Combinaison d'un modèle régional climatique et de modèles à réseaux de neurones pour la prévision du productible photovoltaïque : application au cas de l'île de La Réunion

Yannick Fanchette


Université de La Réunion
École doctorale "Sciences, Technologie et Santé" E.D. 542

THESE DE DOCTORAT

PHYSIQUE ENERGETIQUE

Présentée en vue de l'obtention du grade de :

DOCTEUR EN ÉNERGETIQUE, GENIE DES PROCEDES

COMBINAISON D'UN MODELE REGIONAL CLIMATIQUE ET DE MODELES A RESEAUX DE NEURONES POUR LA PREVISION DU PRODUCTIBLE PHOTOVOLTAÏQUE : APPLICATION AU CAS DE L'ILE DE LA REUNION

Présentée par :

YANNICK FANCHETTE

Soutenue le 29 juin 2023

Composition du jury :

Pr. Ahmed BOURIDANE	Université de Sharjah	**Rapporteur**
Pr. Krishna BUSAWON	Université des Mascareignes	**Rapporteur**
Dr. HDR Daniela CHRENKO	Université de Bourgogne Franche-Comté	**Examinatrice**
Pr. Michel BENNE	Université de La Réunion	**Co-directeur de thèse**
Dr. HDR Béatrice MOREL	Université de La Réunion	**Co-directrice de thèse**
Dr. HDR Harry RAMENAH	Université de Lorraine	**Encadrant de thèse**

Contact
Yannick FANCHETTE: yannick.fanchette@univ-reunion.fr
Pr. Michel BENNE : michel.benne@univ-reunion.fr
Dr. HDR Béatrice MOREL: beatrice.morel@univ-reunion.fr
Dr. HDR Harry RAMENAH: harry.ramenah@univ-lorraine.fr

Adresse
ENERGY^{LAB}
15 Av. René Cassin, CS 92003
97744 Saint Denis Cedex 9

École doctorale "Sciences, Technologies et Santé" - ED 541
15 Av. René Cassin, CS 92003
97744 Saint Denis Cedex 9

« J'ai inventé une lampe de poche qui fonctionne à l'énergie solaire, elle n'a qu'un dernier défaut, elle ne marche qu'en plein soleil. »

Gaston Lagaffe – André Franquin (1924-1997)

« Créer est le seul domaine où il faut se déposséder pour s'enrichir. »

Malcolm de Chazal (1901-1981)

À Aurélie, Timothée et Kupo

Remerciements

Ce projet a bénéficié du soutien financier de la Région Réunion sous le numéro de convention Direction de l'Éducation DIRED/263136, et de la Commission Européenne sous le numéro de convention Programme Opérationnel Fonds Européen de Développement Régional PO FEDER/2014-2020 - Ile de La Réunion - Fiche Action 1.06 "Améliorer les compétences au service de l'économie de la connaissance - Allocations Régionales de Recherche".

De prime abord, la rédaction des remerciements semble être la partie la plus simple de ce manuscrit, mais il demande finalement un effort de concentration et de mémoire afin de ne pas oublier les personnes, qui de près ou de loin, ont apporté leur pierre à cet édifice. Avant de plonger dans les méandres académiques de cette thèse, permettez-moi donc de prendre un moment pour remercier chaleureusement les individus qui ont rendu cette aventure enrichissante et mémorable à bien des niveaux.

En premier lieu, je tiens à exprimer ma reconnaissance envers le Professeur **Michel BENNE** et la Docteur **Béatrice MOREL**, mes deux directeurs de thèse et actuellement respectivement directeur et directeur adjoint du laboratoire ENERGY-Lab. Michel, le *guide sage*, tes conseils avisés ont été ma boussole. Tes encouragements m'ont poussé à dépasser mes propres limites, comme un algorithme d'optimisation qui refuse de converger tant que le résultat n'est pas optimal. Béatrice, la *gardienne inflexible du droit chemin*. Tes remarques perspicaces et tes rappels à l'ordre ont été comme des repères lumineux sur une autoroute sombre et sinueuse. Ton dévouement à mon succès académique m'a ramené sur le droit chemin à chaque fois que je me suis égaré. Et pour compléter ces remerciements envers mon encadrement : le Docteur **Harry RAMENAH**, mon encadrant de thèse de l'université de Lorraine et fier compatriote mauricien. Harry, l'*éclair incandescent*. De mon point de vue, seule la lumière va plus vite que ta pensée et tes innombrables idées (et encore…). Ton esprit vif et créatif m'a inspiré à repousser les limites de ma réflexion et à explorer des horizons inexplorés. Interagir avec toi a été comme tenter de capturer la foudre dans une bouteille, et chaque échange a été une aventure intellectuelle en soi.
Vous trois avez permis à ces travaux de recherches de garder le cap et d'arriver à bon port contre vents et marées.

Je tiens aussi à remercier le Professeur **Jean-Pierre CHABRIAT** qui m'a poussé à tenter cette aventure au départ et le Professeur **Miloud BESSAFI**, qui m'a assidûment accompagné pendant une grande partie de ces travaux.

Merci également à mes deux rapporteurs : les Professeurs **Ahmed BOURIDANE** et **Krishna BUSAWON**. Je tiens à vous remercier pour vos retours sur le manuscrit et nos échanges lors de ma soutenance. Je remercie également mon examinatrice la Docteur **Daniela CHRENKO** pour vos questions pertinentes lors de cette soutenance. Merci encore pour ce moment d'échanges et de partage.

Il est temps de remercier les personnes qui m'ont soutenu et encouragé pendant ces années de travaux au sein du laboratoire ENERGY-Lab. Il y a tant de personnes que j'aimerais citer ici et j'espère n'oublier personne. D'abord pour leur soutien indéfectible autant moralement qu'intellectuellement pendant ces dures années, des collègues devenus amis : **Étienne DIJOUX**, **Christophe LIN-KWONG-CHON**, **Meziane AIT ZIANE**, **Farid AUBRAS**, **Pauline MIAHLE** et **Fabrice K/BIDI**. Je tiens à saluer et remercier du fond du cœur également tous les autres : **Maël RIOU**, **Edouard ROCHEFEUILLE**, **Sébastien BOULEVARD**, **Chao TANG**, **Tifenn JEGADO**, **Ludovic ODDOZ, Alexandre GRAILLET, Mathieu DELSAUT, Kelly GRONDIN, Yannis HOARAU, Christian BROUAT, Patrick JEANTY, P-O, Dominique, Loïc** et bien d'autres. Petite pensée aux autres doctorants que j'ai croisés plus furtivement mais que je remercie aussi, **Tristan**, **Olivia**, **Julie** et tous ceux qui arrivent derrière. Courage !

En ce qui concerne ma famille et mes amis, je ne serai jamais arrivé au bout sans leur soutien quotidien !

À ma compagne exceptionnelle, **Aurélie** : ton amour, ta patience, ton soutien constant et ta bienveillance ont été ma source d'inspiration. Tu as été le pilier, le roc sur lequel je me suis appuyé, et tes mots d'encouragement ont illuminé même les jours les plus sombres de cette aventure, et cela, malgré la maladie et la fatigue ! MERCI ! Cette thèse porte en elle l'empreinte de ton amour et de tes sacrifices !

À mon fils, **Timothée** : *ti bonhomme*, âgé de seulement trois ans mais déjà porteur d'une curiosité infinie, je te dédie tout mon amour et mes remerciements. Quand je me noyais dans un océan d'informations et de codes, ton énergie et ton rire insouciant étaient une bouffée d'oxygène. Puisses-tu grandir dans un monde où la poursuite du savoir est une quête d'épanouissement perpétuel.

À **Kupo**, comme un fils, un compagnon de vie et de route qui m'a accompagné pendant presque toute la durée de cette thèse. Tu es resté à mes côtés pendant des nuits entières. Quand je n'y arrivais plus, tu me donnais le courage qu'il me manquait. Puissent mes mots arriver à tes oreilles même si tu n'es plus près de nous.

À **Florent (Viau !)** et **Cyril** : ça y est les amis, je suis parvenu au bout du tunnel et je vous remercie pour tout.

À ma famille : **Papa**, **Maman**, **Stéphanie** et **Annabelle**, vous avez toujours cru en moi, même quand je n'y croyais plus et je vous en serai éternellement reconnaissant. Comme vous le savez, cette consécration est également la vôtre.

Aux autres membres de ma famille, j'espère que la lecture de ce manuscrit vous éclaircira sur ce à quoi j'ai pu m'atteler ces dernières années.

Enfin à **Jay**, puisses-tu de là-haut, toi aussi, entendre mes paroles et mes remerciements. J'espère que ce travail te rendra fier.

Je tiens aussi à remercier tous les collègues professeurs que j'ai pu rencontrer lorsque j'enseignais en Physique/Chimie ou Mathématiques et Sciences Physiques au lycée pendant les deux dernières années de thèse.

Il a été très dur de terminer cette thèse en étant salarié et avec une famille qui m'attendait et comptait sur moi à la maison mais cela renforce, sans nul doute, ma satisfaction d'être arrivé au bout de cette épopée et il est temps de tourner la **P**age…

Résumé

L'évolution du contexte énergétique mondial et la recherche de solutions d'atténuation et d'adaptation pour compenser les effets des changements globaux stimulent le recours aux ressources renouvelables. Ces sources d'énergies sont caractérisées par une forte variabilité due à leur dépendance aux conditions météorologiques et climatiques. Pour atteindre les objectifs français, et plus particulièrement réunionnais, de réduction des émissions de gaz à effet de serre, ainsi que d'amélioration de l'efficacité énergétique et l'augmentation de la part des énergies renouvelables dans le mix énergétique, le perfectionnement de la connaissance de cette variabilité représente un enjeu incontournable. Dans le cas de la conversion photovoltaïque (PV) à l'échelle de La Réunion, la maîtrise de la variabilité de la production passe par le développement et la mise en place de méthodes permettant de prévoir la production sur un horizon de planification et sur l'ensemble du territoire. Ces prévisions concourent à améliorer le niveau de pénétration de l'électricité PV grâce à une meilleure gestion des centrales PV, installées ou futures, permettant d'optimiser l'intégration de cette ressource dans le mix énergétique.

L'objectif de ces travaux de recherche est de contribuer à améliorer la prévision à court et moyen terme de la production PV. L'étude est basée sur une analyse statistique et un modèle de prévision de la production PV, faisant intervenir des paramètres météorologiques. Étant donné la topographie complexe et marquée de l'île de La Réunion, un nombre important de sites, accueillant des capteurs météorologiques, disposés judicieusement sur l'ile serait indispensable pour mener à bien ces recherches. Toutefois, le coût nécessaire à une telle installation étant trop élevé, l'approche adoptée est de générer ces données météorologiques sur toute l'île à une haute résolution, de l'ordre du km², et sur plusieurs années grâce au modèle régional climatique WRF (Weather Research Forecasting). Une stratégie de sélection des variables, utilisant la causalité au sens de Granger, est développée afin de sélectionner les variables météorologiques (prédicteurs) utiles et pertinentes à l'avancement de la prévision de la production PV. Un modèle statistique de prévision, basé sur la méthode de cointégration de Johansen, est proposé utilisant la corrélation entre variables météorologiques explicatives précédemment sélectionnées et production PV. Ce modèle permet d'estimer ainsi la production PV sur tout le territoire réunionnais. Un modèle spatio-temporel utilisant des réseaux de neurones récurrents à mémoire court-terme persistante (LSTM) et LSTM bi-directionnel (Bi-LSTM) est développé afin de produire des prévisions performantes sur toute l'île à trois horizons temporels susceptibles d'intéresser un gestionnaire de réseau : M+1, j+1 et h+1. Les modèles neuronaux sont confrontés au modèle de persistance et au modèle statistique SARIMA.

Les méthodologies développées pourraient offrir à terme une opportunité d'assurer des garanties supplémentaires au gestionnaire du réseau. Si d'avenir des solutions de prévision performantes se généralisaient, cette opportunité pourrait permettre d'ouvrir le marché au-delà du seuil réglementaire de 35% d'énergie renouvelable imposé actuellement.

Mots-clés : *Prévision spatio-temporelle ; Production photovoltaïque ; Modèle régional de climat WRF : Causalité de Granger ; Cointégration de Johansen ; Réseaux de neurones LSTM ; Ile de La Réunion*

Abstract

The changing global energy context and the search for mitigation and adaptation solutions to offset the effects of global change are driving the use of renewable resources. These energy sources are characterised by high variability due to their dependence on weather and climate conditions. In order to achieve the French objectives, and more specifically those of Reunion Island, of reducing greenhouse gas emissions, improving energy efficiency and increasing the share of renewable energy in the energy mix, improving knowledge of this variability is a key challenge. In the case of photovoltaic (PV) conversion on the scale of Reunion Island, the control of production variability requires the development and implementation of methods to forecast production over a planning horizon and over the whole territory. These forecasts help to improve the level of penetration of PV electricity thanks to better management of installed and future PV power plants, thus optimising the integration of this resource into the energy mix.

The objective of this research is to contribute to the improvement of short and medium term forecasting of PV production. The study is based on a statistical analysis and a forecasting model of PV production that uses meteorological parameters. Given the complex and marked topography of Reunion Island, a large number of sites, hosting meteorological sensors, judiciously placed on the island would be essential to carry out this research. However, as the cost of such an installation would be too high, the approach adopted here is to generate meteorological data over the whole island at a high resolution, of the order of one square kilometre, and over several years using the regional climate model WRF (Weather Research Forecasting). A variable selection strategy, using Granger causality, is developed in order to select the meteorological variables (predictors) that are useful and relevant for improving the forecast of PV production. A statistical forecasting model, based on the Johansen cointegration method, is proposed using the correlation between previously selected meteorological predictors and PV production. This model allows to estimate the PV production on the whole territory of Reunion. A spatio-temporal model using recurrent neural networks with persistent short-term memory (LSTM) and bi-directional LSTM (Bi-LSTM) is developed in order to produce efficient forecasts over the whole island at three time horizons likely to be of interest to a network manager: M+1, d+1 and h+1. The neural models are compared with the persistence model and the SARIMA statistical model.

The methodologies developed could eventually offer an opportunity to provide additional guarantees to the network operator. If efficient forecasting solutions become widespread in the future, this opportunity could open up the market beyond the regulatory threshold of 35% of renewable energy currently imposed.

Keywords: *Spatio-temporal forecasting; Photovoltaic production; Regional climate model WRF; Granger causality; Johansen Cointegration; LSTM Neural networks; Reunion Island*

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

ACF	AutoCorrelation Function
ADF	Augmented Dickey-Fuller
AGCM	Atmospheric Global Circulation Model
AIC	Akaike Information Criteria
AM	Air Mass
ANN	Artificial Neural Network
AOGCM	Atmospheric-Oceanic Global Circulation Model
ARIMA	Autoregressive Integrated Moving Average
ARIMAX	Autoregressive Integrated Moving Average with exogenous input
ARMA	Autoregressive Moving Average
ARMAX	Autoregressive moving average model with exogenous input
ARX	Autoregressive model with exogenous input
Bi-LSTM	LSTM Bi-directionnel
BPNN	Back-Propagation Neural Network
CARDS	Coupled Auto-Regressive and Dynamical System
CFS	Climate Forecast System
CM-SAF	Climate Monitoring Satellite Application Facility
DF	Dickey-Fuller
DHI	Diffuse Horizontal Irradiance
DNI	Direct Normal Irradiance
DW	Durbin Watson
ECM	Error Correction Model
ECMWF	European Centre for Medium-Range Weather Forecasts
EG	Engle and Granger
ERA-5	Fifth generation of the European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis
GCM	Global Circulation Model
GDAS	Global Data Assimilation System
GES	Gaz à effet de serre
GFS	Global Forecast System
GHI	Global Horizontal Irradiance
GRU	Gated Recurrent Unit
HQ	Hannan-Quinn
IEA	International Energy Agency
INSEE	Institut national de la statistique et des études économiques
JB	Jarque Bera
LB	Ljung Box
LM	Lagrange Multiplier
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MCEV	modèle de correction d'erreurs vectoriel
MCO	Méthode des moindres carrés ordinaires
ML	Machine Learning
MRC	Modèle régional du climat

NAM	North American Mesoscale
NOAA	National Oceanic and Atmospheric Administration
nRMSE	normalized Root Mean Square Error
OER	Observatoire Energie Réunion
OMM	Organisation météorologique mondiale
OS	Opération Scientifique
PACF	Partial AutoCorrelation Function
PPE	Programme pluriannuel de l'énergie
r	Coefficient de Pearson
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RTE	Réseau de transport d'électricité
SARAH-2.1	Surface Solar Radiation Data record - Heliosat Version 2.1
SARAH-E	Surface Solar Radiation Data record - Heliosat – Meteosat East
SARIMA	Seasonal Autoregressive Integrated Moving Average
SIC	Schwartz Information Criteria
StD	Standard Deviation
SVM	Support Vector Machine
VAR	Vector Auto Regression
VECM	Vector Error Correction Mechanism
WRF	Weather Research and Forecasting
ZNI	Zone non interconnectée

Introduction générale

Contexte environnemental et énergétique mondial

Les enjeux énergétiques deviennent le centre d'attention principal dans un monde où la croissance démographique et économique ne cesse d'augmenter les besoins en énergie et où les émissions de gaz à effet de serre (GES) induites ainsi générées conduisent au réchauffement climatique de la planète. Ce changement climatique est aujourd'hui un fait admis par la communauté scientifique internationale et acté par les gouvernements. C'est une prise de conscience de la grande majorité des décideurs politiques, industriels, acteurs du monde économique et énergétique et même des citoyens de l'effet de la production énergétique sur le phénomène climatique qu'est le réchauffement planétaire. En novembre 2015 a eu lieu la 21ème conférence des Parties de la Convention-cadre des Nations unies sur les Changements climatiques (CCNUCC) ou COP21 qui a recentré les débats sur le secteur de l'Énergie. Le but était de conclure un accord climatique juridiquement contraignant pour tous qui entrerait en vigueur en 2020. Pour cela, il fallait à la fois adopter un cadre sur l'avant-2020 pour accélérer l'action climatique et de limiter les émissions de gaz à effet de serre et contenir la hausse de la température moyenne mondiale en dessous de 2°C voire 1,5°C. Le secteur de l'Énergie est responsable de près de deux tiers de ces émissions de GES. Selon le dernier rapport du Groupe d'experts Intergouvernemental sur l'Évolution du Climat (GIEC) (Shukla et al., 2022), le réchauffement planétaire a déjà dépassé de 1°C les niveaux préindustriels, en raison des émissions passées et actuelles de gaz à effet de serre. Il existe un nombre considérable de preuves indiquant que le changement climatique a de graves conséquences sur les écosystèmes et les populations. L'océan se réchauffe, devient plus acide et moins fécond. La fonte des glaciers et des calottes glaciaires entraîne une élévation du niveau de la mer et les phénomènes côtiers extrêmes sont de plus en plus intenses. Pour préserver l'environnement, sans entraver la demande en énergie grandissante, une transition énergétique est nécessaire. La pandémie de COVID-19, puis l'invasion de l'Ukraine par la Russie en 2022 a déclenché une crise énergétique mondiale. La Russie était de loin le plus gros pays exportateur de combustibles fossiles. Les répercussions géopolitiques de cette guerre ont créé des tensions sur les chaînes d'approvisionnement mondiales, dont celle de l'énergie entraînant une flambée des prix du gaz naturel, du pétrole et du charbon. Après deux révolutions industrielles où le charbon puis le pétrole et le gaz naturel s'imposent comme sources d'énergie primaires, le secteur de l'énergie se transforme avec un basculement d'un système de production basé sur les énergies fossiles issues de ces révolutions industrielles vers un système basé sur des sources d'énergie faiblement carbonées, les énergies renouvelables, que Rifkin (2011) appelle la troisième révolution industrielle.

