aux méthodes appliquées et aux résultats attendus. Durant cette thèse, notre positionnement a évolué. Au début, nous souhaitions optimiser les prédicteurs pour diminuer au maximum l'erreur commise. Ce type d'approche est scientifiquement intéressant, mais cette spécialisation des prédicteurs va à l'encontre de la généralisation des résultats. La supériorité de ces modèles est une conséquence positive de ce que l'on appelle le « data mining » voire le « data snooping » (White, H., 2000. A Reality Check for Data snooping. Econometrica vol 68, n°5, p1097-1126, 2000). En effet, il est nécessaire que les données utilisées soient suffisamment nombreuses pour que les inférences statistiques et les modèles

découverts soient valides (problème d'overfitting). Forts de ce constat, mais toujours soucieux de l'amélioration de la technique, nous nous sommes concentrés sur l'élaboration d'une méthode qui ne soit pas trop complexe et adaptable concrètement. La complexification nécessaire pour diminuer au maximum la marge d'erreur est difficilement compatible avec les applications et les objectifs que nous envisagions pour cette étude. Intellectuellement, l'enjeu de parvenir à diminuer même de 0,1 % le *nRMSE* de prédiction, est bien sûr intéressant. Cependant il est probable que les personnes qui ont besoin de ces prédictions soient plus intéressées par un meilleur rapport contraintes/résultats. Ce qui est recherché avant tout est une méthode simple à mettre en place, rentable et exploitable sur plusieurs sites.

2. Perspectives et améliorations

Les résultats de cette étude ouvrent de nombreuses perspectives, à court ou moyen terme. La principale étant certainement d'essayer d'égaler la qualité de prédiction déjà existante, avec des modèles plus simples (surtout concernant le cas h+1). Pour ce faire, dans un premier temps, des améliorations relativement rapides à mettre en place sont à envisager :

- concernant l'horizon h+24, en travaillant sur une variété de réseaux plus propices à l'utilisation de délais (*TDNN* ; time delay neural network), on pourrait certainement améliorer encore d'avantage les résultats en intégrant plus facilement les différentes journées à considérer ;

- concernant les horizons h+1, m+5, la prédiction de la puissance *PV* doit être faite avec des modèles de conversion entre le rayonnement global et la puissance *DC/AC* plus sophistiqués.

D'autres améliorations, plus complexes à mettre en place, peuvent cependant être pressenties :

- nous avons montré que notre méthodologie de prédiction basée sur l'emploi d'un prétraitement et d'une modélisation connexionniste est suffisamment robuste pour permettre de compenser la perte d'information engendrée par la non-cohérence spatiale entre test et apprentissage. Ce résultat que nous n'avons pu exploiter correctement dans le cadre de cette thèse pourrait ouvrir une nouvelle voie de recherche sur la prédiction à horizon sur des

sites non pourvus en historique de mesures. Cette approche pourrait permettre de nouvelles investigations liées au foisonnement ;

- l'hybridation des modèles *ARMA* et *PMC* donne de bons résultats au niveau h+1. Cependant, le mixage effectué doit pouvoir être amélioré notamment par l'utilisation de règles de sélection plus abouties. Les inférences Bayesiennes et une sélection basée soit sur les facteurs de Bayes, soit sur la vraisemblance marginale (appelée « evidence » en anglais) pourraient être un très bon critère de choix. Aux vues de nos recherches, la possibilité de réellement mixer les modèles en utilisant les concepts précédents et une moyenne Bayésienne apparait, elle aussi, comme une piste d'expérimentation intéressante.

D'autres améliorations, encore plus évoluées sont envisageables, on peut citer :

- l'utilisation des SIG (Système d'Information Géographique) dans les prédictions afin de réellement prendre en compte le relief et ainsi mieux quantifier le masque solaire, les zones géographiquement équivalentes, etc. ;

- l'utilisation d'images satellites de rayonnement global. Chaque pixel donne les informations relatives à une zone géographique précise. Une carte de NxN pixels donne simultanément N^2 mesures à l'instant t. Si l'on génère les cartes périodiquement, cela signifie que l'on peut transposer ces cartes en N^2 séries temporelles de rayonnement global. Il devient ainsi possible d'utiliser tout ce qui a été montré dans ce manuscrit (prédiction à horizon). Les N^2 prédictions permettent à leur tour de générer une carte de rayonnement global NxN à l'horizon considéré.

Outre l'ensemble de ces améliorations et des perspectives de travaux de recherche qui en découlent, il existe une application directe des méthodes de prédiction sur un système physique de grande envergure. Le site de Vignola sur Ajaccio, qui est le lieu de nombreuses manipulations présentées dans ce rapport, sera prochainement doté d'une plateforme appelée *MYRTE* (Mission hYdrogène Renouvelable pour l'inTégration au réseau Electrique), dont l'inauguration est prévue en octobre 2011 (voir introduction, section 3-1 et 3-2). Il s'agit d'une plateforme de recherche dédiée à l'étude du stockage de l'énergie photovoltaïque dont le but est de développer une stratégie optimale de fonctionnement entre un champ *PV* et une chaîne hydrogène. Cette plateforme s'inscrit dans un projet global nommé « plateforme solaire et de stockage du site de Vignola », ce projet a été labellisé en janvier 2007 par le pôle de