En 2021, la consommation d'énergie primaire[1] mondiale s'élevait à 595 Exajoules (EJ), soit une augmentation de 31 EJ par rapport à 2020, ce qui représente la plus forte progression de l'histoire et 8 EJ (+1,3%) par rapport à 2019, avec une consommation moyenne par habitant de 76 Gigajoules (GJ). La figure I.1, tirée de l'outil de cartographie énergétique du British Petroleum, représente l'évolution de cette consommation d'énergie primaire. En effet, le système énergétique a fortement rebondi alors que l'économie mondiale se remettait de la pandémie de COVID-19, expliquant la chute de la consommation globale énergétique. Cette augmentation s'explique par le rebond économique et énergétique post-pandémique.

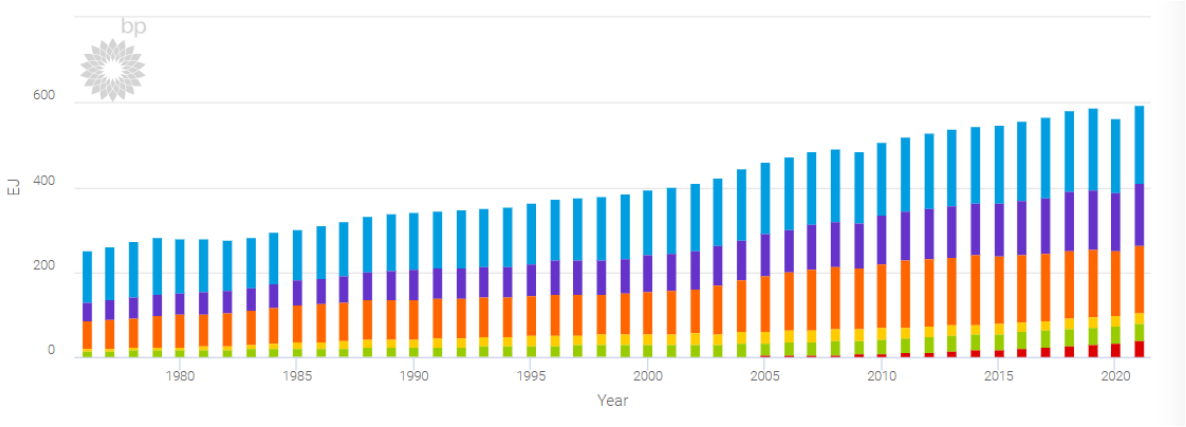


Figure I.1 La consommation d'énergie primaire mondiale entre 1975 et 2021 (en EJ) avec en bleu la consommation de pétrole, en violet de gaz naturel, en orange de charbon, en jaune de nucléaire, en vert d'hydroélectricité et en rouge de renouvelable. Source : (British Petroleum, 2022)

Ainsi, les ressources fossiles, c'est-à-dire le pétrole, le charbon et le gaz naturel représentent 82.2% de cette demande. Bien que la pandémie de COVID-19 ait réduit la demande en 2020 de 4%, la plus grande diminution depuis la Deuxième Guerre mondiale, en 2021 avec la levée progressive des restrictions dans le monde, la demande en énergie mondiale a retrouvé ses valeurs de 2019 notamment pour le pétrole et le gaz qui avaient vu leur consommation diminuer suite à la baisse de la consommation de l'aviation mondiale, particulièrement affectée par la pandémie de COVID-19.

Toutefois, la part d'énergie renouvelable augmente de façon significative. Elle représente, toujours selon le rapport du British Petroleum (2022), 13,5 % de la consommation d'énergie totale mondiale, pour un total de 40 EJ, soit plus du triple d'il y a 10 ans, avec 12 EJ en 2011. Cette tendance ne cesse de se confirmer notamment avec la crise énergétique mondiale provoquée par l'invasion de l'Ukraine en 2022 qui déclenche un engouement sans précédent pour les énergies renouvelables. Cette transition énergétique se manifeste par des investissements accrus dans les sources d'énergies faiblement carbonées, une réduction de la demande énergétique par une amélioration de l'efficacité énergétique des technologies et des bâtiments et par un changement des modes de vie. Il s'agit autant d'une transition énergétique que d'une transition comportementale et sociale (Maresca & Dujin, 2014).

[1] L'énergie primaire est l'énergie théoriquement disponible au sein des ressources naturelles (charbon, gaz, pétrole…) avant toute transformation ou combustion

Il est devenu, ces dernières années, indispensable et inéluctable de diminuer la part des énergies issues de ressources fossiles non renouvelables qui tendent à diminuer dangereusement, et de les remplacer par des énergies vertes dites renouvelables non émettrices de GES. En effet, ces énergies se développent ces dernières années. En 2022, selon l'International Energy Agency (IEA), les biocarburants et le charbon de bois sont la source d'énergie renouvelable la plus utilisée, car indispensable pour la cuisson et le chauffage dans certains pays en voie de développement, mais sont suivis par l'énergie hydraulique et les autres sources d'énergie comme l'éolien, le solaire, la marée.

Concernant la production mondiale électrique en 2021, celle-ci représente 28 466 Térawattheures (TWh) dont 61% à partir des énergies fossiles, 15% à partir de l'hydroélectricité, 6,5% à partir de l'éolien et 3,6% à partir de l'énergie solaire. Parmi les énergies renouvelables, l'hydraulique a encore pesé plus lourd dans les mix énergétique et électrique mondiaux que l'ensemble des autres filières renouvelables. La consommation d'énergies renouvelables hors hydraulique a augmenté de 15% en 2021 par rapport à 2020. Cette année-là, près de 226 GW de capacités éoliennes et solaires ont entre autres été installées dans le monde (proche du record de 236 GW en 2020), mais la part des filières renouvelables hors hydraulique dans le mix électrique mondial n'atteignait encore que 13% en 2021 (contre 15% pour l'hydraulique).

Selon le scénario principal de l'IEA d'analyse et de prévision à l'horizon 2027 (International Energy Agency, 2022), la capacité électrique renouvelable augmentera de 2400 GW d'ici 2027, soit l'équivalent du parc actuel électrique chinois. Les énergies renouvelables deviendront la première source de production d'électricité mondiale au début de 2025, dépassant le charbon. Leur part dans le mix électrique devrait augmenter de 10% au cours de la période de prévision, pour atteindre 38 % en 2027. Les énergies renouvelables sont la seule source de production d'électricité dont la part devrait augmenter, la part du charbon, du gaz naturel, du nucléaire et du pétrole étant en baisse. L'électricité d'origine éolienne et solaire photovoltaïque fera plus que doubler au cours des cinq prochaines années, fournissant près de 20 % de la production mondiale d'électricité en 2027.

La Chine demeure le plus grand marché, avec 40 % de toute la croissance de la capacité renouvelable au cours de la période de prévision, suivie de l'Union européenne, des États-Unis et de l'Inde. L'énergie solaire photovoltaïque, y compris les applications à l'échelle des services publics et les applications distribuées, représente près de 60 % de toute l'expansion de la capacité renouvelable au cours de la période de prévision 2022-2027, suivie par l'énergie éolienne, l'hydroélectricité et la bioénergie.

Selon ce même scénario de l'IEA, la capacité de production d'énergie renouvelable mondiale devrait augmenter massivement entre 2022 et 2027. Ces capacités de production, estimées en 2021 à 3063 GW selon l'International Renewable Energy Agency (IRENA, 2022) devraient augmenter de 350 à 400 GW par an, le solaire photovoltaïque et l'éolien représentant près de 90 % de toutes les nouvelles installations d'énergie renouvelable. L'atteinte du niveau supérieur d'ajouts de capacité dépendra principalement du rythme de mise en service des projets photovoltaïques à grande échelle et distribués en Chine et dans l'Union européenne.

L'éolien en mer (ou offshore) contribue à hauteur de 4 % de l'augmentation de la part des énergies renouvelables (EnR), sa capacité devant tripler d'ici 2024, stimulée par les enchères concurrentielles dans l'Union européenne et les marchés en expansion en Chine et aux États-

Unis. La capacité bioénergétique croît autant que l'éolien en mer, les plus fortes expansions étant enregistrées en Chine, en Inde et dans l'Union Européenne. La croissance de l'hydraulique ralentit, bien qu'elle représente encore un dixième de l'augmentation totale de la capacité renouvelable.

De cette volonté de transition énergétique découle la politique énergétique française

Politique énergétique en France

La politique énergétique adoptée par l'État français regroupe l'ensemble des orientations et décisions prises en matière de production et de consommation d'énergie. La loi relative à l'énergie et au climat adoptée et présentée par le Président de la République en novembre 2019 a créé une loi de programmation sur l'énergie et le climat (LPEC) qui fixe les grands objectifs des Programmations pluriannuelles de l'énergie (PPE) et de la Stratégie nationale bas-carbone (SNBC).Ces trois documents forment ainsi la stratégie française pour l'énergie et le climat et sont les outils de pilotage de la politique énergétique. Ces PPE concernent la métropole continentale mais aussi les ZNI, à savoir La Réunion, la Corse, la Guyane, etc. (Ministère de la Transition écologique, 2023). Elle est basée sur plusieurs fondements, les 17 objectifs de développement durable défendus par les Nations Unies (United Nations, 2015), dont l'accès à l'énergie de tout le territoire français, la sécurité de l'approvisionnement en énergie et la préservation de l'environnement et la lutte contre l'effet de serre.

La production électrique en France

Dans le domaine de l'électricité en France, cette politique s'est démarquée par l'implantation du nucléaire dans le paysage énergétique français qui a conduit à une relative autonomie électrique tout en garantissant des prix parmi les plus bas d'Europe et un bilan carbone plus faible et par une volonté de développer drastiquement les énergies renouvelables. En 2021, la production totale d'électricité en France s'établit à 445,2 TWh , soit une forte baisse de 15% (-77 TWh) par rapport à 2021 (Réseau de Transport d'électricité (RTE), 2023). Cela est imputable à la faible disponibilité du parc nucléaire (-82 TWh par rapport à 2021) et des contraintes sur la production hydraulique (-12TWh). La production totale est répartie entre les différentes filières de production comme suit et est représentée dans la figure I.2 :

- 63% (279 TWh) de nucléaire.
- 11% (49.6 TWh) d'hydraulique
- 9% (38.1 TWh) d'éolien (terrestre et en mer)
- 10% (44,1 TWh) de gaz
- 1% (5.1 TWh) de charbon et fioul
- 4% (18.6 TWh) de solaire
- 2% (10.6) de thermique renouvelable et déchets

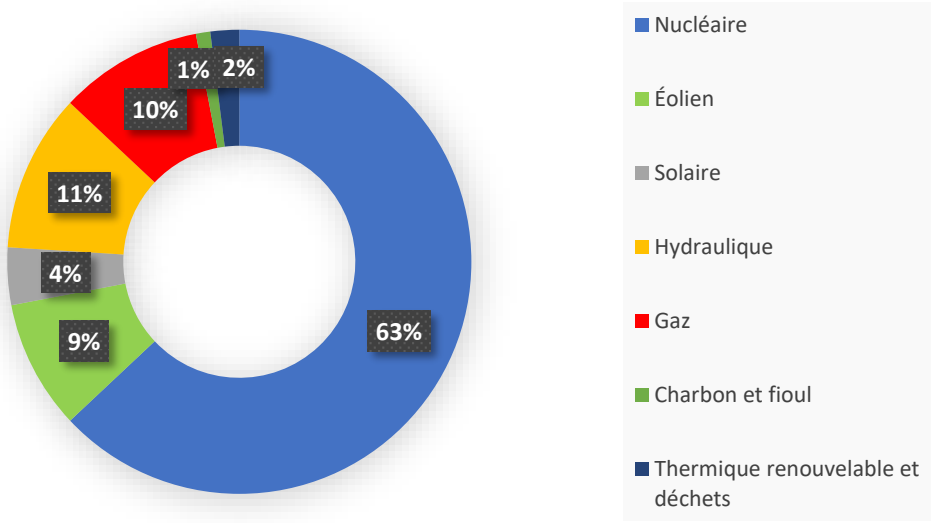


Figure I.2 Répartition par filière de la production totale d'électricité en 2022 en France

Le cas de la France est particulier dans le contexte actuel mondial car plus des 2/3 de sa production sont issues de la filière nucléaire suivie par la filière renouvelable (24%).

Le développement des EnR électriques en France

La France s'est engagée à réduire ses émissions de gaz à effet de serre et à développer ses EnR dans le cadre de la Directive européenne 2009/28/EC (2009) qui a imposé un objectif de 23% d'énergies renouvelables dans la consommation totale d'énergie finale en 2020, puis de la loi de la transition énergétique pour la croissance verte (LTEVC) (Ministère de la Transition écologique, 2015) promulguée en août 2015. Cette dernière, pivot de la transition énergétique dans laquelle s'est inscrite la France, fixe plusieurs objectifs à moyen et long termes, notamment :

- Réduire les émissions de gaz à effet de serre de 40% entre 1990 et 2030 ;
- Réduire la consommation énergétique finale de 50% en 2050 par rapport à 2012 avec un objectif intermédiaire de 20% en 2030 ;
- Réduire la consommation énergétique primaire d'énergies fossiles de 30% en 2030 par rapport à 2012 ;
- Réduire la part du nucléaire dans le mix électrique à 50% d'ici 2025 ;
- Porter la part des énergies renouvelables à 23% de la consommation finale brute d'énergie en 2020 et à 32% de la consommation finale brute d'énergie en 2030.

En 2022, les énergies renouvelables représentent 24% de l'énergie électrique totale, avec selon le rapport de la RTE, une hausse de la production solaire de 30% par rapport à 2021.

Le développement de la filière photovoltaïque (PV)

La France a fait le choix d'une transition énergétique, avec une moindre dépendance du nucléaire et des combustibles fossiles, bâtie sur une croissance maîtrisée des énergies renouvelables. Parmi elles, l'énergie solaire photovoltaïque, au même titre que l'éolien et

l'hydraulique, contribue au développement énergétique propre et durable, car elle n'émet pas de gaz à effet de serre en phase d'exploitation. Cette politique a amené à une très forte croissance du nombre d'installations PV raccordées en France. La figure I.3 montre que la production photovoltaïque annuelle est passée de 105 MWh en 2008 à 18600 MWh en 2022, et a donc été multipliée par un facteur 175 sur ces quinze dernières années. La production solaire sur l'année 2022 a augmenté de 30 % par rapport à 2021. Le rythme de développement du photovoltaïque a atteint un niveau record en 2022, avec près de 2,6 GW nouvellement installés pour un parc total de 15.7 GW en fin d'année. Ce rythme est plus de trois fois plus élevé que celui observé sur les cinq années précédentes (815 MW en moyenne entre 2016 et 2020) et a permis une augmentation de 26% de la puissance installée par rapport à fin 2020. Cependant, l'objectif de 20,1 GW de photovoltaïque fixé par la PPE pour 2023 n'est pas atteint mais 4.4 GW devraient être mis en service d'ici la fin de 2023. Pour atteindre les objectifs fixés par la PPE en 2028 le rythme de développement du parc solaire de 3.2 à 4.7 GW/an devra être conservé (Réseau de Transport d'électricité (RTE), 2023).

L'essor spectaculaire de la filière PV est le fer-de-lance de la transition énergétique amorcée par l'État français des énergies fossiles vers des énergies renouvelables.

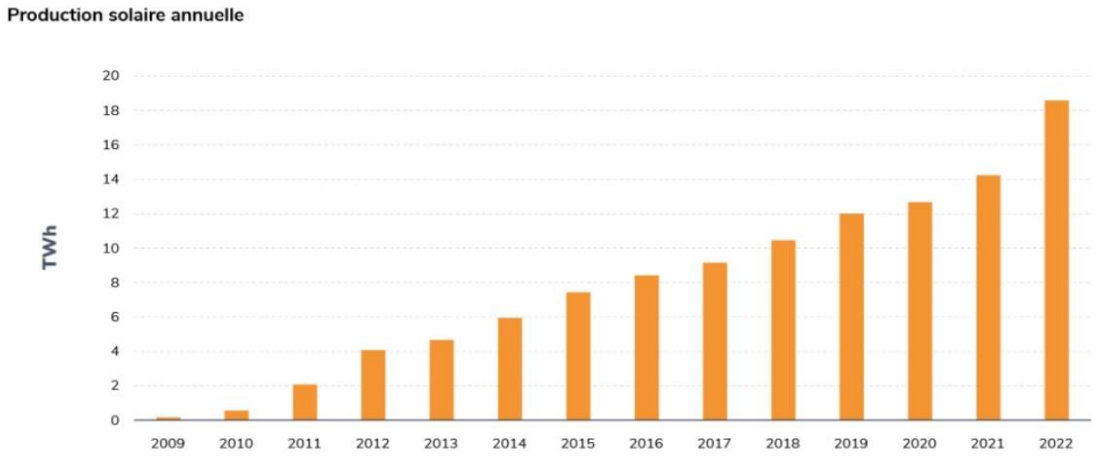


Figure I.3 Évolution de la production solaire annuelle en France

L'insularité et le contexte énergétique de La Réunion

Les îles de par le monde apparaissent comme des territoires contraignants d'un point de vue énergétique. Isolées et dépourvues de ressources fossiles, elles sont dépendantes d'un approvisionnement énergétique extérieur.

C'est notamment le cas des îles du Sud-Ouest de l'Océan Indien (SOOI) regroupées au sein de la Commission de l'Océan Indien (COI), une organisation intergouvernementale créée en 1982 et composée de Maurice, l'île de La Réunion, Madagascar, des Seychelles et des Comores. Ces états qui sont massivement dépendants des combustibles fossiles importent au moins 80% de l'énergie primaire est importée (pétrole et charbon) (COI, 2017). Parmi eux, Madagascar et les Comores importent plus de 90% de leur énergie et les Seychelles, 95% de produits pétroliers pour leur approvisionnement énergétique. L'impact financier sur les compagnies d'électricité et le budget de l'état ainsi que leur dépendance au prix du marché

rendent ces états encore plus vulnérables énergétiquement et économiquement. Un des enjeux des membres de la COI est justement de permettre un meilleur accès à l'énergie. La région possède un grand potentiel en énergies renouvelables et c'est dans cette optique que de nombreuses initiatives ont vu le jour ces dernières années afin d'harmoniser et développer le secteur de l'énergie dans la région. Entre autres, le projet « ENERGIES », programme pour la promotion des énergies renouvelables et l'efficacité énergétique, a été conçu afin d'améliorer l'accès à des sources modernes, durables et renouvelables d'énergie à ces états membres de la COI (Projet ENERGIES-COI, 2019).

Ces territoires ne sont pas connectés au réseau électrique continental et voient ainsi leur approvisionnement en électricité spécifiquement contraint : on les regroupe sous l'appellation de zones non interconnectés (ZNI). En prenant en compte les problèmes d'approvisionnement que peuvent rencontrer ces territoires (l'impossibilité d'une connexion aux sources d'énergie du continent, fluctuations de la demande énergétique en fonction du tourisme saisonnier, difficultés de l'approvisionnement liées au relief et l'isolement, …) (Roche et al., 2018) et l'abondance de ressources renouvelables présentes sur ces mêmes territoires, développer les moyens de production renouvelables sur ces îles devienne une solution évidente et inéluctable.

Ressources énergétiques à l'île de La Réunion

Le programme pluriannuel de l'énergie (PPE) pour l'île de la Réunion, mis en place en 2015, vise à tendre vers une autonomie énergétique par la maîtrise de la demande en énergie et le développement des énergies renouvelables. Cette perspective permettra l'émergence de nouveaux métiers et la création d'emplois. À partir d'un état de la situation en 2013, elle établit les conditions permettant entre 2016 et 2023 (Observatoire Énergie Réunion, 2020) :

- d'augmenter les gains annuels d'efficacité énergétique pour atteindre 360GWh électriques économisés à l'année 2023 ;
- de développer massivement la production d'électricité à partir des énergies renouvelables garanties (+190%) dans le mix énergétique ;
- de poursuivre le développement à partir de sources d'énergies renouvelables intermittentes ;
- et pour cela, de faire, évoluer, le seuil de déconnexion des énergies intermittentes : de 30% en 2014 à 35% en 2018 puis viser une fourchette de 40 à 45% en 2023 ;
- d'initier la transition vers un système de transport propre et efficace.

La Réunion est ainsi poussée vers des objectifs ambitieux : porter la part des énergies renouvelables à 50% au niveau de la production électrique en 2020, puis 100% d'ici 2030. Toutefois ces dernières années, la part des énergies renouvelables bien qu'en progression n'a pas atteint l'objectif fixé de 50% de la production électrique et est très loin d'atteindre cette autonomie énergétique souhaitée. L'évolution du mix énergétique de cette région d'outre-mer est déjà perceptible.

Selon l'Observatoire Énergie Réunion (OER) dans le Bilan Énergétique de la Réunion en 2021, la part des EnR est estimée à 28.2% (contre 30.3% en 2020) comme illustrée dans la figure I.4, tirée du Bilan énergétique de la Réunion 2021 (Observatoire Énergie Réunion, 2022), les énergies fossiles s'établissant à 71,8%. L'hydroélectricité est la solution renouvelable la

plus déployée sur son territoire avec une production électrique représentant 11.6% du mix énergétique. L'ensemble constitué du photovoltaïque, de l'éolien et du biogaz représente 9.4% et est la deuxième part de production locale. Il est suivi par la bagasse qui représente 7% du mix énergétique. La diminution de la part de production régionale d'électricité par les énergies renouvelables (28.2% en 2021 contre 31.2% en 2019), malgré une légère augmentation de la production issue de la bagasse, s'explique par la chute de la production hydroélectrique de plus de 30% en un an due à une année 2019 et 2020 particulièrement sèche. La Réunion ne perd pas pour autant son objectif d'autonomie énergétique. Ainsi, l'entreprise Albioma a décidé en 2021 (Albioma, 2020) de réaménager sa centrale de Bois-Rouge dans l'optique d'abandonner totalement le charbon et l'alimenter ainsi à 100% en biomasse dès le second semestre 2023 permettant à terme de faire passer à plus de 50% la part du renouvelable dans le mix énergétique à la Réunion et de réduire les émissions de gaz à effet de serre d'environ 640 000 tonnes équivalent CO_2 par an soit une basse de 84% des émissions directes par rapport au fonctionnement actuel de la centrale.

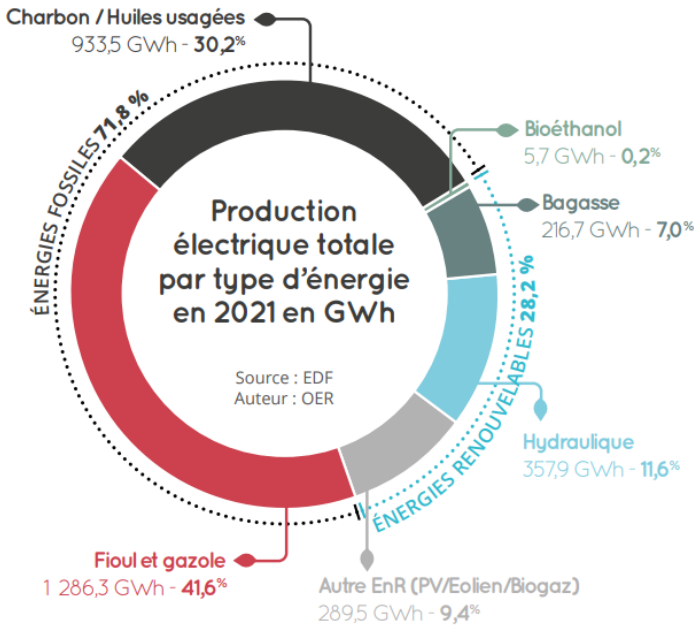


Figure I.4 Bilan énergétique de l'île de la Réunion de 2021. Source :(Observatoire Énergie Réunion, 2022)

De manière générale, l'augmentation constante de la part des EnR dans le mix énergétique est due en grande partie à la forte progression du secteur photovoltaïque. Alors qu'en métropole, 3% de la production totale électrique est issue du PV, il représente sur l'île de La Réunion 8,7% de la production électrique totale. Le parc PV a augmenté de 20% entre 2015 et 2021 avec une production de 267,6 GWh pour 223,6 MW raccordés. La puissance photovoltaïque installée au 31 décembre 2021 (installations raccordées au réseau) est de 249,4 Wc par habitant sur l'île, ce qui classe l'île au 6^{e} rang des régions de France en termes de puissance installée par habitant. Aujourd'hui, malgré une croissance, la filière PV connaît des limites qu'il faut dépasser afin d'en faire un pilier de la transition énergétique.

Verrous scientifiques et technologiques

L'énergie photovoltaïque est une technologie mature qui est de plus en plus compétitive économiquement par rapport aux sources d'énergie traditionnelles, et qui est de plus en plus déployée dans le monde. Le photovoltaïque offre plusieurs avantages, notamment une absence de bruit et de pollution de l'air, un relatif faible coût d'installation et de maintenance, une réduction de la dépendance vis-à-vis des ressources naturelles non renouvelables et une réduction des émissions de gaz à effet de serre. Cependant, il reste encore des verrous scientifiques et technologiques à surmonter pour améliorer l'efficacité et la rentabilité de l'énergie solaire photovoltaïque, notamment :

- **Efficacité de conversion** : L'efficacité de conversion des panneaux solaires dépend de la capacité des cellules photovoltaïques à convertir l'énergie solaire en électricité. Les cellules photovoltaïques actuelles ont une efficacité de conversion moyenne d'environ 15-20% (essentiellement les cellules à base de silicium qui couvre plus de 80% du marché mondial). Certaines cellules de pointe aujourd'hui peuvent atteindre 24% ou plus pour améliorer l'efficacité de conversion (Al-Ezzi & Ansari, 2022). Des recherches sont menées afin de mieux comprendre les propriétés des matériaux utilisés dans les cellules solaires, telles que le silicium ou encore le pérovskite d'halogénure organométallique dont la stabilité et le rendement élevé ont suscité l'attention des chercheurs. La réduction des coûts de production et l'amélioration de l'efficacité ont été obtenues grâce à l'introduction de nouvelles techniques de contrôle du point de puissance maximale. Ainsi, des cellules solaires PV hybrides ont été fabriquées avec de l'oxyde de nanostructure poly 3-hexylthiophène (P3HT) (Almosni et al., 2018).
- **Durabilité** : Les panneaux solaires doivent être résistants aux intempéries, aux conditions climatiques de manière générale, comme c'est le cas dans des zones tropicales comme l'île de La Réunion, à la corrosion et à l'exposition à la lumière du soleil pour une durée de vie de 20 à 25 ans. Des études sont menées sur des matériaux plus durables tels que les cellules solaires organiques ou les cellules solaires pérovskites biface (Song et al., 2022), ainsi que sur des technologies de protection pour augmenter la durée de vie des panneaux solaires.
- **Stockage de l'énergie** : Les panneaux solaires produisent de l'électricité seulement pendant les heures d'ensoleillement. Cette électricité est soit transportée et utilisée ou stockée sur place. Cela nécessite des moyens de stockage d'énergie pour une utilisation ultérieure. (Worku & Abido, 2015) a utilisé un système de stockage d'énergie par supercondensateur (SCESS) dans une stratégie de gestion de la puissance et de contrôle pour un système PV intégré au réseau, afin d'augmenter l'efficacité et la rentabilité de l'énergie PV.
- **Coûts de production** : Les coûts de production des panneaux solaires ont considérablement diminué ces dernières années (Kavlak et al., 2018) avec une réduction de 89% du coût de fabrication sur les dix dernières années (Shafiullah et al., 2022) mais la production à grande échelle reste coûteuse. Réduire les coûts de fabrication rendrait l'énergie photovoltaïque encore plus compétitive. Cela inclut l'utilisation de processus de fabrication plus efficaces, ainsi que la mise au point de technologies de recyclage pour récupérer les matériaux des panneaux solaires en fin de vie.
- **Stabilité du réseau électrique** : La stabilité du réseau, surtout s'il est alimenté par des énergies renouvelables est un sujet important. Un micro-réseau n'a en effet pas l'inertie d'un

grand réseau, et ne peut compter que sur ses propres ressources. Ainsi, l'intermittence et la variabilité de la production PV peuvent poser problème car un équilibre permanent entre production et consommation doit être assuré à tout instant (Widén et al., 2010). Afin de pallier à cette notion d'intermittence et de variabilité de la ressource solaire et donc de l'énergie solaire PV, une solution est la prévision à court, moyen et long terme de cette production (Li et al., 2019).

Positionnement des travaux

Cette thèse se pose comme un trait d'union entre deux des opérations scientifiques au sein du laboratoire ENERGY^{LAB} : le premier, étudiant la ressource solaire, sa variabilité et ainsi sa prévision (Opération Scientifique 1) à l'île de La Réunion et dans la région SOOI et l'autre (Opération Scientifique 2), les micros réseaux et sa stabilité étant donné que l'île est une zone non interconnectée (ZNI).
La prévision de l'énergie photovoltaïque joue un rôle important dans l'intégration de cette source d'énergie variable dans le réseau électrique. Étant donné que la production d'énergie solaire dépend des conditions météorologiques, elle peut fluctuer considérablement au fil du temps. La prévision de l'énergie photovoltaïque permet de prédire avec une certaine précision la quantité d'énergie solaire qui sera produite à un moment et lieu donnés, permettant ainsi aux gestionnaires de réseau de mieux planifier la production d'énergie, l'achat et la vente d'énergie, la régulation de la fréquence et de la tension, et l'optimisation du stockage d'énergie. Le caractère aléatoire et stochastique dans le temps et l'espace de la ressource solaire particulièrement en milieu tropical et insulaire entraînent une limitation de l'insertion de cette énergie renouvelable à 35% de la puissance active totale sur le réseau électrique. La solution envisagée, ici, est la prévision de la production photovoltaïque. Les travaux présentés dans ce manuscrit s'inscrivent dans une volonté de plus en plus évidente aujourd'hui d'intégrer massivement des énergies renouvelables dans un réseau électrique insulaire.

Problématiques

La problématique de la prévision de la production PV est multiple de par la topographie marquée de l'île de La Réunion. Les données au sol à notre disposition s'avèrent insuffisantes en l'état pour obtenir quelque soit la méthode et l'outil employés des résultats performants. Pour résoudre cette problématique, nous faut-il poser des capteurs sur l'ensemble du territoire afin d'obtenir assez de données pour nos modèles de prévision ? Les coûts d'une telle approche et la gestion du parc ainsi déployé seraient beaucoup trop conséquents. Notre approche dans ce mémoire est la génération de données météorologiques à une résolution très fine, de l'ordre du kilomètre carré, par un modèle régional de climat, notamment dans notre cas le modèle WRF, afin de pallier à cette première difficulté. Ensuite, nous devons utiliser un modèle d'estimation de production photovoltaïque à partir des données météorologiques simulées et un modèle de prévision PV qui exploitera cette base de données ainsi générée.

Objectif de la thèse

L'objectif principal de ce travail de recherche est l'élaboration d'un nouveau modèle de prévision spatio-temporelle de production photovoltaïque combinant un modèle de réseau de neurones et un modèle statistique utilisable sur l'ensemble du territoire réunionnais, malgré la topographie marquée de l'île. Pour cela, des données simulées sur toute l'île de La Réunion avec une granularité de l'ordre du km² alimentent ce modèle. Les travaux de recherche s'inscrivent dans le cadre d'une identification des variations spatiales et temporelles de la production d'énergie PV en tenant compte des facteurs tels que la météorologie et la topographie.

Organisation du manuscrit

Pour cela, le mémoire de thèse sera organisé comme suit.

Le chapitre 1, présente le contexte mondial et local énergétique et la place du photovoltaïque dans le mix énergétique diversifié de la Réunion. L'enjeu du photovoltaïque et de sa prévision permet de montrer l'apport de ces travaux dans le contexte énergétique. Un état de l'art des modèles de prévision de la production PV décrits dans la littérature est mis en avant.

Le chapitre 2 porte sur le contexte particulier de l'île de La Réunion et des données qui alimenteront nos modèles de prévision de production PV choisis. Une présentation des données disponibles ainsi que du modèle régional de climat WRF sera faite. Ce modèle WRF est utilisé comme générateur de données météorologiques à une échelle très fine sur toute l'île de la Réunion afin de respecter les variations locales de climat.

Le chapitre 3 concerne les choix des méthodes et des outils retenus et développés. Ce chapitre permet d'évaluer les limites des performances des deux méthodes retenues vis-à-vis des données à notre disposition et leur complémentarité. Le modèle statistique de Johansen, modèle basé sur l'analyse des séries temporelles, ainsi que le modèle utilisant des réseaux de neurones récurrents à mémoire court-terme persistante (LSTM) sont plus longuement abordés.

L'objectif du chapitre 4 est de décrire en détail la méthode statistique de Johansen qui, utilisée fortement en économétrie, n'a jamais été utilisée sur de la prévision PV. Ce chapitre traite des simulations de la prévision utilisant le modèle statistique de Johansen ainsi que les résultats de la prévision et de ses erreurs.

Le chapitre 5, quant à lui, donne une description détaillée de la procédure du modèle de prévision spatio-temporelle appliquée à nos données développées dans le chapitre 3. L'évaluation de sa performance prédictive sera étudiée selon un horizon temporel mensuel, journalier et horaire.

CHAPITRE 1

VERS LA PREVISION PHOTOVOLTAÏQUE

Introduction

1.1 Énergie solaire photovoltaïque

L'énergie solaire occupe une place de choix parmi les énergies renouvelables. Elle est en effet disponible partout à la surface de la Terre et en très grande abondance. Par ailleurs, sa conversion directe en énergie électrique ou thermique offre un rendement plus important que l'énergie éolienne, hydraulique ou maritime.

1.1.1 Le potentiel solaire

L'énergie solaire est fiable, durable et renouvelable La Terre continue de recevoir environ 170 000 TW d'énergie solaire, soit environ 10000 fois le total des besoins en énergie primaire de l'humanité (Lakatos et al., 2011).

Le rayonnement extraterrestre

Le Soleil est à l'origine de 99,98% de l'approvisionnement énergétique mondial (Labouret & Villoz, 2006), y compris la conversion thermique, photovoltaïque, photochimique, photobiologique et hybride de l'énergie solaire, hydraulique, éolienne, houlomotrice et de la biomasse. Cette dernière étant à l'origine des ressources fossiles que nous exploitons. Parmi les autres sources d'énergie, il existe l'énergie marémotrice, l'énergie géothermique et l'énergie nucléaire. L'énergie du soleil provient de réactions de fusion dans son noyau. Ces réactions ont lieu depuis 4,5 milliards d'années et devraient se poursuivre pendant encore 6,5 milliards d'années jusqu'à la mort de l'étoile. La puissance totale rayonnée dans l'espace par le soleil est d'environ $3{,}86 \times 10^{26}$W. Étant donné que le soleil se trouve à environ $1{,}5 \times 10^{11}$m de la terre et que celle-ci a un rayon d'environ $6{,}3 \times 10^{6}$m, il n'intercepte que 0,000000045% de cette puissance (Kennewell & McDonald, 2015). Cela représente tout de même une puissance de $1{,}75 \times 10^{17}$W. La majeure partie de ce rayonnement se situe dans la partie visible et infrarouge du spectre électromagnétique, moins de 1% étant émis dans les bandes spectrales radio, UV et rayons X. Le Soleil est une source d'énergie pour l'humanité. Composé à 90% d'hydrogène, à 10% d'hélium et des traces de carbone, d'oxygène et autres éléments lourds, le Soleil a un profil de température qui passe de 15 100 000K en son noyau à 6 000K à sa surface (Amari et al., 2015). Une cascade de réactions de fusion thermonucléaire entre les atomes d'hydrogène ou entre les atomes d'hydrogène et d'hélium a lieu sans cesse au cœur du Soleil à des millions de degrés de température. La perte de matière pendant cette fusion libère une colossale quantité d'énergie sous forme d'ondes dites électromagnétiques couvrant un domaine de longueur d'onde allant de 0,2 à $4\mu m$. Ces longueurs d'onde forment le spectre solaire et transportent l'énergie du Soleil à la Terre (Ramalingam & Indulkar, 2017). L'irradiance solaire totale arrivant sur l'atmosphère terrestre est de 1361 $W.m^{-2}$, valeur moyenne plus ou moins constante mesurée sur les dernières décennies (Schmutz, 2021).

Le rayonnement à la surface de la Terre

Le rayonnement solaire extraterrestre incident n'arrive pas en totalité à la surface de la Terre. Il interagit avec différents éléments atmosphériques tels que les gaz, les aérosols, les gouttelettes nuageuses ou autres molécules. Ces éléments absorbent et diffusent le rayonnement à travers l'atmosphère. Il est ainsi atténué lorsqu'il passe à travers l'atmosphère, avec des pics d'absorption particuliers correspondant aux gaz atmosphériques dominants (O_3, H_2O, CH_4 et CO_2 notamment) ou autres particules présentes comme les aérosols. Ces derniers ont un impact non négligeable (Riihelä et al., 2018). La trajectoire des rayons, en fonction de la position du Soleil, ne sera pas la même dans l'atmosphère. La longueur du trajet à travers l'atmosphère dépend du lieu et du moment de la journée et de l'année. Par convention, cette modification de la longueur du trajet est représentée par la masse d'air, (ou Air Mass, AM), qui est le rapport entre l'épaisseur atmosphérique traversée par le rayonnement solaire direct pour atteindre le sol et l'épaisseur d'atmosphère traversée à la verticale du lieu.
Dans la figure 1.1, le spectre solaire représenté au sommet de l'atmosphère est appelé AM0 tandis qu'AM1 représente le chemin le plus court à travers l'atmosphère, c'est-à-dire normal à la surface de la Terre, et AMx désigne x fois ce chemin le plus court. Formellement, cette distance se calcule en fonction de l'angle d'incidence θ par rapport au zénith. La masse d'air vaut :

$$AM = \frac{1}{\cos\theta} \tag{1}$$

À la surface de la Terre et dans des conditions de ciel-clair pour une masse d'aire de 1,5, soit une traversée de l'atmosphère terrestre d'une fois et demie, AM1.5 est utilisé. Cela permet les tests de cellules et de modules. Par convention également, la valeur standard européenne de la masse d'air utilisé pour les tests de cellules et de modules PV est AM1.5. En prenant en compte le rayonnement direct à incidence normale reçu directement du disque solaire, le spectre AM1.5D (où D signifie direct) et le rayonnement global regroupant le rayonnement direct et diffus, le spectre AM1.5G (où G signifie global) est utilisé alors pour un soleil à 37°. Dans la figure 1.1, les spectres de référence avec masse d'air (AM1.5 global et direct) et extraterrestre terrestre (AM0) sont comparés.