compétitivité CAPENERGIES[16]. L'objectif de la plateforme *MYRTE* est, entre autre, de permettre une injection contrôlée d'énergie électrique dans le réseau. Les premières manipulations couplant la modélisation numérique d'un tel outil (code breveté *ORIENTE* développé par l'université de Corse ; Optimization of Renewable Intermittent Energy with hydrogeN for auTonomous Electrification) et les prédictions basées sur les méthodologies présentées dans ce manuscrit, ont été présentées dans le cadre de la thèse de M. Christophe Darras soutenue en novembre 2010 (Université de Corse ; Modélisation de systèmes hybrides Photovoltaïque / Hydrogène : Applications site isolé, micro-réseau, et connexion au réseau électrique). Les prédictions de la ressource énergétique solaire comme présentées dans ce rapport ont été positionnées en entrée du code *ORIENTE* afin de quantifier et de maîtriser l'évolution des gaz dans les réservoirs et donc l'énergie électrique potentiellement disponible. D'après les premiers résultats, nous pouvons affirmer que la prédiction en période estivale, des différents éléments de la chaîne de production (énergie absorbée par l'électrolyseur, quantité d'hydrogène produite et consommée, temps de fonctionnement de l'électrolyseur, énergie fournie par la pile à combustible, etc.) donne des résultats corrects et très certainement exploitables. Cependant, concernant la période hivernale, la qualité des prédictions n'est pas encore suffisamment satisfaisante. Dès que ce problème sera résolu, ces prédictions pourront être insérées dans le système de contrôle - commande de la plateforme technologique du grand instrument *MYRTE*.

De plus, les méthodes de prédictions présentées, pourraient aussi être utilisables dans le cas du micro-réseau *PAGLIA-ORBA*. Ce projet consiste à développer un micro-réseau électrique, qui associe toutes les formes de stockage de l'énergie à des moyens de production d'électricité d'origine solaire. Les mauvaises prédictions de la ressource solaire sont rédhibitoires au développement d'un tel projet, il est donc nécessaire de continuer les travaux de prédictions afin de rendre les estimations directement exploitables.

[16] Pôle dédié aux énergies non génératrices de gaz à effet de serre ; http://www.capenergies.fr/.

ANNEXES

SOMMAIRE

1. ANNEXE A : Validation des modèles *SOLIS*, *CLIMED2* et Klucher

Avant d'utiliser les approches liées au modèle ciel clair *SOLIS* et au modèle Klutcher pour la pris en compte de l'inclinaison du détecteur, il est nécessaire de les tester. La validation que nous avons retenue, repose sur la comparaison des résultats obtenus avec ces approches (modélisation Matlab™) et ceux issus de modélisations faites avec le logiciel PVSYST™ (disponible sur http://www.pvsyst.com/). Ce dernier est une référence en matière de modélisation de potentiel solaire pour le dimensionnement des systèmes *PV*. La confrontation des résultats est faite pour quatre localisations avec des caractéristiques assez spécifiques comme le montre le Tableau A-1. Les paramètres testés sont, la longitude, la latitude, l'altitude, l'inclinaison, l'azimut et le temps considéré (*TL* ou *TSV*). La période simulée correspond à l'année 2010 par pas horaire.

Localisation	Long	Lat	Alt(m)	Inclinaison (°)	Azimut (°)	TL/*TSV*	*nRMSE* (%)
Ajaccio	8°44'E	41°55'N	4	0	0	TL	**5,80**
Ajaccio	8°44'E	41°55'N	4	0	0	TSV	**5,17**
NewYork	74°00' O	40°43'N	57	50	0	TSV	**7,97**
Nairobi	36°49'E	1°17'S	1640	0	50	TSV	**2,19**

Tableau A-1 : caractéristiques des sites retenus pour la comparaison entre notre méthodologie et le logiciel PVSYS

La colonne *nRMSE* renseigne sur l'écart quadratique existant entre les deux méthodologies étudiées (calculé sur une année entière). Les résultats obtenus avec la méthodologie *SOLIS*, *CLIMED2* et Klucher sont relativement proches de ceux donnés par le logiciel commercial. On a un maximum de 8 % de différence pour le site de New-York. Lors de cette modélisation, l'inclinaison est très importante (50°) et donc assez délicate à prendre en compte. Le détail des profils simulés est représenté par les graphiques de la Figure A-1. Visuellement, on constate qu'il n'y a pas énormément de différence entre les deux modèles testés. Les zones présentant des écarts maximaux sont liées à Ajaccio et New-York et sont essentiellement situées au niveau du midi solaire.