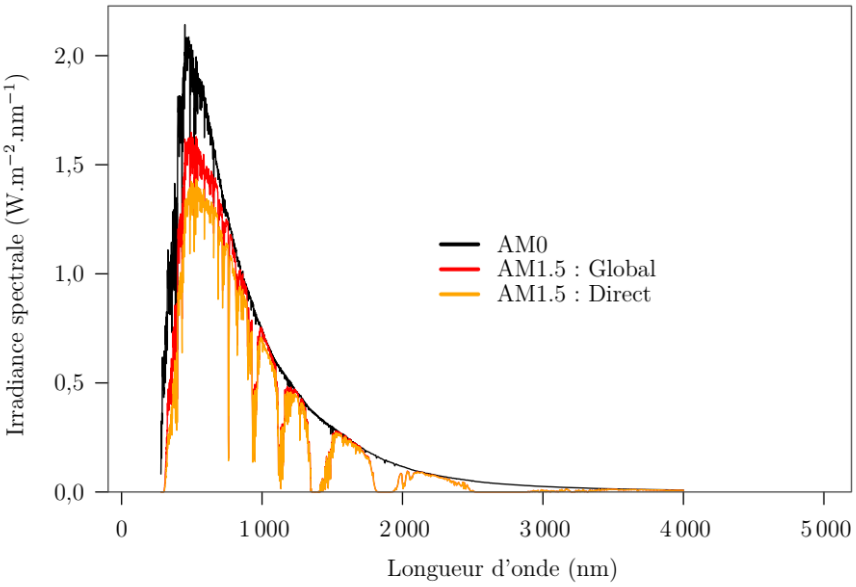


Figure 1.1 L'irradiance spectrale du spectre solaire au sommet de l'atmosphère (AM0) et à la surface terrestre pour le rayonnement global et le rayonnement direct (AM1.5). *Données utilisées :* (ASTMG173-03, 2012)

Les rayonnements global, direct et diffus

A la surface de Terre, le rayonnement solaire total incident par unité de surface mesuré sur une surface horizontale terrestre est appelé irradiance horizontale globale (GHI). Il est composé de deux composantes : l'irradiance normale directe (DNI) et l'irradiance horizontale diffuse (DHI). La figure 1.2 représente la décomposition du rayonnement solaire total suite aux interactions du rayonnement avec l'atmosphère en sa partie directe et diffuse. La relation entre GHI, DNI et DHI s'exprime dans l'équation 2 :

$$GHI = DHI + DNI \times \cos(\alpha_{zenith}) \quad (2)$$

avec α_{zenith} l'angle zénithal utilisé pour déterminer la position du soleil par rapport à un endroit spécifique de la surface de la Terre.

- L'irradiation normale directe ou DNI est la partie du GHI reçu directement la Terre
- L'irradiation horizontale diffuse (DHI ou DIF) est la partie du rayonnement solaire totale qui atteint la surface la Terre indirectement. La vapeur d'eau, les aérosols et les nuages (comme mentionné précédemment) réfléchissent et absorbent le rayonnement solaire et la diffusent dans l'atmosphère.

L'albédo est la partie du rayonnement réfléchie par le sol.

Le GHI est une variable essentielle pour l'exploitation de l'énergie solaire et par conséquent la production photovoltaïque (PV).

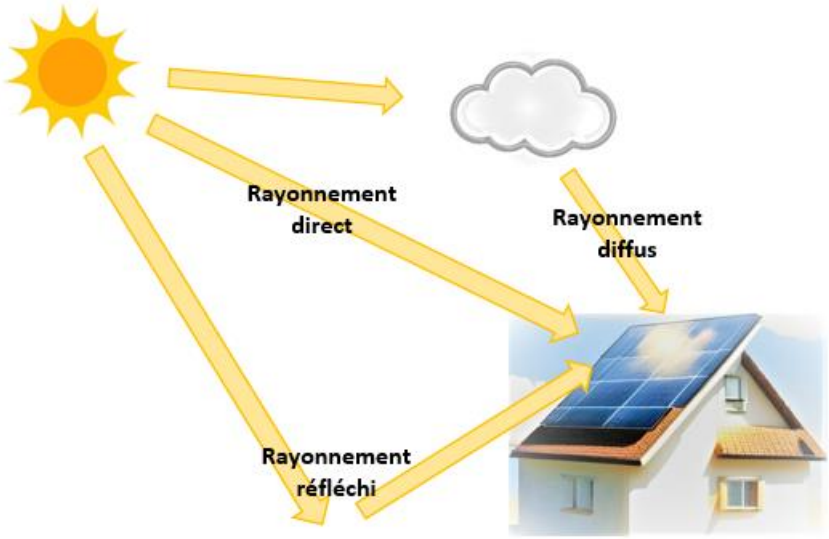


Figure 1.2 Les composants du rayonnement solaire total à la surface de la Terre

1.1.2 Énergie photovoltaïque

Le mot « photovoltaïque » provient du grec « phôtos » qui signifie *lumière* et de « Volta » du nom du physicien italien qui découvrit, en 1800, la pile électrique. Il désigne le processus physique consistant à convertir l'énergie lumineuse en énergie électrique par transfert de l'énergie des photons aux électrons d'un matériau. L'effet photovoltaïque est un principe fondamental découvert par Antoine Becquerel en 1839 et expliqué plus tard en 1905 par Albert Einstein.

1.1.2.1 Principe de fonctionnement

Une cellule photovoltaïque est l'unité essentielle d'un système de production d'énergie solaire dans lequel la lumière du soleil est rapidement convertie en énergie électrique. La cellule solaire est un dispositif à jonction p-n. Le type n fait référence aux électrons chargés négativement donnés par les atomes d'impureté donneurs et le type p fait référence aux trous chargés positivement créés par les atomes d'impureté accepteurs (Smith et al., 2018).
Le principe de fonctionnement des cellules solaires est basé sur l'effet photovoltaïque. Comme le montre la figure1.3, ce dernier peut être divisé en trois procédures essentielles (Al-Ezzi & Ansari, 2022):

- L'absorption de photons dans un semi-conducteur électronique à jonction p-n pour générer des porteurs de charge (paires électron-trou). L'absorption d'un photon dont l'énergie ($E = h\upsilon$) est supérieure à l'énergie de gap "Eg" du matériau semi-conducteur dopé signifie que son énergie est utilisée pour exciter un électron de la bande de valence "Eυ" vers la bande de conduction "Ec", laissant un vide (trou) au niveau de valence. L'énergie cinétique supplémentaire est donnée à l'électron ou au trou par l'énergie excédentaire du photon ($h\upsilon - h\upsilon_0$). $h\upsilon_0$ est l'énergie minimale ou la fonction de travail du semi-conducteur requise pour générer une paire électron-trou. La fonction de travail représente ici l'écart d'énergie. L'énergie excédentaire est dissipée sous forme de chaleur dans le semi-conducteur (Smets et al., 2016).

- Dans un circuit solaire externe, les trous peuvent s'éloigner de la jonction à travers la région p, et les électrons peuvent s'écouler à travers la région n et traverser le circuit avant de se recombiner avec les trous.
- Enfin, les électrons séparés peuvent être utilisés pour alimenter un circuit électrique. Après avoir traversé le circuit, les électrons se recombinent avec les trous. Le type n doit être conçu plus fin que le type p. Ainsi, les électrons peuvent traverser le circuit en peu de temps et générer du courant avant de se recombiner avec les trous. En outre, un revêtement antireflet est appliqué sur la couche n afin de réduire la réflexion de la surface et d'améliorer la transmission de la lumière vers le matériau semi-conducteur.

La figure 1.3 illustre ce phénomène.

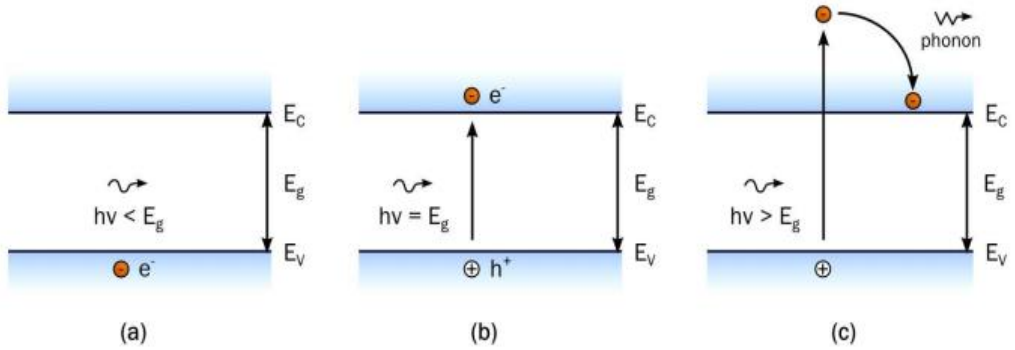


Figure 1.3 (a) Transmission d'un photon d'énergie $E_{photon} < E_g$ dans un matériau semi-conducteur. (b) Absorption d'un photon d'énergie $E_{photon} = E_g$ et formation d'une paire électron-trou. (c) Absorption d'un photon d'énergie $E_{photon} > E_g$, formation d'une paire électron-trou et thermalisation de l'électron par émission de chaleur (phonon). Source :(Roger, 2014)

1.1.2.2 Les cellules photovoltaïques et leurs technologies

Les cellules solaires photovoltaïques ont évolué au cours du temps et les générations de cellules décrivent fondamentalement les étapes de leur évolution à ce jour. Depuis l'invention des cellules solaires au cours des dernières décennies, il existe quatre catégories principales, connues sous le nom de "générations de technologies de cellules photovoltaïques" (Jayawardena et al., 2013).

i. Première génération : Le silicium monocristallin et polycristallin ainsi que l'arséniure de gallium (GaAs) sont les technologies de cellules photovoltaïques incluses dans cette catégorie. Cette génération est donc limitée aux "technologies basées sur le silicium cristallin". L'efficacité de ces cellules solaires PV à base de silicium a augmenté jusqu'à 24% depuis leur découverte dans les laboratoires Bell (Green et al., 2005). Toutefois, bien que 86% du marché de l'énergie solaire PV est couvert des cellules en silicium, les modules commerciaux n'ont qu'une efficacité maximale de conversion de 15 à 20% (Saha, 2015).

ii. Deuxième génération : Cette génération couvre les développements des technologies de cellules solaires PV de première génération ainsi que les développements des cellules solaires à couches minces en silicium microcristallin (µc-Si) et en silicium amorphe (a-Si), les cellules solaires au séléniure de cuivre, d'indium et de gallium (CIGS) et les cellules solaires au tellurure de cadmium et au sulfure de cadmium (CdTe/CdS) (Britt & Ferekides, 1993; Peumans et al., 2003). Ces cellules moins coûteuses, plus fines et plus flexibles ont une efficacité d'environ 20%.

iii. Troisième génération : Les technologies photovoltaïques basées sur des composés plus récents ont été incluses dans cette génération. En outre, les technologies basées sur les films nanocristallins GaAs/GaInP qui sont des points quantiques actifs, les cellules solaires sensibles aux colorants, les cellules solaires à base de polymères organiques, etc. sont également incluses dans cette génération. Leur efficacité global a été augmenté jusqu'à 12% à l'échelle du laboratoire (Keis et al., 2002).

iv. Quatrième génération : La faible flexibilité ou le faible coût des polymères en couches minces font partie de cette génération, de même que la fermeté des nanostructures inorganiques innovantes telles que les oxydes métalliques et les nanoparticules métalliques ou les nanomatériaux organiques, c'est-à-dire le graphène, les nanotubes de carbone et les dérivés du graphène. Cette catégorie est généralement appelée organique. Les points quantiques semi-conducteurs inorganiques ont été inclus dans les cellules photovoltaïques polymères/organiques. Les nouveaux dispositifs de production d'énergie photovoltaïque sont considérés comme faisant partie de la quatrième génération de la technologie des cellules solaires photovoltaïques, ces dispositifs souvent appelés "nano photovoltaïques" peuvent devenir l'avenir des cellules solaires photovoltaïques avec de grandes perspectives, bien que pour l'instant l'efficacité de ces panneaux soit faible, de l'ordre de 7% (Jung Xu, 2012).

1.1.2.3 Application

En plus de son caractère renouvelable, l'énergie solaire PV a pour intérêt d'être extrêmement modulable. En effet, il est possible de concevoir des centrales électriques PV de plusieurs centaines de MWc (Mégawatts crête), tout comme des alimentations pour dispositifs autonomes en énergie, dont la surface peut descendre jusqu'à quelques cm^2. Quelques exemples sont présentés dans la figure 1.4. Dans le cas des dispositifs nomades, le photovoltaïque est très prometteur, car l'énergie solaire est disponible en tout point du globe. De plus, comparé à d'autres formes de production d'énergie (éolien, nucléaire, solaire thermique, biomasse...), le photovoltaïque nécessite relativement peu de matière et de volume. Aucun élément mécanique mobile n'est indispensable, ce qui exclut les problèmes d'usure. Ainsi, les applications du PV en tant que source d'énergie autonome sont aujourd'hui fortement diversifiées. Cela inclut des objets de petite dimension ayant une faible consommation, tels que les chargeurs d'appareils électroniques nomades, les alarmes ou pour alimenter l'Internet des objets (Internet of Things, IoT) grâce à des circuits à très faible consommation d'énergie (El-Damak & Chandrakasan, 2015).

Mais l'énergie photovoltaïque est aussi de plus en plus utilisée comme source d'énergie pour des objets plus grands, tels que les véhicules. Initialement utilisée comme énergie complémentaire pour les équipements électriques d'automobiles, de bateaux ou d'avions, son utilisation comme source principale a aussi été démontrée sur ces mêmes types d'appareils équipés de moteurs électriques. L'exemple le plus remarquable est le projet suisse Solar Impulse 2, au sein duquel un prototype d'avion solaire a pu voler durant un cycle complet jour/nuit sans autre source d'énergie et traverser l'Atlantique en 71 heures en 2016 ("Solar Impulse Completes Historic Round-the-World Flight," 2016).

Figure 1.4 Exemples d'applications du PV à de multiples échelles. a) cellule photovoltaïque dans les calculatrices à énergie solaire. b) Ferme photovoltaïque de Pierrefonds, au sud de la Réunion. Puissance de 3MWc. c) Exemple d'intégration au bâtiment : la voile photovoltaïque de la Cité de la Musique, construite sur l'île Seguin. d) La voiture solaire appelée « Lightyear One » (Mathijsen, 2021).

Dans le secteur du bâtiment, la conception de systèmes photovoltaïques intégrés au bâtiment (Building Integrated PhotoVoltaics, BIPV) permet une transition plus rapide vers des bâtiments à consommation énergétique nette nulle (Hamzah & Go, 2023).

Toutefois, l'utilisation du PV ne se limite pas au domaine terrestre. Son application initiale concernait le domaine spatial (Verduci et al., 2022). Ainsi, les satellites, surtout ceux en orbite autour de la Terre ou destinés aux planètes du système solaire (comme par exemple, Vénus, Mercure, etc.) et autres équipements nécessitaient une alimentation stable et interrompue (Datas & Martí, 2017). Le photovoltaïque est particulièrement adapté à ce champ

d'applications et ces missions, grâce à l'encombrement et au poids limité et à la disponibilité de l'énergie solaire en dehors de l'atmosphère terrestre.

1.2 Variabilité de l'énergie solaire et intérêt de la prévision photovoltaïque

La production des énergies renouvelables (EnR) a connu une énorme croissance mondiale ces dernières décennies. Toutefois, contrairement à de nombreuses sources de production conventionnelles, de nombreuses ressources renouvelables, notamment l'énergie solaire PV, sont considérées comme de la production variable. Elles ont une limite de production maximale qui change dans le temps, c'est l'intermittence de la ressource solaire qui conduit indubitablement à la variabilité de la production solaire PV.

La variabilité du rayonnement solaire au sol s'exerce à différentes échelles temporelles et un ou plusieurs phénomènes physiques peuvent être associés à ces variabilités.

Comme cela a été mentionné dans l'Introduction générale, le mouvement des nuages et des constituants de l'atmosphère contribuent à une forte variabilité du rayonnement solaire infra journalier. La complexité des mouvements atmosphériques dynamise le cycle des nuages. Cette variabilité climatique impactant le Sud-Ouest de l'Océan Indien, et plus précisément l'île de La Réunion, sera détaillée dans la partie Contextualisation du chapitre 2.

La rotation de la Terre autour de son axe reliant les pôles géographiques entraine aussi une variabilité journalière par l'alternance jour/nuit ; les pôles recevant en moyenne moins d'ensoleillement que les zones intertropicales. De plus, la révolution de la Terre autour du Soleil rajoute une variabilité saisonnière (hiver/été austral) à ces variations du rayonnement solaire au sol.

1.2.1 Prise en compte et conséquence de l'intermittence

Les énergies renouvelables représentent plus de la moitié des investissements dans la production électrique et une part de plus en plus importante de ces énergies dans le mix énergétique des pays européens comme l'indique l'INSEE, avec 62,6% d'énergie renouvelable dans la consommation finale brute d'énergie en 2021 (Insee, 2023). La stabilité du réseau électrique doit être assurée grâce d'abord à des moyens de production modulable d'électricité (charbon, gaz, biomasse, etc.) dans le mix énergétique. Cela permet de garantir la stabilité du réseau en compensant des chutes éventuelles de l'offre. En France, le principal gestionnaire du réseau de distribution est ENEDIS qui a pour but de distribuer l'électricité aux consommateurs de l'échelle régionale à l'échelle locale par ligne à moyennes et basses tensions, tout cela en équilibrant l'offre et la demande d'électricité. De ce fait, pour les gestionnaires de réseau, la variabilité et l'incertitude de la production solaire PV sont un challenge pour les gestionnaires de réseau que ce soit pour la qualité de la production en courant, tension et fréquence mais aussi en coûts opérationnels d'ajustement en temps réel (Jha & Shaik, 2023).

Traditionnellement, les gestionnaires optent pour des prévisions de la demande en électricité pour planifier les unités de productions en utilisant des algorithmes d'optimisation (Bashir & El-Hawary, 2007) mais face à cette croissance de la part des énergies renouvelables intermittentes dans le mix énergétique, il est indispensable en plus de pouvoir prédire de

manière précise et fiable leur production à tout instant. Cela devient encore plus problématique pour des réseaux insulaires (Diagne, 2015; Voyant, 2011) où l'absence ou le manque d'interconnexion avec un réseau voisin ou continental nécessite un pilotage intelligent des réserves.

1.2.2 Intérêt de la prévision photovoltaïque pour le réseau électrique

Le photovoltaïque apporte une contribution de plus en plus importante à l'approvisionnement en électricité à long terme. La prévision de la ressource solaire et de la production solaire PV devient un outil incontournable du paysage énergétique et de la transition verte. Les stratégies d'équilibre entre la production et la consommation d'électricité à chaque instant et la rentabilité reposent sur la précision, la fiabilité et la puissance de la méthode et de l'outil de prévision.