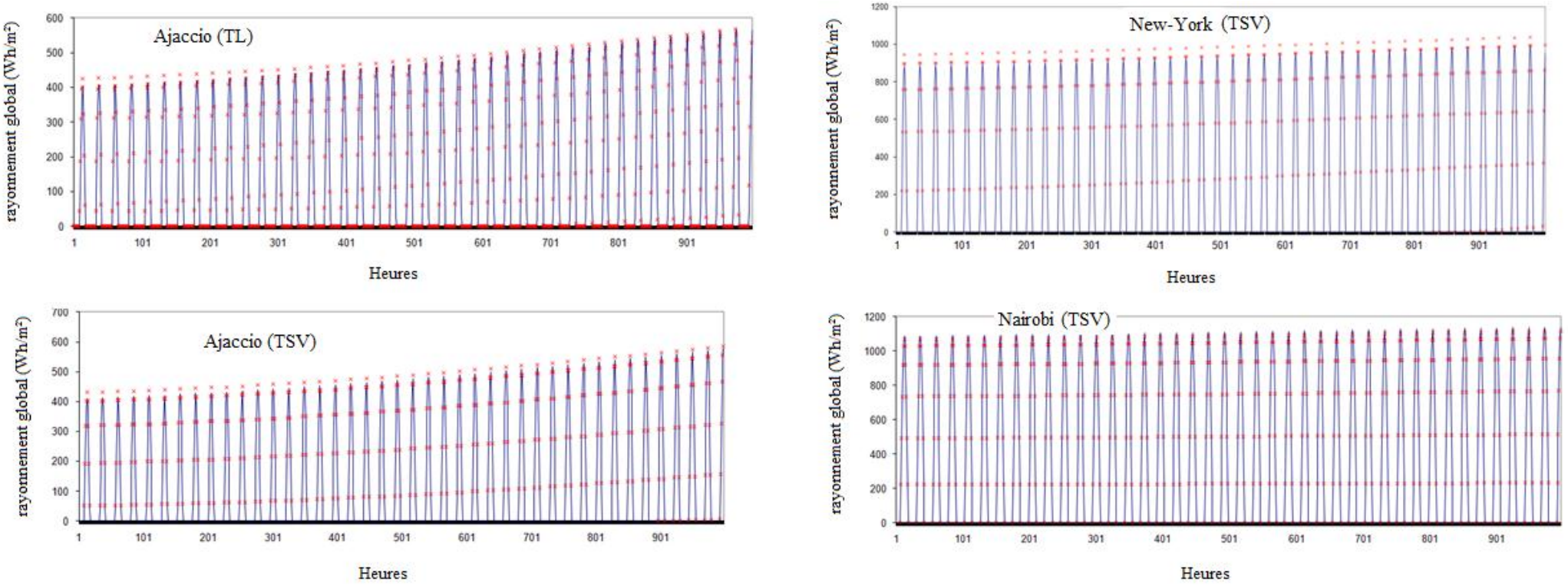


Figure A- 1 : comparaison des simulations issues de PVSYS (trait plein) et de notre méthodologie (croix rouge)

Le test sur ces quatre localisations permet de comparer les modélisations ciel clair sur surface inclinée obtenues avec les deux simulateurs étudiés. Dans ce cas précis, la prise en compte de l'inclinaison est relativement simple à appréhender. En effet, dans notre approche, l'utilisation du modèle *SOLIS* donne séparément la part de rayonnement direct, diffus et total. Le modèle de Klucher permet lui de tenir compte de l'inclinaison pour la partie diffusée. Pour tester le model *CLIMED2* et vérifier notre méthodologie d'acquisition de la fraction de diffusé, le test précédent n'est pas performant. Nous avons choisi d'utiliser pour un même lieu et durant la même période, des données de rayonnement horizontal et inclinées. Le but de la manipulation étant, à partir des données horizontales, de calculer les données inclinées et de les comparer avec la réalité. Nous avons utilisé des mesures horaires issues du laboratoire de Vignola à Ajaccio datant de 1998 (0°, 45° et 60°). La confrontation des mesures inclinées avec les résultats des simulations donnent des résultats satisfaisants. Cependant, l'erreur (*nRMSE*) croît avec l'angle d'inclinaison. La Figure A-2 montre la comparaison entre le calcul et la mesure pour juin 1998. Le modèle est assez fidèle à la réalité. De plus, les erreurs obtenues sont relativement conformes avec ce que l'on peut trouver dans la littérature (Notton G. et al. Performance evaluation of various hourly slope irradiation models using Mediterranean experimental data of Ajaccio. Energy Conversion and Management. 2006 I;47(2):147:173). L'erreur sur le mois de juin (*nRMSE*) entre la simulation et la mesure est de 7,1 % pour le rayonnement global incliné à 45°, et de 18,5 % pour le rayonnement incliné à 60°. Notons qu'avec ces inclinaisons, les masques solaires peuvent générer des interférences, non prises en compte avec la méthodologie de calcul appliquée ici.