La prévision solaire PV est importante pour plusieurs raisons :

- Optimisation de la production d'énergie : La production d'énergie photovoltaïque dépend des conditions météorologiques, comme l'ensoleillement, la température et la vitesse du vent. La prévision de ces conditions permet de prédire la production d'énergie à court et moyen terme, ce qui permet aux gestionnaires de centrales photovoltaïques de planifier la production d'énergie et d'optimiser leur rendement. Cela peut également aider à réduire les coûts d'exploitation en évitant les surproductions ou sous-productions d'énergie.
- Intégration dans le réseau électrique : La production d'énergie photovoltaïque est variable et dépendante des conditions météorologiques. La prévision de la production d'énergie peut aider les gestionnaires de réseau électrique à mieux intégrer l'énergie photovoltaïque dans le réseau électrique, ce qui peut contribuer à améliorer la stabilité et la fiabilité du réseau.
- Planification et développement de projets : Les prévisions de production d'énergie photovoltaïque peuvent également aider les développeurs de projets à planifier l'installation de nouvelles centrales photovoltaïques. En utilisant des modèles de prévision, les développeurs peuvent évaluer la faisabilité de nouveaux projets et planifier leur production d'énergie future.
- Réduction des coûts d'exploitation : Les prévisions de production d'énergie photovoltaïque peuvent également aider à réduire les coûts d'exploitation. En prévoyant la production d'énergie, les gestionnaires de centrales photovoltaïques peuvent ajuster leur production en temps réel pour éviter les coûts élevés associés à l'utilisation de combustibles fossiles pour compenser les fluctuations de la production PV.

La prévision photovoltaïque est donc essentielle pour maximiser le rendement de la production d'énergie photovoltaïque, améliorer l'intégration de l'énergie photovoltaïque dans le réseau électrique et réduire les coûts d'exploitation (Shafiullah et al., 2022).

1.3 Horizons temporels et état de l'art de la prévision de production photovoltaïque

Pour évaluer les avantages de la réponse à la demande, il est important que les décideurs soient conscients de l'infrastructure physique et des exigences opérationnelles nécessaires à la construction et à l'exploitation fiable d'un réseau électrique, ainsi que des différences régionales dans la structure du marché et l'organisation de l'industrie. Dans toutes les structures de marché,

la gestion des réseaux électriques est largement déterminée par deux propriétés physiques importantes de la production d'électricité. Premièrement, l'électricité n'est pas économiquement stockable, ce qui nécessite de maintenir l'équilibre entre l'offre et la demande au niveau du système en temps réel. Les déséquilibres entre l'offre et la demande peuvent menacer l'intégrité du réseau électrique sur des zones extrêmement vastes en quelques secondes. Deuxièmement, l'industrie de l'énergie électrique est très capitalistique. Les investissements dans les systèmes de production et de transmission sont des projets complexes et de grande envergure, dont la durée de vie économique prévue est de plusieurs décennies et dont le développement, la mise en place et la construction prennent souvent de nombreuses années. Ces caractéristiques des systèmes d'énergie électrique nécessitent une gestion de l'électricité sur plusieurs échelles de temps, allant d'années (voire de décennies) pour la planification et la construction de la production et du transport, à des secondes pour l'équilibrage de la fourniture d'énergie par rapport aux fluctuations de la demande. Les décisions sont prises à plusieurs moments de cette échelle de temps (figure 1.5). En règle générale, la quantité de charge engagée à chaque étape diminue à mesure que l'horizon temporel se rapproche de la livraison de l'électricité. Par exemple, 70 à 80 % de la charge fournie est souvent engagée par des contrats d'énergie à terme, des mois, voire des années avant la livraison. La quantité d'énergie prévue pour le jour suivant varie, mais représente généralement 10 à 25 % des besoins totaux. Dans la plupart des cas, moins de 5 % de l'approvisionnement est engagé dans les deux dernières heures avant sa livraison (U.S. Department of Energy, 2006).

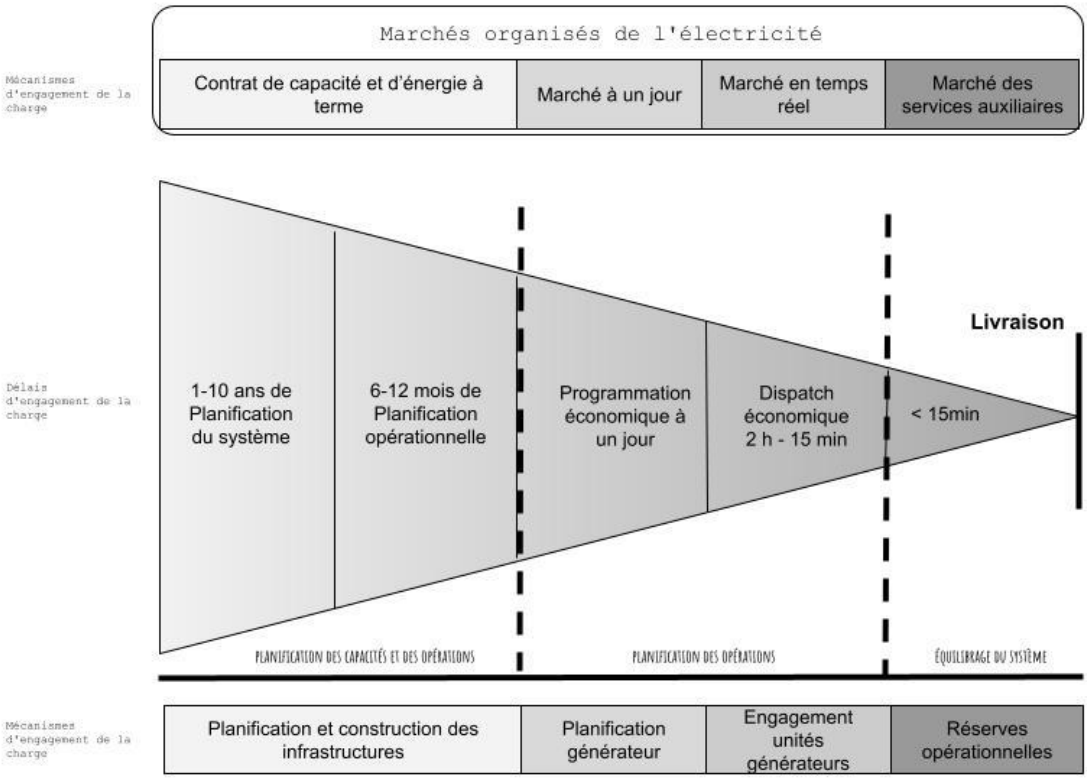


Figure 1.5 Planification et programmation des systèmes électriques : Échelles de temps et mécanismes de décision. Source :(U.S. Department of Energy, 2006)

Ainsi, le choix d'une méthode de prévision dépend de plusieurs paramètres à savoir le besoin auquel répond la prévision (participation au marché, planification, etc.), l'horizon de prévision et les données à disposition. Ces trois éléments sont intimement liés. Le besoin définit les horizons de prévision à choisir ainsi que les données nécessaires à cette prévision de production PV. Il existe diverses sources de données utilisables dans le cadre de la prévision de production PV, à savoir les mesures de production et de variables météorologiques comme l'irradiance solaire, les prévisions météorologiques, les images de caméras ou de satellites. Une approche intéressante est de regrouper les modèles de prévisions par horizons croissants de quelques minutes à plusieurs jours. Il faut toutefois noter la différence fondamentale entre les termes horizon et résolution temporelle de prévision. L'horizon désigne la période de temps sur laquelle se porte la prévision tandis que la résolution correspond à sa fréquence des données de prévision, sa granularité.

Dans le cadre de la prévision PV on retrouve la classification suivante en fonction des horizons de prévision (Zamo et al., 2014a) :

— des prévisions intra-horaires (de 15 min à 1 heure avec un pas temps de 1min) ;
— des prévisions à très court-terme pour des horizons de quelques heures (≤ 1h - 6h) ;
— des prévisions à court terme pour des horizons de quelques jours (1 jour - 3 jours) ;
— des prévisions à moyen-terme (1 semaine - 3/4 mois) ;
— des prévisions à long-terme (≥ 1 an).

Les prévisions intra-horaires et très court-terme qui couvrent des horizons allant de moins de quelques minutes à quelques heures sont essentielles aux évènements extrêmes (rampes), de la variabilité impactant les opérations, de suivi de la production, d'ajustement de la charge et de gestion du stockage. La prévision à court terme est utilisée dans le cadre des démarrages des unités de production, du management et du trading d'énergie. Elle est populaire sur le marché de l'électricité, où les décisions comprennent la répartition économique de la charge et l'exploitation du système électrique. Il est également utile pour le contrôle des systèmes de gestion intégrée de l'énergie des énergies renouvelables. La prévision à long terme quant à elle permet une meilleure couverture, planification et optimisation des ressources (M. Diagne et al., 2013). De nombreuses études ont démontré que les horizons de prévision et qu'à modèle de prévision et autres paramètres constants, la précision de la prévision varie en fonction de l'évolution de l'horizon temporel (Das et al., 2018). On retrouve dans la littérature des comparaisons de méthodes de prévision pour des horizons court et très court-terme (Zamo et al., 2014a, 2014b) et des analyses détaillées de ces méthodes suivant le type de données d'entrées (Inman et al., 2013). Dans la suite, un état de l'art exhaustif des méthodes de prévision de la production solaire PV.

Dans le paragraphe suivant sont décrites les différentes méthodes de prévision de production d'énergie photovoltaïque. Dans la littérature, les techniques de prévision PV sont généralement classées en quatre grandes catégories, à savoir la méthode de persistance, les méthodes statistiques, les méthodes physiques et les méthodes hybrides (Antonanzas et al., 2016; Inman et al., 2013; Pedro et al., 2018; Sobri et al., 2018; Voyant et al., 2017). La figure

1.6 représente ces quatre grandes familles et leurs méthodes de prévision de production solaire PV.

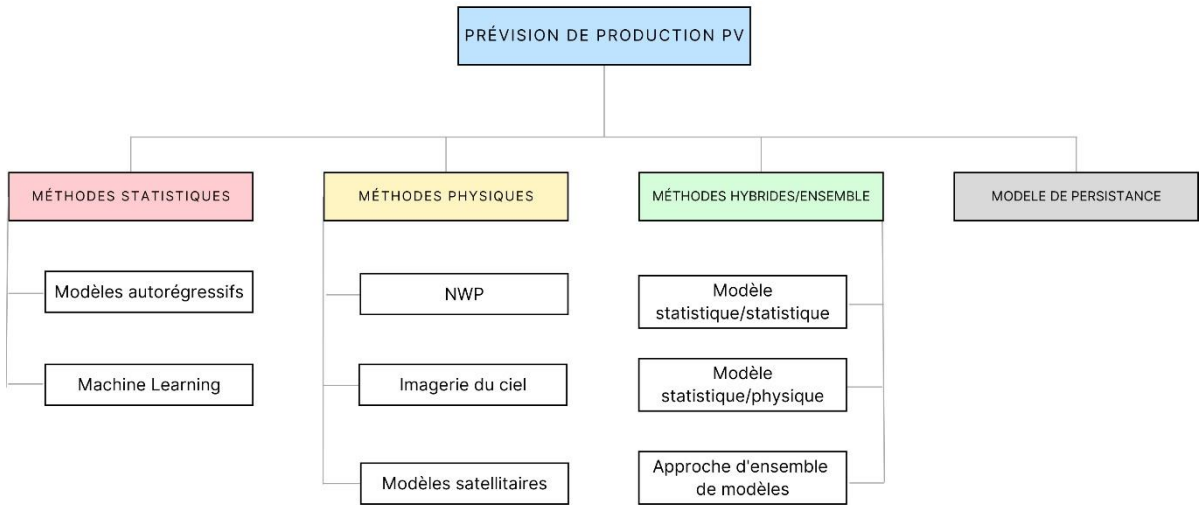


Figure 1.6 Représentation des familles de méthodes de prévision de production PV

1.3.1 Le modèle de persistance

Le modèle de persistance est un prédicteur de type « naïf » très souvent utilisé pour les prévisions à très court terme et à court terme du rayonnement global et de la puissance de production PV. Il présente un coût de calcul réduit, un faible délai et une précision raisonnable. Cette technique adopte le concept d'aujourd'hui égal à demain. En d'autres termes, les conditions climatiques (c'est-à-dire l'irradiance solaire) d'un jour à l'autre devraient rester similaires à celles de la veille (Dutta et al., 2017). Ainsi, si aujourd'hui est une journée ensoleillée avec une température de 30°C, le modèle prévoit que demain sera une journée ensoleillée avec une température de 30°C. Par conséquent, le modèle peut calculer instantanément le GHI en décomposant le GHI de prévision en GHI de ciel clair de calcul et en projetant l'indice de ciel clair. Cependant, l'indice de ciel clair ne réagit pas à la variation de l'angle zénithal du soleil due aux conditions météorologiques telles que les nuages, les précipitations, les tempêtes, etc. dans la fenêtre temporelle de prévision (Kumler et al., 2019).

En fait, il s'agit de la répétition d'une mesure de l'instant t à l'instant t+h (dans le cas h = 1 on obtient l'équation 3). 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 (Équation 4 pour le cas h = 1).

$$\hat{x}_{t+1} = x_t \tag{3}$$

$$\hat{x}_{t+1} = x_{t+1-T} \tag{4}$$

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. Ce modèle confirme généralement les performances d'autres paradigmes de prévision. C'est pourquoi différents modèles sont comparés à lui lorsque les conditions climatiques persistent jusqu'au lendemain ; cependant, lorsque l'horizon temporel augmente, la précision de ce modèle diminue considérablement.

Il sera confronté, dans le chapitre 5, aux modèles développés au cours de cette thèse.

1.3.2 Les modèles physiques

Les modèles physiques utilisent des modèles mathématiques qui prennent en compte les paramètres météorologiques et internes à la centrale PV, tels que la position du soleil, l'angle d'inclinaison des panneaux solaires, la température et la vitesse du vent. Ces modèles sont appelés aussi modèles de performance. Ces modèles sont très précis, mais ils nécessitent une grande quantité de données in situ et sont souvent complexes à mettre en œuvre. Ces données sont issues de modèles présentés si après :

- **Numerical Weather Prediction (NWP)**

La prévision physique implique : la pression atmosphérique, la rugosité de la surface, la température, les obstacles et les perturbations pour les prévisions futures (Zhao et al., 2016) Cette technique est généralement plus fiable pour les prévisions à long terme (Yang & Wang, 2018). Basée sur cette méthode, la technique NWP fusionne les informations météorologiques et les équations du modèle atmosphérique pour obtenir des prévisions. Ce modèle est généralement classé en deux catégories en fonction de l'échelle, à savoir le modèle régional et le modèle global. Le modèle régional traite des caractéristiques atmosphériques pour une zone restreinte telle que les régions, les pays ou les continents (Monteiro et al., 2013). Parmi les modèles régionaux, citons le modèle WRF (Weather Research and Forecasting Model) qui sera développé dans le chapitre 2 et utilisé comme simulateur de données météorologiques générées de manière cohérente entre elles, NAM (North American Mesoscale), RAP ou RAR (Rapid Refresh) qui offre un accès gratuit aux données, tandis que le modèle global délimite les caractéristiques de l'atmosphère à l'échelle mondiale. En outre, pour ce modèle mondial de prévision numérique du temps, il existe environ 15 services météorologiques actifs dans l'acquisition de données : Global Forecast System (GFS), Climate Forecast System (CFS) et Global Data Assimilation System (GDAS), etc., qui sont gérés par des organisations gouvernementales telles que la NOAA et l'European Centre for Medium-Range Weather Forecasts (ECMWF). Les modèles de prévision numérique du temps peuvent prévoir l'état du climat plus de 15 jours à l'avance (Lorenz et al., 2011) et utilisent un ensemble d'équations numériques pour l'état physique et les caractéristiques dynamiques de l'atmosphère, simultanément. Le tableau 1.1 présente les principales caractéristiques de ces modèles

Cependant, pour la prévision de la puissance de sortie PV, le modèle utilise des caractéristiques météorologiques particulières telles que l'indice de serre, l'humidité relative, la vitesse et la direction du vent (Lorenz et al., 2011; Pelland et al., 2013) et la qualité de la prévision est donc meilleure si les variables météorologiques restent stables. En revanche, des prévisions erronées se produisent lorsque les valeurs des variables météorologiques changent brusquement.

Tableau 1.1 : Caractéristiques de quelques modèles NWP

Modèles	ECMWF	GFS	NAM	RAP
Résolution spatiale horizontale (km)	16	50	12	13
Résolution temporelle en sortie (heures)	3	3	1	1
Horizon de prévision	6 j	8 j	36 h	18 h

Les sorties des modèles NWP sont très rarement utilisées brutes. Elles peuvent être couplées à des observations d'images satellites ou de caméras hémisphériques qui permettent de décrire la couverture nuageuse au niveau des centrales par l'étude des mouvements des nuages (Perez et al., 2007, 2010). Ces sorties peuvent aussi être injectées dans des modèles statistiques afin de prévoir la production PV. Ces modèles présentent deux limitations majeures. La première concerne le temps de calcul important. Et la deuxième concerne la résolution spatiale de ces modèles qui rend impossible la descente à des niveaux microscopiques de la physique qui sont associés à la formation des nuages.

- **L'imagerie du ciel**

L'intégration de ces technologies avec des données historiques présente un avantage en termes de connaissances disponibles, car elle permet de déterminer la position et le mouvement des nuages. (Magnone *et al.* 2017) divise le processus de génération de prévisions d'irradiance solaire à l'aide d'images du ciel entier (All-Sky Images, ASI) en plusieurs étapes : le prétraitement des images, puis la détection des nuages, l'identification du mouvement des nuages, et enfin l'élaboration des prévisions d'irradiance solaire. Pour atteindre cet objectif, les informations sur le flux de nuages sont généralement traitées à l'aide d'algorithmes de vision par ordinateur afin de prévoir les niveaux futurs d'occultation du soleil en raison de la couverture nuageuse (fraction de nuage). Les méthodologies basées sur l'image du ciel consistent principalement en deux processus : la détection des nuages et la prévision de l'irradiance solaire (Kamadinata et al., 2019). La méthode utilise généralement un imageur de ciel au sol pour classer les nuages en fonction des informations sur les caractéristiques de texture des images du ciel, puis peut utiliser des algorithmes (c'est-à-dire l'algorithme de correspondance des blocs ou l'algorithme de flux optique) pour construire le modèle de calcul pour la prédiction de l'irradiance solaire. La plupart des techniques de prédiction utilisant des images du ciel se sont concentrées sur la détection et le suivi des nuages.

- **L'imagerie satellitaire**

L'imagerie satellitaire est l'une des techniques physiques les plus utilisées pour générer des données sur la couverture nuageuse et fournir ainsi des informations sur l'irradiance solaire et sur d'autres paramètres météorologiques locales comme la température et la vitesse du vent. Grâce à sa haute résolution spatiale (jusqu'à 2,5 km × 2,5 km) et temporelle (jusqu'à une résolution temporelle de 30 min), elle constitue une base fiable pour la prévision de l'irradiance solaire et in fine de la production solaire PV. Les méthodes de prévision par satellite peuvent

fournir des données sur le rayonnement solaire à très court terme, jusqu'à environ 6 heures à l'avance (Arbizu-Barrena et al., 2017).

1.3.3 Les modèles statistiques

Les méthodes statistiques utilisent des données historiques de production photovoltaïque pour prédire la production future. Les modèles statistiques comprennent entre autres les méthodes de Machine Learning (ML), qui comprend les réseaux de neurones, les Support Vector Machine (SVM) et les arbres de décision entre autres, et les méthodes régressives et autorégressives. Ces modèles sont plus faciles à mettre en œuvre que les modèles basés sur la physique et n'ont pas besoin d'informations sur les états internes du système pour le modéliser mais ils peuvent manquer de précision si les conditions de production sont très différentes de celles des données historiques (comme une modification de la structure ou la taille de la centrale PV).