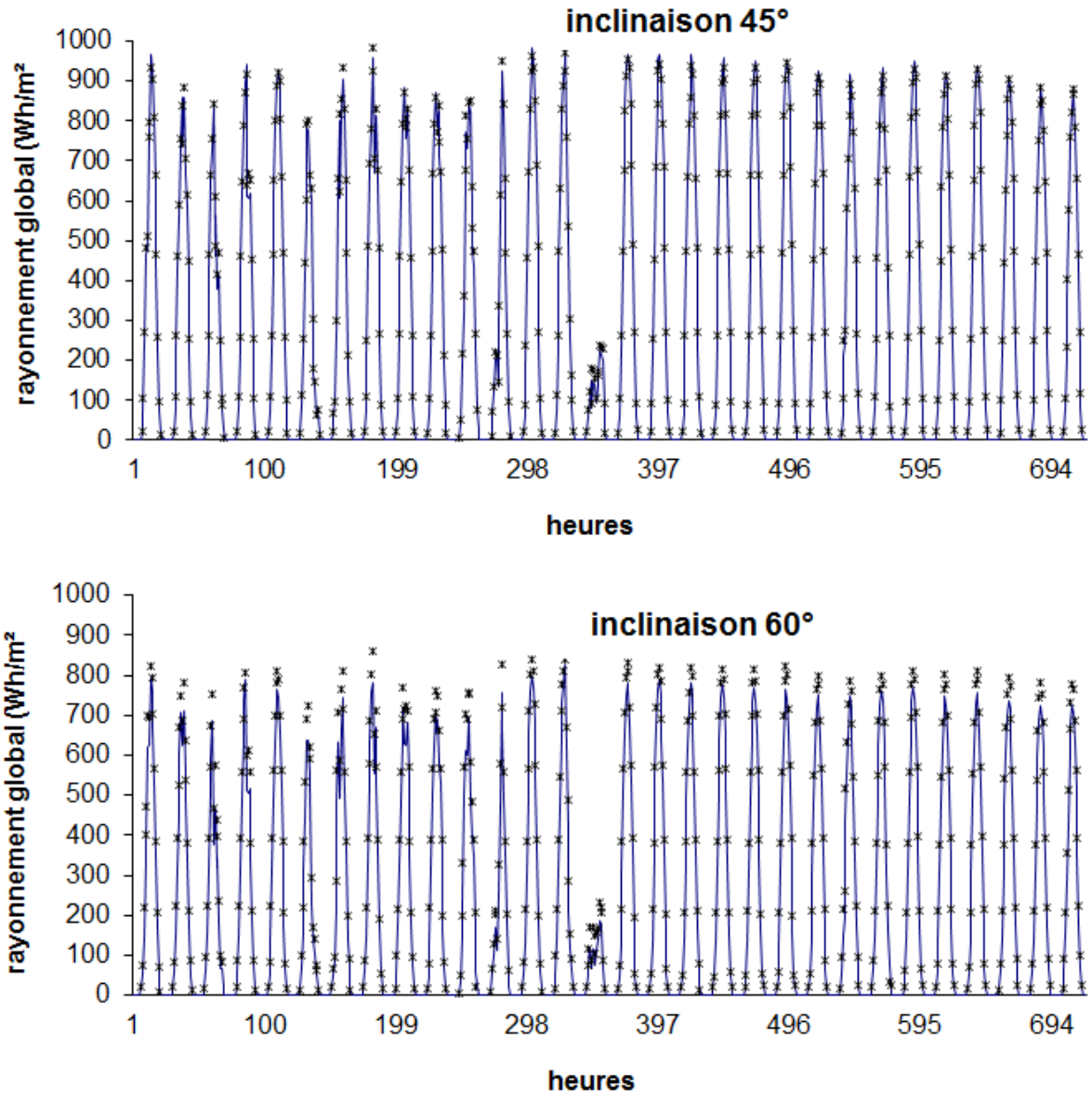


Figure A-2 : comparaison entre la mesure de rayonnement global incliné (trait plein, 45° et 60°) et le calcul de celui-ci avec la méthodologie *CLIMED2*/Klucher (croix noires)

2. ANNEXE B : Méthodologie de nettoyage des séries temporelles ; erreurs aléatoires

Le but de cette annexe est de montrer la méthodologie de nettoyage des séries temporelles que l'on a utilisé durant cette thèse. Nous allons simplement détailler l'étape succédant à l'acquisition. En théorie, on devrait avoir des séries directement exploitables, mais en pratique ce n'est pas le cas : entre 3 et 5% de mesures sont à corriger. C'est certainement la partie la plus laborieuse de l'étude des séries temporelles de rayonnement global. Il est fréquent que le système d'enregistrement soit momentanément interrompu, dérive (nécessitant une calibration), subisse un « bugg » informatique, qu'un composant de la chaîne soit remplacé ou, plus généralement, que l'on soit en présence d'une erreur de mesure aléatoire ou systématique. Pour corriger les erreurs de mesures aléatoires, il faut tout d'abord les détecter. Cette tâche peut se faire pour les valeurs grossières (série de *0* par exemple) en vérifiant manuellement chaque série. Cependant, dès que l'on s'intéresse aux séries de plus 50000 occurrences cela devient très vite impossible. La solution que nous avons adoptée repose sur une assistance informatique, qui génère pour chaque journée (ou chaque heure, suivant l'horizon considéré) la valeur moyenne et l'écart type. Pour chaque élément de la périodicité *T,* il suffit de calculer x^{max}et x^{min} (P périodes disponibles) comme le montre l'équation B-1. Dans le cas d'une série horaire où l'on a 10 ans de données, alors T=8760 (24x365) et P=10.