- **Les modèles régressifs et autorégressifs**

Les modèles régressifs ont été largement utilisés pour la prévision, de la régression linéaire multiple aux modèles non linéaires. Pour la prévision à court terme de l'irradiance solaire, les méthodes statistiques linéaires les plus utilisées sont celles basées sur la méthode autorégressive de Box-Jenkins; la technique fournit un cadre rapide, compréhensible et statistiquement rigoureux pour identifier, estimer et valider les modèles de prévision des paramètres météorologiques tels que la température ambiante, l'irradiance solaire, la vitesse du vent, etc. (Bin Shams et al., 2016). Parmi ces méthodes, nous pouvons évoquer des modèles qui sont notamment utilisés dans d'autres secteurs comme en Économétrie mais qui ont été réappropriés pour les besoins au domaine énergétique. Le modèle de cointégration d'Engle & Granger (Engle & Granger, 1991) a été repris par (Ramenah et al., 2017) pour établir une relation non fallacieuse entre le rayonnement solaire et la production PV afin d'en faire la prévision. Afin d'appliquer plus de paramètres climatiques à la prévision de production PV (Fanchette et al., 2020; Ramenah et al., 2023) ont utilisé cette fois la méthode de cointégration de Johansen. Elle s'appuie statistiquement sur plusieurs variables météorologiques explicatives afin de prévoir une variable expliquée notamment, ici, la production PV.

Les séries temporelles semblent avoir une performance efficace en termes de précision pour des prédictions de l'ordre de quelques minutes ou de quelques heures et peuvent être combinées avec d'autres méthodes (c'est-à-dire NWP, ANN et techniques d'apprentissage automatique) pour créer des modèles hybrides capables de reproduire des comportements dans les données d'entrée (Reikard & Hansen, 2019). En fonction des facteurs à imiter, différentes variantes du modèle autorégressif à moyenne mobile (AutoRegressive Moving Average, ARMA) de Box-Jenkins ont été développées. Des modèles comme le modèle à moyenne mobile intégrée autorégressive (AutoRegressive Integrated Moving Average, ARIMA) en prenant en compte des informations externes pour l'analyse (Zhang, 2003), SARIMA (incluant la caractéristique de saisonnalité), ARMAX et ARIMAX, sont les méthodes les plus utilisées dans la prévision de l'irradiation solaire grâce à la flexibilité de modéliser des modèles

complexes et à la simplicité du modèle à assimiler et à ajuster pour leur application dans les nouvelles technologies (Bin Shams et al., 2016).

Ainsi, le modèle ARIMA est une généralisation du modèle de moyenne mobile intégrée autorégressive qui représente un exemple important de l'approche de Box et Jenkins (Box & Jenkins, 1976) pour la modélisation des séries chronologiques. En particulier, le modèle SARIMA contient une composante saisonnière et il est largement utilisé pour l'analyse et la prévision des séries chronologiques (cf. chapitre 4.1).

La première étape avant de traiter des modèles complexes de type 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 ARIMA puis SARIMA.

On appelle d'un processus MA(q) un processus linéaire x_t (avec ϵ_t un bruit blanc de variance σ^2 et θ le coefficient de régression) qui vérifie une relation de type :

$$x_t = \sum_{i=0}^{q} \theta_i . \epsilon_{t-1} \tag{5}$$

Une série temporelle peut être prédite par un modèle autorégressif paramétrique. La prédiction pour une série x_t peut être décrite par un modèle AR(p) (avec ψ le coefficient de régression) :

$$x_t = \sum_{i=1}^{p} \psi_p . x_{t-1} + \epsilon_t \tag{6}$$

En couplant le modèle AR(p) et MA(q), nous obtenons le modèle ARMA(p,q). Il s'agit d'un processus x_t vérifiant la relation :

$$\psi(L) x_t = \theta(L) \epsilon_t \tag{7}$$

où L est l'opérateur retard

Le modèle ARIMA(p,d,q) est un processus x_t pour lequel le processus différencié d'ordre d, noté $(1 - \mathrm{L})^{\mathrm{d}} x_t$ vérifie la relation :

$$\psi_p(L)(1 - L)^d x_t = \theta(L) \epsilon_t \tag{8}$$

Dans le cas des modèles saisonniers, un type de modèle est basé sur le même principe que le modèle ARIMA tout en s'affranchissant de la périodicité saisonnière s. Ce modèle s'appelle SARIMA(p,d,q)(P,D,Q) et s'écrit sous la forme :

$$\psi_p(L) \psi_P(L)^s (1 - L)^d (1 - L^s)^D x_t = \theta_q(L) \theta_Q(L)^s \epsilon_t \tag{9}$$

où N est le nombre d'observations, p, d, q, P, D et Q sont des ordres et s la périodicité saisonnière.

Sur de longues périodes, la prévision numérique a démontré une meilleure performance dans la prévision de l'irradiation solaire, cependant, lorsque les informations historiques doivent être collectées sur des intervalles plus courts, les séries temporelles sont connues pour être plus précises (Reikard & Hansen, 2019). Comme le modèle de persistance, ce modèle SARIMA sera aussi confronté, dans le chapitre 5, aux modèles développés au cours de cette thèse.

- **Les méthodes de Machine Learning**

La production d'énergie photovoltaïque dépend fortement des conditions météorologiques telles que l'irradiance solaire et la température. Par conséquent, la production de cette source d'énergie PV fluctue, ce qui rend difficile pour les gestionnaires de réseau électrique d'équilibrer la production et la consommation d'électricité lors de l'utilisation de systèmes photovoltaïques. C'est pourquoi plusieurs algorithmes ML ont été mis en œuvre pour prévoir l'irradiance solaire et la puissance de sortie des systèmes PV (Gaviria et al., 2022).

Les machines à vecteurs de support (Support Vector Machine, SVM) sont une technique qui trouve un modèle linéaire pour la classification (Géron, 2022). Le critère de la SVM est basé sur le concept de marge maximale, qui se réfère à la distance entre l'hyperplan de séparation et les observations les plus proches dans l'une ou l'autre classe.

Les forêts d'arbres de décisions (Random Forest, RF) sont des modèles basés sur des arbres de décision qui sont construits en parallèle en utilisant des sous-ensembles aléatoires de variables d'entrée. Une fois les arbres de décision formés, un modèle RF prend un exemple, le fait passer par tous les arbres et lui attribue la classe avec le plus grand nombre de votes (D. Liu & Sun, 2019).

Parmi les ML les plus utilisés, nous retrouvons la famille des réseaux de neurones artificiels (Artificial Neural Network, ANN) qui imite le mécanisme de traitement de l'information du cerveau humain. Il a la capacité unique d'approximer des fonctions non linéaires avec une grande fidélité et une grande précision. Il est utilisé dans des domaines aussi divers que les prévisions météorologiques, la finance, la physique, l'ingénierie et la médecine. L'architecture de base de l'ANN est divisée en trois sections : la couche d'entrée, la couche cachée et la couche de sortie, comprenant des neurones artificiels et des connexions. À l'instar des neurones biologiques, chaque neurone artificiel d'un réseau neuronal est un nœud d'activation où se déroulent toutes les activités de traitement de l'information et de prise de décision (Lo Brano et al., 2014). Son architecture sera toutefois plus détaillée dans le chapitre 3 sur les méthodes employées dans cette thèse.

Dans la grande famille des ANN, le réseau de neurones à perceptron multicouche (MLP) est souvent considéré comme le modèle de référence (Mellit & Kalogirou, 2008). Il s'agit d'une technique d'approche ANN élémentaire et efficace pour la conception et la prédiction. Il est toutefois très puissant, de sorte que ce réseau est utilisé pour l'approximation universelle, la modélisation non linéaire et les problèmes complexes qui ne peuvent pas être résolus par un réseau neuronal ordinaire à une seule couche (Azimi et al., 2016). En général, le MLP est un composite de trois couches ou plus de nœuds activés de manière incohérente. Ces nœuds dans n'importe quelle couche sont connectés par un certain poids à d'autres nœuds dans la couche suivante. Il a donc la capacité d'établir une corrélation entre l'entrée et la sortie grâce à un

apprentissage adéquat. La corrélation entre le nombre de nœuds et la couche cachée est essentielle.

Le réseau de neurones récurrent (RNN) est une classe importante d'ANN qui peut apprendre et traiter différentes relations complexes et composées ainsi que des structures informatiques. Ce réseau s'appuie provisoirement sur des données de séries temporelles par le biais d'un système de rétroaction afin d'hériter des valeurs du pas de temps précédent, ce qui démontre des caractéristiques dynamiques temporelles. Le modèle a une structure simple avec une boucle de rétroaction intégrée qui lui permet d'agir comme un moteur de prévision. La sortie du RNN de la couche neuronale concernée est additionnée au vecteur d'entrée suivant et réinjectée dans la même couche, qui est la seule couche de l'ensemble du réseau. (Hertz et al., 1991) ont examiné en détail les applications de base du modèle RNN. Ces applications sont incroyablement polyvalentes, allant du pilotage de voiture sans conducteur à la reconnaissance vocale. Dans un RNN, chaque neurone actif est connecté à tous les autres neurones de traitement et à lui-même. Par conséquent, le résultat de la sortie du RNN dépend du signal de rétroaction du pas de temps précédent et du signal d'entrée. Parmi les RNN, se trouve le modèle à mémoire court et long terme (LSTM) qui est décrit comme une version améliorée et étendue des RNN qui a été appliqué avec succès aux difficultés de prévision de séries temporelles. Il a été signalé que les réseaux RNN sont incapables de traiter les dépendances à long terme dans les données en raison de la disparition du gradient et du problème d'explosion du gradient (B. Hochreiter et al., 2009). Cependant, ce problème a été résolu avec l'introduction des réseaux LSTM introduits par Sepp Hochreiter et Jürgen Schmidhuber (Hochreiter & Schmidhuber, 1997) et ces modèles sont de plus en plus employés dans la littérature pour la prévision de production solaire PV (Gao et al., 2019; Kumar Dubey et al., 2021; Qing & Niu, 2018; F. Wang et al., 2020a).

Le réseau neuronal convolutif (CNN) est une nouvelle approche dans laquelle les principes de connectivité entre les neurones synthétiques imitent l'organisation du cortex visuel animal. Ainsi, il apprend à reconnaître des modèles généralement en mettant en évidence les bords et les comportements des pixels qui sont généralement observés dans diverses images dans ses couches. Il s'agit actuellement du meilleur outil disponible pour la reconnaissance automatique de l'écriture manuscrite, la détection des visages, la reconnaissance du comportement, la reconnaissance vocale, les systèmes de recommandation et la classification des images. Le modèle de calcul utilisé est appelé « Neocognitron » et s'appuie sur des filtres linéaires ou non linéaires pour extraire les caractéristiques d'une image (Jiang et al., 2018). Pour fonctionner efficacement, le CNN se compose de plusieurs blocs tels que la convolution, l'activation et la mise en commun, qui fonctionnent tous ensemble pour l'extraction et la transformation des caractéristiques. Les CNN sont de plus en plus utilisés dans la prévision de production PV (Mellit et al., 2021) et de plus en plus dans des modèles hybrides. Ces modèles ANN et en particulier les modèles LSTM sont pertinents et très utilisés dans la littérature pour la prévision du solaire et de production PV à partir de données sous forme de séries temporelles, d'où l'intérêt d'étudier les données climatiques à notre disposition.

1.3.4 Les modèles hybrides et l'approche d'ensemble de modèles

- **Les modèles hybrides**

Les modèles hybrides combinent un modèle basé sur la physique et un modèle statistique (modèles physique/statistique) ou plusieurs modèles statistiques entre eux (modèles statistique/statistique). La technique de modélisation hybride permet la synergie de deux ou plusieurs méthodes différentes afin d'exploiter les meilleurs attributs de chacune d'entre elles L'essor de ces modèles a mis en avant leur performance élevée face aux modèles unique ou « stand-alone ». Par exemple, CNN et LSTM ont été incorporés dans un modèle hybride par (Zang et al., 2020), le couplage d'un modèle d'interférence floue avec RNN pour la production d'énergie solaire a été réalisé par (Yona et al., 2013)la combinaison d'une transformée en ondelette (Wavelet Transform, WT) et un modèle CNN amélioré par une régression quartile est mis en avant par (Wang et al., 2017), et une prévision solaire PV horaire combinant un modèle ANN autorégressif avec un algorithme de regroupement k-means a été montrée par (Benmouiza & Cheknane, 2013), l'hybride GA et LSTM bi-directionnel (Bi-LSTM) pour la prévision à court terme de production PV sans données météorologiques a été discuté par (Zhen et al., 2021)et d'autres types de techniques hybrides de prévision ont été détaillés par (Rajagukguk et al., 2020). Toutefois, l'inconvénient majeur de ces modèles réside dans leur temps de calcul, puissance de calcul nécessaire accrus et la complexité de leur mise en œuvre par rapport aux modèles uniques.

- **L'approche d'ensemble de modèles**

L'approche d'ensemble de modèles, qui regroupe les prédictions de plusieurs modèles de prédiction de base indépendants, s'est avérée efficace pour la prévision de production PV (Dietterich, 2000). L'ensemble des ANN en est un exemple (Al-Dahidi et al., 2019). Les vecteurs de poids, paramètres internes de l'ANN, relient les nœuds d'entrée à la couche cachée, établissant ainsi l'effet des entrées sur la réponse et les biais de la couche neuronale cachée. La couche de sortie, reliée à la couche cachée par des poids, prédit la sortie du PV par le biais d'une fonction d'activation. Elle est traitée en parallèle par un groupe d'ANN, c'est-à-dire des modèles de prédiction de base ; la somme linéaire finale est la sortie de l'approche d'ensemble (pour l'ANN) à une heure donnée d'un jour donné. Ainsi, la précision des prévisions de l'ensemble peut être améliorée en générant de la diversité parmi ses modèles de prévision de base (Ren et al., 2015). Pour ce faire, il faut 1) en utilisant des méthodes de prédiction disparates (SVM, ANN, etc.) ; 2) en utilisant le même modèle de prédiction, mais avec des paramètres différents (par exemple, en faisant varier le nombre de neurones cachés et de couches neuronales) ; et 3) en formant les modèles individuels en utilisant des ensembles de données différents. Pour la dernière approche, des techniques telles que boosting (Kuzmiakova, 2017; Verbois et al., 2018), Bootstrapping AGGregatING (BAGGING) (El-Baz et al., 2018) et Adaboost (Bai et al., 2021) sont populaires.

Les modèles uniques manquent de robustesse et de flexibilité et ne peuvent donc pas réaliser de manière cohérente des prévisions de production PV précises, en particulier lors des

fluctuations météorologiques (Al-Dahidi et al., 2019). C'est le principe de l'approche d'ensemble qui a toujours montré une plus grande robustesse et précision de prévision.

Conclusion

La production photovoltaïque dépend de plusieurs paramètres météorologiques dont principalement la ressource solaire. Toutefois, la variabilité et l'intermittence de cette ressource ont fait de sa modélisation et sa prévision un défi de poids pour la recherche. À l'instar de l'ensemble des énergies renouvelables, le photovoltaïque est en plein essor. La capacité cumulée de la production mondiale d'énergie solaire ne cesse de progresser et donc sa nécessité de la maitriser.

Il existe de nombreux modèles de prévision de la production photovoltaïque, chacun ayant ses avantages et inconvénients. Toutefois, le but de cette thèse est d'élaborer un outil de prévision de production PV utilisable temporellement et spatialement sur l'ensemble du territoire réunionnais dont la topographie est marquée.

Les données à notre disposition conditionnent les méthodes de prévision que nous allons utiliser par la suite. Pour cela, le chapitre suivant va discuter dans un premier temps du contexte climatique et topographique particulier du territoire insulaire qu'est l'île de La Réunion. Dans un deuxième temps, le modèle WRF est décrit. Il est utilisé dans le cadre de cette thèse comme un générateur de données météorologiques brutes à très haute résolution (~ 1 km²) nécessaires à l'alimentation en entrée de nos modèles que nous développerons dans un autre temps. Ces données simulées seront décrites, vérifiées, analysées, confrontées et validées par rapports aux données météorologiques à notre disposition.

CHAPITRE 2

CONTEXTE DE L'ILE DE LA REUNION ET BASE DE DONNEES CLIMATIQUES ET ENERGETIQUES : CHOIX DE DONNEES ET VALIDATION

Introduction

Le chapitre précédent a traité de l'intérêt de la prévision de la production PV et un état de l'art des classes de techniques de prévision a été établi. Toutefois, le choix définitif du modèle dépend de plusieurs éléments : le contexte particulier de l'île de La Réunion et la base de données dont nous disposons.

Ce chapitre est dédié au contexte réunionnais et aux données nécessaires et utilisées dans la suite de ces travaux de recherche. La première partie est consacrée au contexte de l'île de La Réunion. La deuxième partie traite des bases de données d'observation directes et indirectes climatiques et énergétiques à notre disposition. Le choix a été de se tourner vers un modèle climatique régional. À ce titre, un modèle en particulier, le modèle WRF, a été sélectionné et ce choix constitue la troisième partie du chapitre. Elle traite de la méthode, de la préparation des données avec WRF prepropressing system, de la paramétrisation physique ainsi que des données en sortie exploitées dans la suite de la thèse. La dernière partie du chapitre est destinée à la validation de ces données de simulation avec la base de données climatiques au sol et satellitaire.

2.1 Contexte de l'île de La Réunion

La Réunion est une île localisée en 21°07'S, 55°32'E et fait partie de l'archipel des Mascareignes avec l'île Maurice et Rodrigues (Duncan, 1981). Toutes trois prennent naissance de l'activité d'un point chaud actif depuis plusieurs millions d'années, connu sous le nom de point chaud de La Réunion, également à l'origine des Trapps du Deccan (ou Trapps de Dekkan). Cette île présente plusieurs particularités : une topographie complexe et un contexte atmosphérique. L'île est soumise à des processus nuageux locaux et une ressource solaire très variable temporellement et spatialement.

2.1.1 Une topographie marquée

Bien que l'île de La Réunion soit connue pour abriter un des volcans les plus actifs au monde, le Piton de La Fournaise, dont la forte activité a participé au relief abrupt du territoire réunionnais, l'origine de l'île provient du Piton des Neiges et du volcan des Alizés (Lénat et al., 2012). La superficie de l'île est de 2512 km² et elle présente une topographie complexe, structurée autour de ses deux volcans culminant tous deux à plus de 2500 mètres d'altitude. Trois cirques : Cilaos, Salazie et Mafate résultent de l'effondrement des caldeiras du Piton des Neiges (figure 2.1).

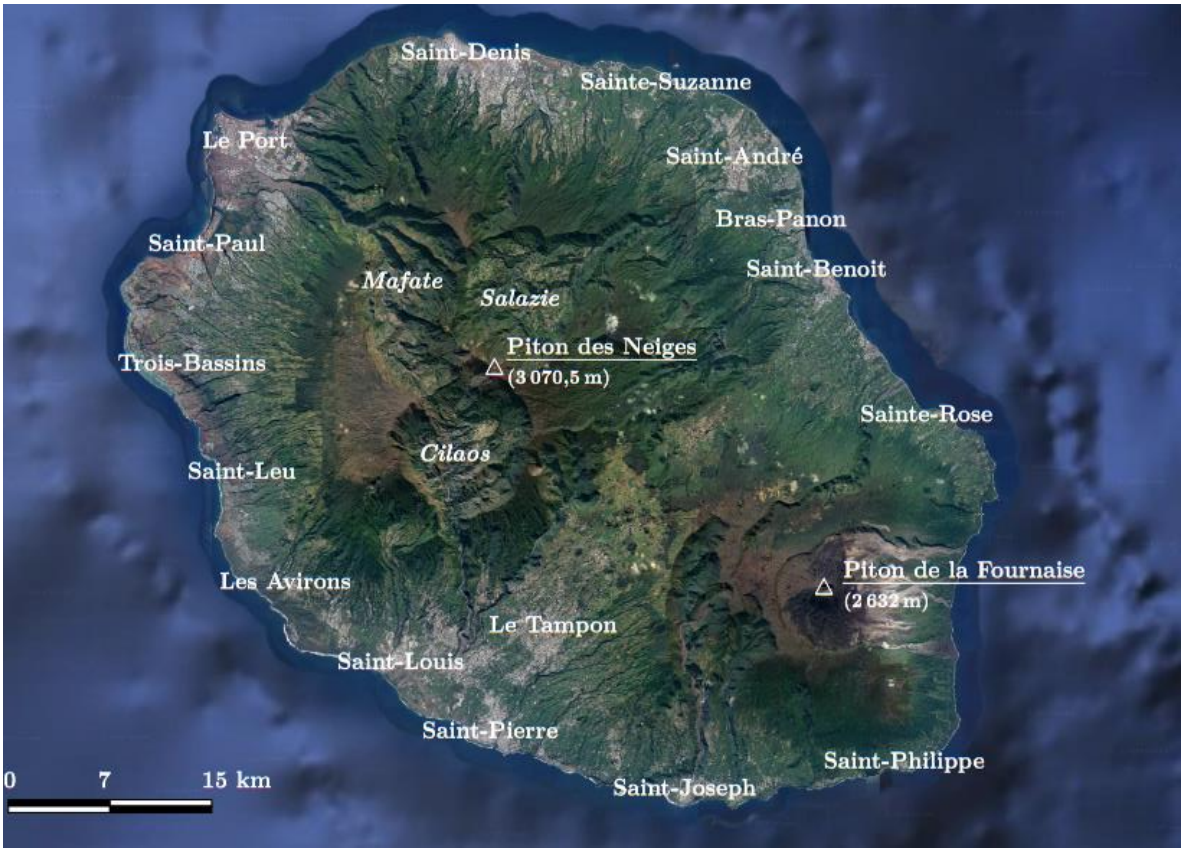


Figure 2.1 Vue satellitaire de La Réunion structurée autour de ses volcans et cirques. Source :(Mialhe, 2018)

2.1.2 Contexte atmosphérique

La Réunion est soumise à la circulation atmosphérique de l'anticyclone des Mascareignes. Ainsi, les conditions sont relativement stables en hiver car l'île est située dans la cellule de Hadley mais est soumise à des vents plus intenses qu'en été (Baldy et al., 1996). L'anticyclone des Mascareignes est plus éloigné de l'île en été et d'ailleurs la formation de cyclones intervient régulièrement sur l'île en période estivale. La Réunion est donc alimentée par des évènements extrêmes et des flux d'humidité provenant de l'Océan Indien causant une forte pluviométrie sur toute l'île (Morel et al., 2014; Pohl et al., 2016).

L'île représente un obstacle à la circulation atmosphérique. Des régimes d'écoulements orographiques sont notés. Selon l'intensité de l'écoulement de la circulation, il peut y avoir soit un régime de soulèvement (pour un écoulement intense) où la zone ouest de l'île est protégée ou un régime de contournement (pour un faible écoulement) exposant le flanc ouest de l'île à des vents intenses induits par effet Venturi (Lesouëf, 2010; Lesouëf et al., 2011). Des écoulements thermiques, des brises de terre et de mer, dynamisent aussi les circulations locales sur l'île. Les propriétés thermiques de la terre diffèrent de la mer créant ainsi un gradient de température. De plus, La Réunion est soumise aux alizés qui se combinent aux régimes de brises.

Deux processus d'ennuagement existent principalement sur l'île de La Réunion (Badosa et al., 2015) : le débordement nuageux et les nuages advectés. La couverture nuageuse sur l'île est la principale source de variabilité du rayonnement solaire reçu au sol.

2.1.3 Une ressource solaire variable

La localisation de l'île de La Réunion au sud-ouest de l'Océan Indien à quelques degrés au nord du Tropique du Capricorne, fait de la ressource solaire une source d'énergie privilégiée (Praene et al., 2012). Toutefois, sa forte variabilité principalement due à la couverture nuageuse sur l'île freine son intégration dans le mix énergétique local. Plusieurs études présentent déjà sur l'île le climat local (Jumaux et al., 2011). Des cartes de la ressource solaire en moyenne annuelle (figure 2.2) et saisonnières y sont présentées par interpolation spatiale. Ces données interpolées et cartographiées sont des données observables au sol in situ et ne représentent pas la complexité du relief de l'île

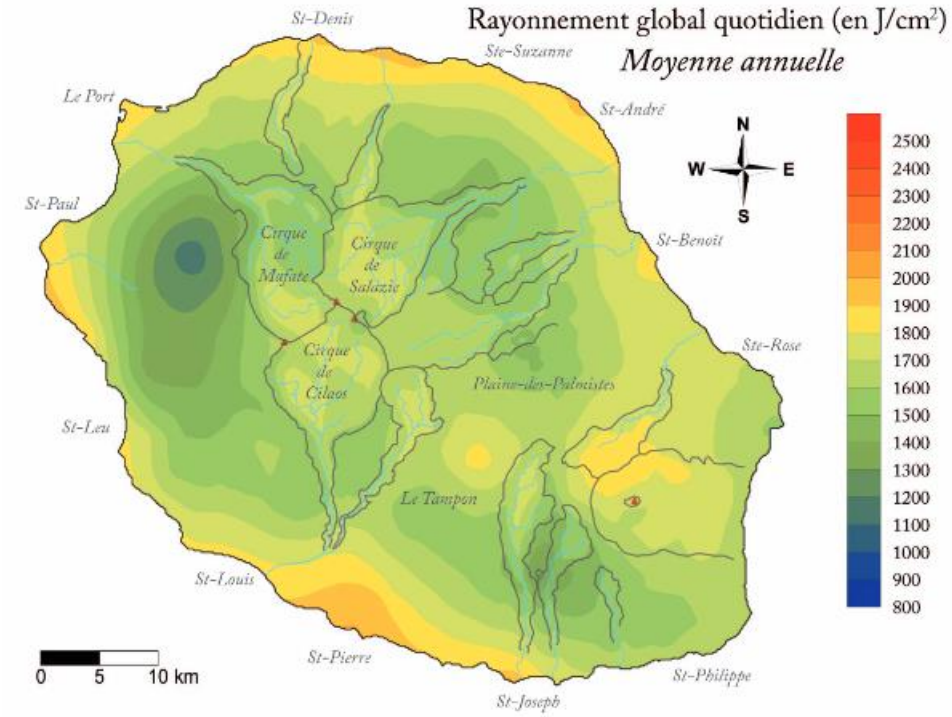


Figure 2.2 Carte du rayonnement global journalier moyenné annuellement reprise de (Jumaux et al., 2011)

(Li et al., 2015) proposent également une approche multi-fractale basée sur la transformation de Hilbert-Huang (HHT) de la ressource solaire locale en associant intermittence et turbulence. (Bessafi et al., 2018) proposent une interpolation spatiale sur des données satellitaires de résolution de ~ 5km² en tenant compte de la topographie de l'île et l'impact du rayonnement solaire direct et diffus et proposent une vision spatiale du cumul moyen annuel du rayonnement solaire au sol. Cependant, des écarts entre les données interpolées et les données observées in situ par Météo-France apparaissent dus à la forte variabilité de la couverture nuageuse sur l'île. D'autres études ont été effectuées afin d'obtenir des régimes de rayonnement solaire en classant les journées selon leur indice de ciel clair (Badosa et al., 2013) (Bessafi et al., 2015).

2.2 Choix de la base de données

Les données jouent un rôle central dans le choix, la conception et la paramétrisation d'algorithmes de prévision. Pour cela, les données à notre disposition au cours de ces travaux de thèse seront étudiées dans cette section. Dans un premier temps, les données climatiques d'observation directe au sol (Météo-France) et les données d'observation indirecte (produits satellitaires) seront présentées. La deuxième sous-section s'intéressera aux données énergétiques (PV) et météorologiques recueillies au niveau de la centrale PV de COREX située à la Possession.

2.2.1 Données climatiques

- **Données d'observations directes : données météorologiques au sol Météo-France**

Les données météorologiques au sol de Météo-France (MF) recueillies sur plusieurs stations au sol entre 2017 et 2018 sont étudiées et concernent 16 stations parmi l'ensemble du réseau au sol Météo-France réparties sur l'intégralité du territoire. Ces 16 stations météorologiques ont été sélectionnées, car elles répondaient à un certain nombre de critères concernant les paramètres météorologiques dont les données étaient indispensables aux travaux de thèse, à leur répartition géographique sur le territoire. Le tableau 2.1 détaille ces stations selon leur nom, leur longitude, leur latitude, leur altitude, leur indice de confiance. Cet indice de confiance prend les valeurs F0 à F5 (F pour flag) avec la plus faible valeur notée F0, où la confiance est très élevée (Bessafi et al., 2018). Toutefois la seule station à ce jour à utiliser cet indice de confiance est la station de l'aéroport de Gillot, rattaché au réseau Global Energy Balance Archive (GEBA) (Wild et al., 2017). Les 15 autres stations météorologiques sont notées F1. La classification des sites et capteurs est faite selon les critères de l'Organisation Météorologique Mondiale (OMM) (OMM, 2014). Ces stations ont en effet été choisies, car sur les années 2017 et 2018 (les années où ont été simulées les données météorologiques issues de WRF que nous présenterons dans la section suivante) possèdent des données, au pas de temps horaire, sur :

- La température (en °C)
- La vitesse du vent (en $m.s^{-1}$)
- Le rayonnement global solaire (en $J.cm^{-2}$)

Le rayonnement (en $J.cm^{-2}$) nécessite, par ailleurs, une conversion pour les comparer ensuite à l'irradiance solaire (en $W.m^{-2}$) d'autres bases de données.

De plus, faisant partie du réseau Météo-France, elles ont une maintenance régulière de leur part et ont une faible erreur d'incertitude sur les mesures. La figure 2.3 représente la carte de l'île de La Réunion et sa topographie indiquant au passage ses deux plus hauts sommets. Le premier est le volcan dormant du Piton des Neiges, culminant à 3 071m d'altitude (représenté par le triangle N) et le second est volcan du Piton de la Fournaise (représenté par le triangle F), culminant lui à 2 560m d'altitude. Ils sont séparés par un plateau de 1 500 m d'altitude, ce qui

fait de l'île un obstacle isolé et particulièrement proéminent aux circulations atmosphériques dans le sud-ouest de l'océan Indien (Mialhe et al., 2020). De plus, sur la figure 2.3 est représentée la localisation de ces 16 stations de Météo-France utilisées sur l'île.

Tableau 2.1 Tableau descriptif des stations météorologiques Météo-France provenant des fiches de poste éditées par Météo-France pour chacune de ses stations. *Toutes les stations ici présentes sont rattachées au réseau Météo-France. La classe des capteurs est celle de l'OMM. Pour les capteurs, 1 signifie "Première classe", avec une précision élevée et une faible incertitude et 2 signifie "Seconde classe", avec une moindre précision. Les pyranomètres sont des KIPP&ZONEN (noté K&N), les capteurs de vent, anémomètres Déolia 96 (Déolia 96) ou Capteur Ultrasonique GILL WindSonic 2D (GILL WS2), et les sondes thermométriques sont soit des platine PT100 T5312 (PT100 T5312) ou des sondes Pyrocontrole au platine T° air (Platine T°air)*

Nom de la station	**Longitude (°E)**	**Latitude (°S)**	**Altitude (m)**	**Capteurs**			**Classe des capteurs**
BELLECOMBE-JACOB	55.687	21.217	2245	K&Z CM6B	Déolia 96	PT100 T5312	1
BELLEVUE BRAS-PANON	55.622	21.005	480	K&Z CM6B	Déolia 96	Platine T°air	2
CILAOS	55.472	21.134	1197	K&Z CM6B	Gill WS2	PT100 T5312	1
COLIMACONS	55.305	21.13	798	K&Z CM6B	Gill WS2	Platine T°air	1
GILLOT-AEROPORT	55.528	20.892	8	K&Z CM6B	Déolia 96	Platine T°air	1
GROS PITON SAINTE-ROSE	55.828	21.179	181	K&Z CM6B	Gill WS2	Platine T°air	1
LE BARIL	55.732	21.359	115	K&Z CM6B	Gill WS2	PT100 T5312	1
LE PORT	55.282	20.946	9	K&Z CM6B	Gill WS2	Platine T°air	1
PETITE-FRANCE	55.342	21.045	1200	K&Z CM6B	Gill WS2	Platine T°air	1
PIERREFONDS-AEROPORT	55.425	21.320	21	K&Z CM6B	Déolia 96	Platine T°air	2
PITON-MAIDO	55.381	21.076	2150	K&Z CM6B	Gill WS2	PT100 T5312	1
PLAINE DES CAFRES	55.572	21.209	1560	K&Z CM6B	Déolia 96	Platine T°air	1
PLAINE DES PALMISTES	55.627	21.136	1032	K&Z CM6B	Gill WS2	Platine T°air	1
POINTE DES TROIS-BASSINS	55.248	21.105	5	K&Z CM6B	Déolia 96	Platine T°air	1
PONT-MATHURIN	55.380	21.265	19	K&Z CM6B	Gill WS2	Platine T°air	1
SAINT-BENOIT	55.719	21.058	43	K&Z CM6B	Gill WS2	Platine T°air	1

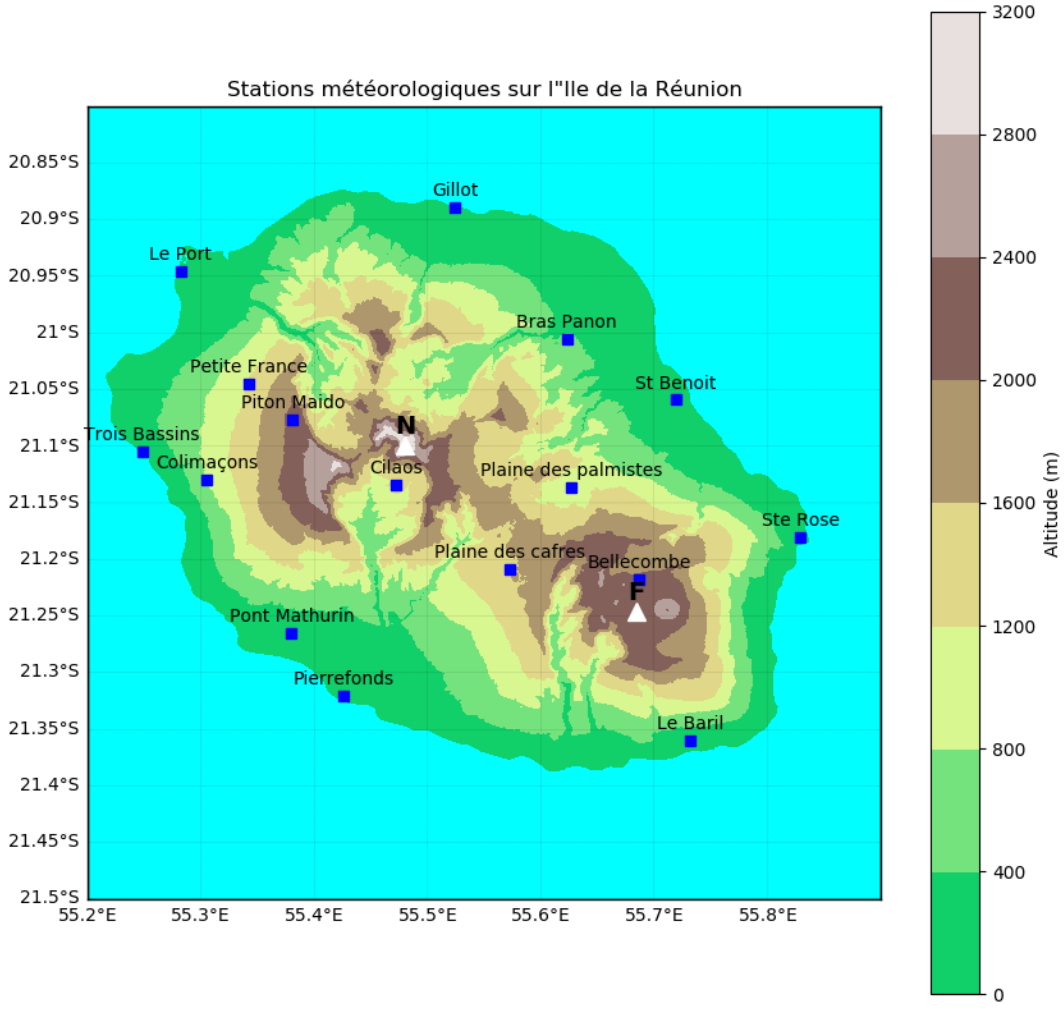


Figure 2.3 Position et altitude des stations météorologiques de Météo-France retenues sur la carte de l'île de La Réunion

Sur la figure 2.4 sont représentées les données de température et de GHI pour toutes les stations Météo-France répertoriées et dont les données ont été récupérées sur la période 2017. Elle permet de mettre en avant la variation liée à la saisonnalité (l'été et l'hiver austral) de ces paramètres météorologiques.