$$\begin{cases} x^{max}(t \in [1;T]) = E[x(t+pT)] + n.\frac{\sigma(t)}{\sqrt{P}} \; et \; p \; varie \; de \; 0 \text{ à } P \\ x^{min}(t \in [1;T]) = E[x(t+pT)] - n.\frac{\sigma(t)}{\sqrt{P}} et \; p \; varie \; de \; 0 \text{ à } P \end{cases} \qquad \textit{Eq B-1}$$

La valeur du paramètre n permet de régler la sensibilité de détection de valeurs aberrantes. Cette méthodologie n'est pas sans rappeler l'expression des intervalles de confiance liés à une variable statistiques (n = 1 intervalle de confiance à 68%, n = 2 intervalle de confiance à 95% et n = 3 intervalle de confiance à 99%). Nous avons utilisé dans notre étude n = 3 pour détecter les points aberrants évidents, puis n = 2 avec une interprétation manuelle pour vérifier s'il s'agit réellement d'un point aberrant. Il peut s'agir d'une succession de mesure réellement basses (ou élevée) sans pour autant être anormale. Pour utiliser cette méthodologie, il convient d'utiliser de gros échantillon, qui comportent au moins 10 périodes, sans quoi les résultats ne sont pas optimums, et l'analyse manuelle (bien que

subjective) devient la seule méthode réellement efficace. Une fois ces points repérés, il faut modifier la valeur de la série. Pour ce faire, nous avons utilisé, l'algorithme représenté sur l'Eq B-2 qui permet de remplacer les points aberrants par la moyenne spécifique à la position intra-périodique considérée.

$$\begin{bmatrix} SI\ x(t) \notin [x^{min}, x^{max}] \\ ALORS\ x(t) = E[x(t+pT)]\ p\ varie\ de\ 0 \text{ à } P \end{bmatrix} \qquad \textit{Eq B-2}$$

3. ANNEXE C : Méthodologie de nettoyage des séries temporelles ; erreurs systématiques

Le nettoyage que l'on a présenté en ANNEXE 2 ne s'applique qu'aux erreurs de mesure de type aléatoire (caractère probabiliste de l'occurrence). Concernant les erreurs de mesures dites systématiques (erreurs qui perdurent), il convient de trouver une autre méthodologie de correction. De ce fait, si pendant la période d'acquisition des mesures, on constate une dérive d'un élément de la chaîne, il est nécessaire d'effectuer une remise en conformité des acquisitions. Cette dernière peut dans certains cas être faite rétrospectivement. Considérons l'exemple concret des séries temporelles de rayonnement global incliné relevées sur le mur du laboratoire de Vignola et présentées au chapitre 1. Durant l'acquisition, les mesures ont montré une dérive des détecteurs, c'est pourquoi il a été indispensable de remettre à l'échelle les acquisitions non conformes. Pour opérer, une campagne de mesures complémentaires utilisant un détecteur de référence de type pyranomètre étalon (un mois d'acquisition par orientation de façade) fut nécessaire. La méthodologie utilisée est relativement simple et consiste à trouver une relation simple entre le rayonnement mesuré avec les deux types de détecteurs. La Figure C-1 montre la courbe d'étalonnage pour la façade orientée plein sud.

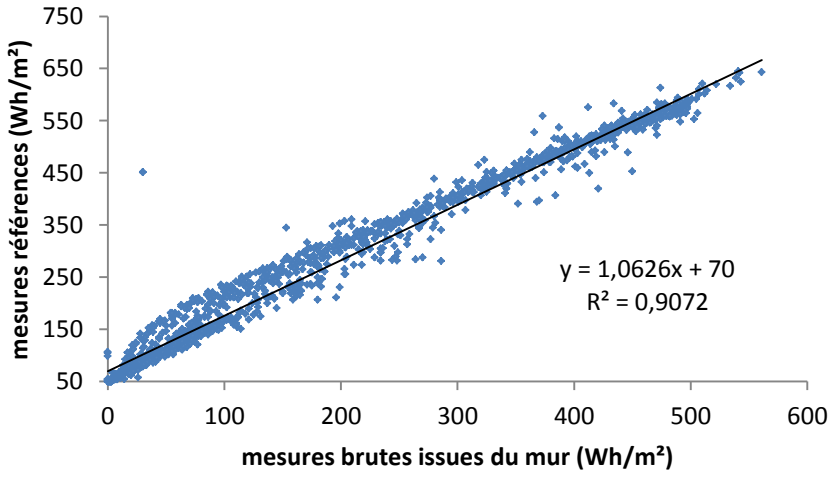


Figure C-1 : courbe d'étalonnage du détecteur de rayonnement global de la façade sud du laboratoire

Pour limiter le problème de bruit de fond qui peut s'avérer important avec les deux détecteurs, les valeurs de rayonnement inférieures à 50 Wh/m² sont considérées comme étant

nulles (offset). La Figure C-2 montre le résultat de la calibration croisée pour la façade Sud qui utilise la courbe d'étalonnage de la Figure C-1.

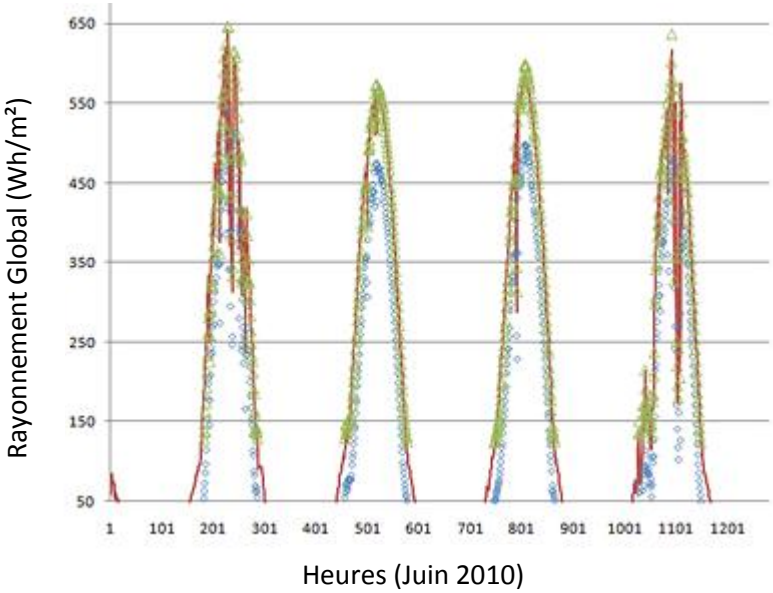


Figure C-2 : résultats de la calibration des détecteurs du mur PV (tilt 80° et azimut 0°). Trait rouge : mesure de référence avec pyranomètre ; rond bleu : mesures brutes issues du mur ; triangle vert : mesures corrigées (étalonnage)

Ce que nous venons de présenter précédemment est préalable à l'utilisation de toutes les séries temporelles pour cette thèse. La méthodologie permet de préparer les séries en vue d'une modélisation.

Nomenclature
(termes en *italique* dans le texte)

β	angle d'inclinaison de la surface par rapport à l'horizontal (deg)
γ	angle azimutal de la surface du plan (angle entre le sud et la projection sur le plan horizontal de la normale à la surface inclinée ; deg)
$\gamma(h)$	fonction d'autocovariance entre les instant *t* et *t+h*
δ	déclinaison solaire (deg)
Δ	erreur de tendance de la prédiction
Δt	temps d'intégration (s)
ε	correction relative de la distance Terre-Soleil
ϵ	erreur entre prédiction et mesure (résidu)
η_{PV}	rendement du système *PV* (%)
η	indice de fiabilité d'une prédiction (%)
θ	angle d'incidence des rayons lumineux sur la surface (deg)
θ_i	paramètres de régression d'un modèle *MA*
λ	longitude (deg)
μ	valeur moyenne
μ^k	paramètre d'amortissement de l'algorithme *LM*
λ_0	longueur d'onde du rayonnement (m)
ρ	albébo de surface (sans unité)
ρ_h	coefficient d'autocorrélation simple (*ACF*) reliant l'instant *t* et *t+h*
ρ_{hh}	coefficient d'autocorrélation partiel (*PACF*) reliant l'instant *t* et *t+h*
σ	écart type
$\tau/\tau_g/\tau_d$	profondeur optique ou coefficient de transmittance général, du rayonnement global et du rayonnement diffusé (sans unité)
φ	latitude (deg)
φ_i	paramètres de régression d'un modèle *AR*
ω_1/ω_2	expression des pulsations liées à la périodicité du rayonnement global (heure^{-1})

$\boldsymbol{\omega_h}$	angle horaire de l'heure $h \in [1,24]$ (deg)
$\boldsymbol{\omega_j}$	angle journalier du le jour $j \in [1,365]$ (deg)
$\boldsymbol{\omega^k}$	vecteurs des valeurs de poids d'un *PMC* au pas d'apprentissage *k*
$\boldsymbol{\omega_{ij}^1}$	poids reliant le neurone d'entrée *j* au neurone caché *i* d'un *PMC*
$\boldsymbol{\omega_i^2}$	poids reliant le neurone caché *i* au neurone de sortie d'un *PMC*
$\boldsymbol{\omega_j^{RL}}$	coefficient de régression linéaire relatif à la *j*-ème variables
a(t)/b(t)	expressions analytiques de l'amplitude du rayonnement global (Wh/m²)
a	azimut solaire (deg)
aod	profondeur optique concernant les aérosols et les radiations de longueur d'onde de 700 nm (sans unité)
ACF	voir ρ_h
b/g/d	constantes d'ajustement du modèle Solis relatives au faisceau direct, global et diffusé
$\boldsymbol{b_i^1}$	Poids reliant le neurone d'offset (biais) et le *i*-ième neurone caché
$\boldsymbol{b^2}$	Poids reliant le neurone d'offset (biais) et le neurone de la couche de sortie
$\boldsymbol{b^{RL}}$	constante de régression linéaire Poids
$\boldsymbol{CP_t}$	rapport à la tendance entre le *CSI* et sa moyenne mobile concernant l'instant *t*, aussi appelé coefficient périodique
$\boldsymbol{CSI_t}$	indice de ciel clair à l'instant *t* (sans unité)
$\boldsymbol{CSI^*_t}$	CSI_t corrigé des variations saisonnières
$\boldsymbol{CSI^*_{i,j}}$	*CSI* corrigé des variations saisonnières pour la *i*-ème période de la *j*-ème mesure dans la période. Pour une série horaire, $CSI^*_{1,1000}$ signifie la 1000ième mesure de la première année
d(a,b)	distance entre a et b correspondant à l'écart quadratique de ces deux valeurs
DC/AC	courant direct et alternatif (A)
EC	erreur cumulée sur une période de prédiction
Esc	constante solaire (1367 W.m²)
ET	équation du temps (heure, min, sec)
$\boldsymbol{E_t}$	composante extra-annuelle de la série temporelle
$\boldsymbol{f}$	fraction de diffusé (sans unité)
$\boldsymbol{f_n}$	expression analytique de la fonction de régression pour *n* observations

f_u	fuseau horaire correspondant au décalage longitudinal du méridien origine (heure)
F	facteur d'inclinaison du modèle de Klutcher (sans unité)
F_c	paramètre calculé du Test de Fisher
F_{seuil}	paramètre discriminant du Test de Fisher
g	fonction d'activation de la couche cachée du *PMC*
h	hauteur solaire (deg) et aussi horizon de prédiction
H_0/H_0'	rayonnement hors atmosphère et rayonnement hors atmosphère corrigé (Wh/m²)
H_{bn}	rayonnement direct monochromatique (Wh/m²)
$H_{gh}/H_{dh}/H_{bh}$	rayonnement global horizontal, diffusé et direct (Wh/m²)
$H_{g\beta}$	rayonnement global incliné d'un angle β
I/I_0	ensemble des paramètres d'un *PMC* à optimiser / valeurs particulières de ces paramètres permettant d'obtenir les meilleurs résultats
I_{cc}	courent de court-circuit (A)
I_{mod}/V_{mod}	caractéristique tension courant d'un module *PV* (A/V)
IC	intervalle de confiance encadrant une prédiction ou une estimation
$J(\omega)$	matrice Jacobienne relatives au poids ω
k_t	indice de clarté à l'instant *t* (sans unité)
k^*_t	k_t corrigé des variations saisonnières
k_0	classe de mesures qui maximise la similarité entre valeurs du passé et du présent
k_t	indice de clarté discret au temps *t* (sans unité)
L	opérateur retard (lag ou backshift operator en Anglais)
$m_{i,j}$	modélisation analytique du $CSI^*_{i,j}$ à partir des deux composantes saisonnières $a_{i,j}$ (périodicité journalière) et $b_{i,j}$ (périodicté annuelle)
M	facteur optique de la masse de l'air (sans unité)
Me	nombre de neurones exogènes positionnés sur la précouche d'entrée du *PMC*
Me^*	groupe de neurones exogènes positionnés sur la couche d'entrée du *PMC*
MM_t	moyenne mobile de *CSI(t)*
N	nébulosité totale (octas)
N	nébulosité (octas)

$\boldsymbol{Nc}$	nombre de neurones positionnés à sur la couche cachée du *PMC*
$\boldsymbol{Ne}$	nombre de neurones endogènes positionnés sur la précouche d'entrée du *PMC*
$\boldsymbol{Ne^*}$	groupe de neurones endogènes positionnés sur la couche d'entrée du *PMC*
$\boldsymbol{N_T}$	nombre de mesures disponibles pour l'optimisation par *MCO* d'une régression linéaire multiple
$\boldsymbol{P/P_0}$	pression atmosphérique ambiante et de référence au niveau de la mer (Pa)
$\boldsymbol{PACF}$	voir ρ_{hh}
$\boldsymbol{PKW}$	vitesse de rafale (m/s)
$\boldsymbol{PR}$	« performance ratio », coefficient qui permet de prendre en compte l'effet de l'onduleur sur la puissance *PV* (sans unité)
$\boldsymbol{P_{max}/P_c}$	puissance maximale/puissance crête (W)
$\boldsymbol{P_{PV}}$	puissance *PV* (W)
$\boldsymbol{P_{ac/dc}}$	puissance *PV* pré-onduleur (*DC*) ou post onduleur (*AC*)
$\boldsymbol{r}$	distance Terre Soleil (m)
$\boldsymbol{r_0}$	distance moyenne Terre-Soleil (150.10^6 m)
$\boldsymbol{r_*}$	incrément temporelle qui minimise la distance entre les mesures du présent et du passé
$\boldsymbol{R}$	coefficient de corrélation de deux grandeurs
$\boldsymbol{R^2}$	coefficient de détermination
$\boldsymbol{R_M}$	matrice de transition de la chaîne de Markov
$\boldsymbol{RH}$	humidité relative (%)
$\boldsymbol{RP}$	hauteur de précipitation (mm)
$\boldsymbol{R_b/R_d/R}$	facteurs correctifs d'inclinaison du rayonnement horizontal relatifs à la composante directe, diffuse et globale (sans unité pour R_b et R_d et Wh/m² pour R)
$\boldsymbol{R_t}$	composante résiduelle de la série temporelle
$\boldsymbol{s}$	périodicité saisonnière d'une série temporelle
$\boldsymbol{S}$	surface des modules *PV* (m²)
$\boldsymbol{S'}$	matrice utilisée pour l'optimisation *MCO* d'une régression linéaire multiple
$\boldsymbol{\Sigma_t}$	composante saisonnière de la série temporelle
$\boldsymbol{S_t}$	indice de ciel clair (sans unité)
$\boldsymbol{ST}$	série temporelle ou chronologique

$\boldsymbol{Su}$	insolation (h)
$\boldsymbol{t}$	incrément de temps ou valeur du *t* de Student
$\boldsymbol{T}$	température ambiante (°C)
$\boldsymbol{T_1/T_2}$	périodes observables au sein de la chronique de rayonnement global
$\boldsymbol{T_{E\_H}}$	transition heure d'été et heure d'hiver (heure)
$\boldsymbol{V_{co}}$	tension de circuit ouvert (V)
$\boldsymbol{w}$	colonne de vapeur d'eau (cm)
$\boldsymbol{W_d}$	direction du vent (deg)
$\boldsymbol{W_s}$	vitesse du vent moyenne à 10 m (m/s)
$\boldsymbol{W}^{\mathrm{RL}}$	matrice regroupant tous les ω_j^{LR}
$\boldsymbol{\hat{x}_t}$	prédiction de la mesure de la série temporelle *x* à l'instant *t*
$\boldsymbol{x_t}$	mesure de la *t*-ième mesure de la série temporelle *x*
$\boldsymbol{< x >}$	valeur moyenne de la série temporelle *x*
$\boldsymbol{X_t}$	mesure de la *t*-ième mesure de *la* série temporelle de rayonnement global (Wh/m²)
$\boldsymbol{\hat{X}_t}$	prédiction de la *t*-ième mesure de *la* série temporelle de rayonnement global (Wh/m²)
$\boldsymbol{V_p}$ / $\boldsymbol{V_R}$	variance particulière calculée sur la périodicité de la série et sur le résidu
$\boldsymbol{VC_x}$	coefficient de variation de la grandeur *x* (%)

Glossaire des acronymes
(termes en *italique* dans le texte)

a-Si/mono-Si Silicium amorphe et monocristallin

ADEC agence de développement économique de la Corse

ADEME agence de l'environnement et de la maîtrise de l'énergie

ALADIN aire limitée adaptation dynamique développement international

ANFIS neuro-fuzzy inference scheme, schéma neuro-flou

ARPEGE action de recherche petite échelle grande échelle

BIC bayesian information criterion, critère d'information bayesienne

CCNUCC convention cadre des nations unies sur les changements climatiques

CIUS conseil international des unions scientifiques

CLER comité de liaison des énergies renouvelables

CLIMED2 climatic changes in the western Mediterranean

CRE commission de régulation de l'énergie

CSI clear sky index, indice de ciel clair

DC/AC courant direct et alternatif

EC erreur cumulée

EDF électricité de France

ER énergies renouvelables

FFER fédération française des énergies renouvelables

GALSI gazoduc algérie-sardaigne-italie

GES gaz à effets de serre

GIEC groupe d'experts intergouvernemental sur l'évolution du climat

IA intelligence artificielle

IC intervalle de confiance encadrant une prédiction ou une estimation

k-NN k nearest neighbors , k plus proches voisins

LM algorithme de Levenberg-Marquardt utilisé pour l'apprentissage d'un *PMC*

MAE/MAPE mean absolute (pourcentage) error, erreur absolue moyenne et sa forme normalisée

MBE mean bias error, biais moyen entre prédiction et mesure

MCO méthode d'optimisation des moindres carrées ordinaires

MSE/RMSE erreur quadratique moyenne et sa racine carrée

MYRTE Mission hYdrogène-Renouvelable pour l'inTégration au réseau Electrique

nRMSE racine carré de l'erreur quadratique moyenne normalisée

NH non hybride

NWP numerical weather prediction, prediction météorologique numérique

OMM organisation météorologique mondiale

ONU organisation des nations unies

ORIENTE optimization of renewable intermittent energy with hydrogen for autonomous electrification

PAGLIA ORBA plateforme avancée de gestion électrique en milieu insulaire associant stockage et énergie renouvelable

PCET plan climat-énergie territorial

PNUE programme des nations unies pour l'environnement

PMC perceptron multicouche

PR performance ratio, rapport de performance

PV relatif au photovoltaïque

RBF réseau de fonctions radiales

RNA réseau de neurones artificiels

SACOI liaison de courant Sardaigne-Corse-Italie

SARCO liaison de courant Sardaigne-Corse

SARIMA modèles autorégressive à moyenne mobile saisonniers

SE/S/SO orientation sud-est, sud et sud-ouest

SOLIS solar irradiance scheme, schéma d'irradiation solaire

SRCAE schéma régional du climat, de l'air et de l'énergie

SPE science pour l'environnement

TL temps local (heure, min, sec)

TSM temps solaire moyen (heure, min, sec)

TSV temps solaire vrai (heure, min, sec)

UTC/GMT temps solaire moyen de la ville de Greenwich

Table des illustrations

Liste des Figures

Liste des Tableaux