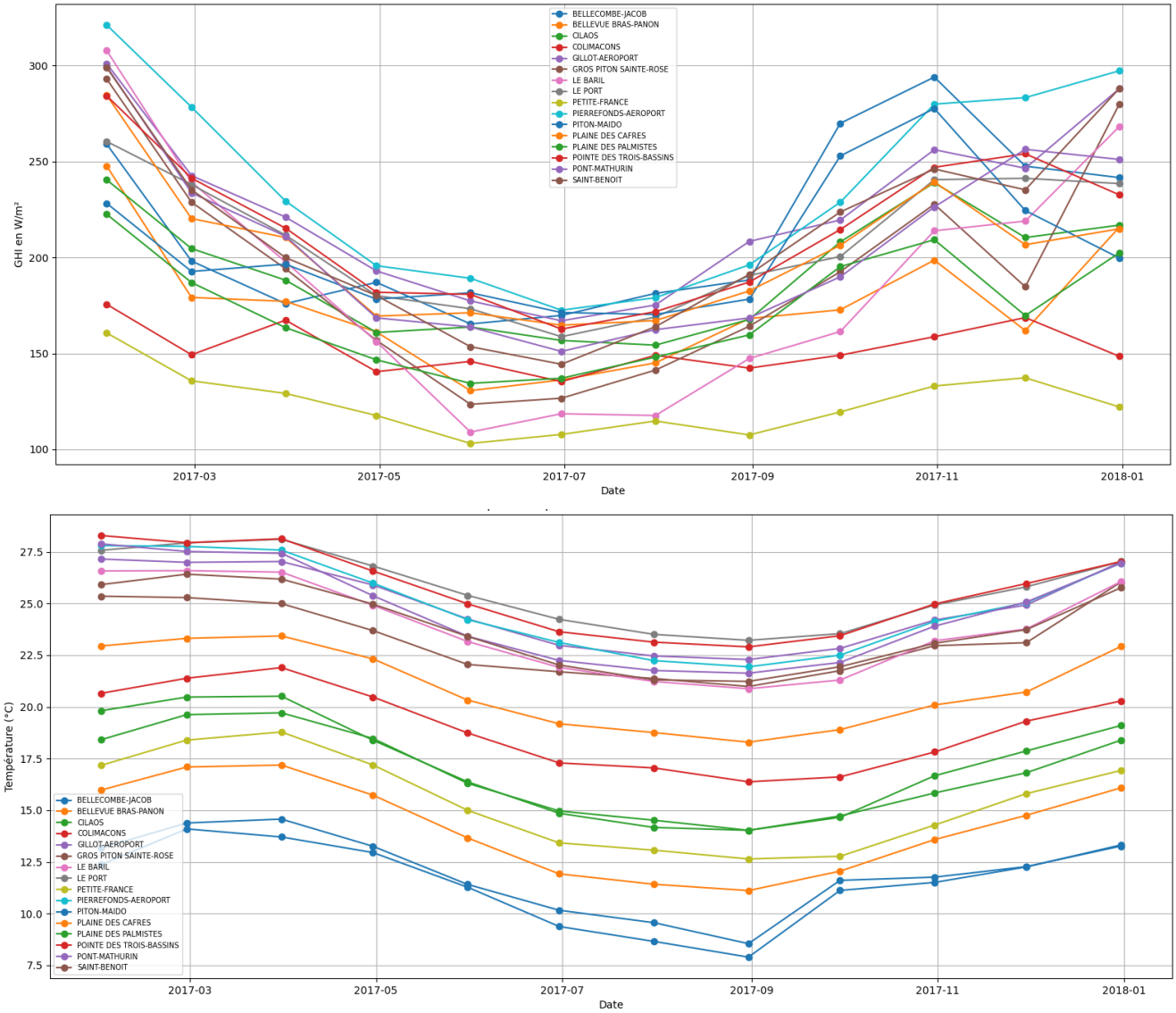


Figure 2.4 Irradiance solaire (en W.m^{-2}) et température (en °C) moyennées mensuelles pour les 16 stations Météo-France détaillées dans le tableau 2.1 sur une période s'étalant de février 2017 à janvier 2018

Toutefois, pour avoir une image « climatique » précise en tout point de l'île, nous ne pouvons pas nous contenter de 16 points sur la carte, qui plus est pour un territoire dont la topographie est marquée, rendant complexe toute tentative d'interpolation spatiale (Bessafi et al., 2018).

- **Données d'observations indirectes : données satellitaires**

Les programmes multi-satellites à long terme fournissent une base de données croissante d'observations qui constituent des contributions essentielles à la surveillance du climat. Les satellites météorologiques Meteosat fournissent des données pour la surveillance du climat en Europe et en Afrique depuis 1978, constituant ainsi l'une des plus longues séries chronologiques de données climatiques collectées par satellite au monde. EUMETSAT traite

ces observations satellitaires afin de développer des enregistrements de données climatiques, qui sont des séries temporelles de mesures satellitaires suffisamment longues et cohérentes pour déterminer la variabilité et le changement climatiques. Il fournit des données et des informations climatiques à ses États membres, au *Copernicus Climate Change Service* et aux utilisateurs du monde entier, y compris la communauté scientifique mondiale.

Le bilan radiatif à la surface de la Terre est un paramètre clé pour la surveillance et l'analyse du climat. Les données satellitaires permettent de déterminer le bilan radiatif avec une haute résolution spatiale et temporelle et offrent une large couverture régionale par la combinaison de différents satellites. Le CM SAF a traité un enregistrement continu de 35 ans (1983-2017) de données climatiques sur le rayonnement de surface, basé sur les observations des satellites Meteosat de première et deuxième génération : Surface Solar Radiation Data record - Heliosat Version 2.1 (SARAH-2.1) (Pfeifroth et al., 2019). Les données sont disponibles sous forme de moyennes mensuelles, journalières et de données instantanées de 30 minutes sur une grille régulière de latitude/longitude avec une résolution spatiale de 0,05° X 0,05°. L'intérêt de cette version par rapport Heliosat basé sur Meteosat-East (SARAH-E) est que les données de rayonnement s'étalent jusqu'au 31 décembre 2017 contre 2015 pour SARAH-E. Or les données WRF (discuté dans la section suivante) ont été simulées sur les années 2017 et 2018.

La limite sur ces données Météosat SARAH-2.1 réside dans le fait qu'elles ne concernent que le rayonnement solaire et qu'elles ne s'étendent pas au-delà de 2017. De ce fait, les données SARAH-2.1 n'ont été recueillies que sur l'année 2017. Pour information, la base de données SARAH-3 est disponible depuis le 05 mai 2023 et couvre une période s'étendant du 1er janvier 1983 au 31 décembre 2020 (Pfeifroth, Uwe et al., 2023). Elles seront comparées aux données WRF afin de les valider dans la dernière section de ce chapitre. La figure 2.5 représente les données d'irradiance solaire moyennées pour le mois de décembre 2017 sur l'île de La Réunion.

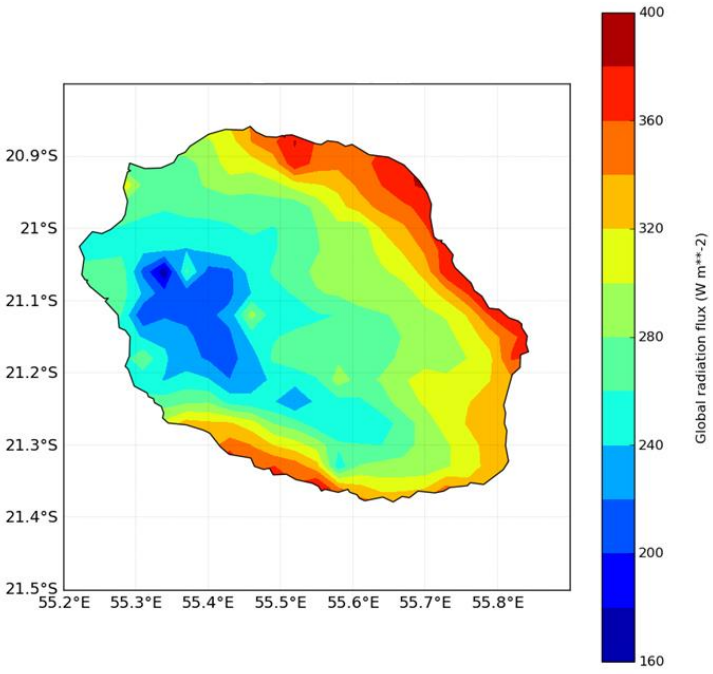


Figure 2.5 Carte d'irradiance solaire moyennée mensuelle (décembre 2017) de l'île de la Réunion issues des données SARAH-2.1

2.2.2 Données énergétiques

- **Données COREX**

Les données COREX sont des données énergétiques et météorologiques. Quatre paramètres ont été relevés : PV, irradiance solaire, température extérieure et vitesse de vent. Ces données ont été obtenues au niveau d'un bâtiment COREX situé à La Possession sur la côte ouest de l'île de La Réunion, près de la ville du Port (représentée sur la figure 2.3). La conception PV est un bâtiment monté sur le système connecté au réseau où les modules qui composent la centrale PV sont inclinés à un angle de 21°, comme la latitude de l'île. Les cellules PV polycristallines de 180W chacune sont équipées de pyranomètres, de sondes thermométriques et d'anémomètres. Pour cette étude, les données moyennes horaires et journalières sont récupérées sur les sept années de mesures entre 2011 et 2018. Ces données ont été principalement utilisées pour l'article de journal constituant le chapitre suivant (Fanchette et al., 2020) permettant de prévoir la production solaire PV en utilisant la méthode de cointégration de Johansen détaillée dans le chapitre suivant . La figure 2.6 est une vue d'en haut de la centrale de production PV, source de nos données météorologiques et PV ayant permis la mise en place de ce modèle de conversion PV.

Figure 2.6 Centrale PV COREX de la Possession vu d'en haut

La limite majeure des données PV de manière générale est la difficulté de les obtenir. Ici, nous avons récolté des données énergétiques sur une centrale PV (sur un point de l'île) mais afin de mener à bien ces travaux de recherche, des données météorologiques et surtout énergétiques étaient nécessaires sur l'ensemble de l'île. L'idée était alors d'utiliser un modèle régional de climat afin de simuler avec une très haute résolution des données sur une période donnée (2017-2018) et d'estimer à la même résolution à partir de ces données météorologiques des données PV. Pour cela, le modèle de cointégration de Johansen, qui sera détaillé dans le chapitre suivant, est employé. Dans la section suivante sont mis en avant les modèles régionaux de climat et plus particulièrement le modèle WRF retenu pour la suite de la thèse.

2.3 Modèles climatiques

Les modèles climatiques actuels sont une représentation numérique de la planète et des phénomènes physiques ou biogéochimiques et leurs couplages à sa surface. La géographie du globe est composée de cases appelées mailles et les interactions entre ces mailles sont représentées par des équations mathématiques. Les modèles climatiques cherchent à reproduire le plus précisément possible la réalité en représentant les forces induisant les mouvements terrestres, atmosphériques et océaniques (Dufresne & Salas y Mélia, 2017).

2.3.1 Les modèles de circulation générale

En premier lieu, la modélisation du climat s'effectue par des modèles dits de circulation générale (General Circulation Models, GCM). Ces modèles représentent les processus physiques de l'atmosphère, des océans et des surfaces terrestres. Ils constituent un outil indispensable à la compréhension du climat passé et présent (Randall, 2000). De ce fait, ils sont aussi capables d'estimer une évolution possible du climat dans le futur à l'échelle du globe. Parmi ces modèles de circulation générale, deux catégories se distinguent :

- Les modèles de circulation générale atmosphérique (AGCM) qui ne prennent en compte que l'atmosphère. Ces modèles sont utilisés pour les prévisions météorologiques.
- Les modèles de circulation générale atmosphérique et océanique (AOGCM) qui prennent en plus de l'atmosphère et l'océan. Ces modèles sont plutôt utilisés en climatologie puisqu'ils tiennent compte des interactions entre l'océan et l'atmosphère dans les prévisions.

Ces GCM, grâce à la montée en puissance des capacités de calcul informatique, atteignent désormais des résolutions horizontales d'une à plusieurs centaines de kilomètres. Les GCM sont fiables à l'échelle continentale et leur résolution permettant de reproduire la variabilité du climat aux grandes échelles. Toutefois, ils ne peuvent tenir compte de l'hétérogénéité à petite échelle de la variabilité du climat et des changements possibles. (X. Zhang et al., 2016) indiquent que les GCM ont tendance à sous-estimer les quantités du nuages à basse et moyenne altitudes qui reflètent le rayonnement solaire tandis que (P. li et al., 2013) mettent en avant une surestimation du rayonnement solaire de surface et cette surestimation est généralement plus importante dans les régions tropicales. Ces hétérogénéités sont indispensables quant à la compréhension des impacts potentiels sur l'hydrologie, la production végétale, la prévision d'énergie solaire, la température au sol, etc. à des échelles régionales (10 à 50 km) ou locales (1 à 5 km). Ainsi, ces GCM restent insuffisants pour la représentation des processus de fine échelle et la modélisation du climat et de ses impacts à l'échelle régionale ou local (Tang et al., 2019).

2.3.2 La descente d'échelle

Il existe ainsi des méthodes dites de descente d'échelle permettant d'affiner les simulations climatiques issues des GCM en des résolutions plus fines, plus pertinentes pour mesurer les différents impacts plus localement. Parmi elles, se trouvent deux catégories : l'approche statistique, utilisant des approches plus empiriques et l'approche dynamique, utilisant des modèles climatiques régionaux. La descente d'échelle statistique consiste à établir

des relations statistiques entre les variables simulées à grande échelle par les GCM et les données observées des paramètres à représenter à des échelles plus fines (Wilby and Wigley, 1997). Les hétérogénéités (topographie, etc.) sont prises en compte tout en appliquant ces relations au domaine modélisé et à la période temporelle étudiée. Cependant, la modélisation par descente d'échelle statistique est purement statistique et ne considère nullement les processus physiques, mais permet d'obtenir des résultats proches des observations à des résolutions très fines. Elle est néanmoins préférable lorsque des estimations de variables spécifiques, en particulier à des endroits précis, sont recherchées pour alimenter des modèles sectoriels comme les modèles en hydrologie (Wood et al., 2004). La descente d'échelle dynamique permet d'augmenter la résolution uniquement à une région précise du globe. Elle est recommandée lorsque ces caractéristiques jouent un rôle important dans le climat régional.

2.3.3 Les modèles climatiques régionaux

Les modèles régionaux de climat (MRC) ont été essentiellement développés dans le but de réduire l'échelle des champs climatiques produits par les modèles climatiques mondiaux à résolution grossière, fournissant ainsi des informations à des échelles de grille fines, sous-GCM, plus adaptées à l'étude des phénomènes régionaux et à l'application aux évaluations de la vulnérabilité, des impacts et de l'adaptation (Giorgi, 2019). La stratégie sous-jacente à cette technique de descente d'échelle est que le GCM peut décrire la réponse de la circulation globale aux forçages à grande échelle, tels que ceux dus aux GES ou aux variations du rayonnement solaire, tandis que le MRC peut affiner spatialement et temporellement ces informations à grande échelle en tenant compte des effets des forçages et des processus à l'échelle de la grille sous-GMC, tels que ceux dus à la topographie complexe, aux côtes, aux masses d'eau intérieures et à la distribution de la couverture terrestre, ou aux processus dynamiques se produisant à méso-échelle.

Les méthodes de descente d'échelle dynamique se sont améliorées et ont été largement appliquées dans diverses régions du monde (Dickinson et al., 1989 ; Giorgi et Bates, 1989 ; Castro et al., 2005 ; Tapiador et al., 2020). La dérivation d'informations climatiques à échelle fine par la descente d'échelle dynamique est basée sur l'hypothèse que le climat local est conditionné par les interactions entre les variables (circulation, température, humidité, etc.) et les particularités locales (masses d'eau intérieures, montagnes, propriétés de la surface terrestre, etc.). Il est possible de modéliser ces interactions et d'établir les relations entre le climat local actuel et les conditions atmosphériques par le biais du processus de descente d'échelle (Morel et al., 2014; Pohl et al. 2016). Il est important de comprendre que ce processus ajoute des informations au résultat grossier du GCM, de sorte qu'elles soient plus conformes à une échelle plus fine, en saisissant les contrastes et les inhomogénéités des sous-échelles de la grille. La figure 2.7 illustre ce modèle de circulation générale. Le développement des supercalculateurs ultra performants permet aujourd'hui une descente d'échelle de modèles de climat régionaux allant jusqu'à 2,5 km de résolution horizontale (Daniel, 2017) avec prise en compte des paysages urbains.

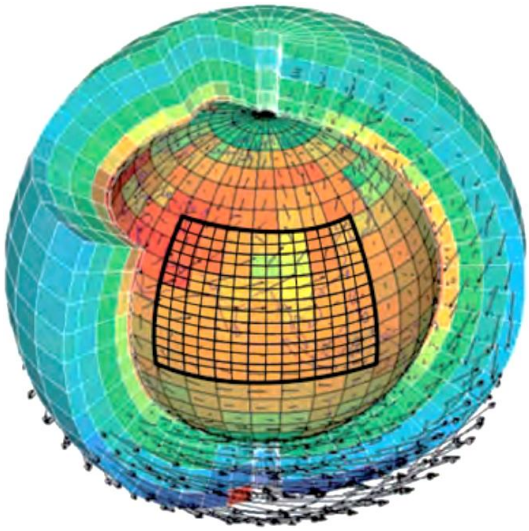


Figure 2.7 Modèle de circulation générale à résolution grossière et modèle régional de climat à haute résolution (grille noire). Source : (Cretat et al., 2011)

2.4 Modèle régional de climat WRF

Dans cette section, nous présentons les caractéristiques du modèle méso-échelle Weather Research and Forecasting (WRF) utilisé par la suite comme générateur de données météorologiques, du noyau dynamique et des paramètres physiques employés ainsi que du protocole expérimental de la simulation.

2.4.1 Choix de la méthode

L'université de La Réunion ne dispose pas d'un modèle particulier régional de climat opérationnel. Il en existe, toutefois, plusieurs susceptibles de nous permettre de générer des données climatiques avec une très haute résolution sur l'île de la Réunion. Cependant, le modèle WRF est historiquement un modèle largement utilisé au sein du laboratoire Energy^{LAB} (Morel et al., 2014; Pohl et al., 2016) afin de modéliser à très haute résolution un territoire comme l'île de La Réunion dont la topographie est marquée et complexe. De plus, il est libre d'accès et possède une large communauté scientifique qui ne cesse de croitre. Enfin, les données d'initiation permettent la préparation des données. Nous allons présenter dans le paragraphe suivant le modèle régional de climat WRF.

2.4.2 Présentation générique du modèle WRF

Le modèle WRF est un modèle climatique régional (Skamarock et al., 2008, 2019), utilisé, comme son nom l'indique, à la fois pour la recherche et la prévision opérationnelle du temps (ou Numerical Weather Prediction (NWP)). Bien qu'il ait été conçu par le National Center for Atmospheric Research (NCAR), la National Oceanic and Atmospheric Administration, l'Air Force Weather Agency (AFWA), le Naval Research Laboratory, l'université d'Oklahoma et la Federal Aviation Administration (FAA) pour prévoir et modéliser la circulation atmosphérique, le modèle WRF est devenu un véritable modèle communautaire par son développement à long

terme grâce à l'intérêt et aux contributions d'utilisateurs mondiaux (Powers et al., 2017). Il résout explicitement les équations de la dynamique qui assure la conservation et implémente les principaux processus physiques en lien avec le climat.

2.4.3 Noyau dynamique et composantes physiques

Ces dernières années, l'utilisation du modèle régional climatique WRF pour la recherche sur le climat régional a connu un essor considérable (Soares de Araujo, 2020, 2021; Zhu & Ooka, 2023). Le point fort du WRF est de résoudre les processus atmosphériques et de surface terrestre à plus petite échelle mieux que les modèles mondiaux traditionnellement utilisés pour les projections climatiques. En utilisant une formulation non-hydrostatique des équations de la mécanique des fluides et de la thermodynamique, les équations de la dynamique à haute résolution sont résolues. La dynamique du WRF est représentée dans ses solveurs de flux de fluides atmosphériques, ou noyaux. WRF propose deux noyaux :

- Le noyau Non-Hydrostatic Meso-scale Model (NMM) (Janjić et al. 2001 ; Janjić 2003) concerne la prévision météorologique opérationnelle et a été developpé par NOAA/NCEP (National Oceanic and Atmospheric Administration/National Centers for Environmental Prediction).
- Le noyau Advanced Research WRF (ARW), développé par le NCAR, est destiné à la recherche climatique et présente l'état de l'art de la modélisation de la dynamique de l'atmosphère à fine échelle.

La version WRF/ARW est une plate-forme de recherche sur la simulation numérique régionale du climat, permettant diverses configurations, allant de cas idéalisés en 2D ou 3D au mode dit « real » (configuration généralement utilisée par la descente d'échelle dynamique climatique) qui est alimentée aux limites par des données simulées à larges échelles (GCM, réanalyses, etc.) et/ou par des données provenant d'observations satellitaires ou au sol. Cette version WRF/ARW est représentée sur la figure 2.8.

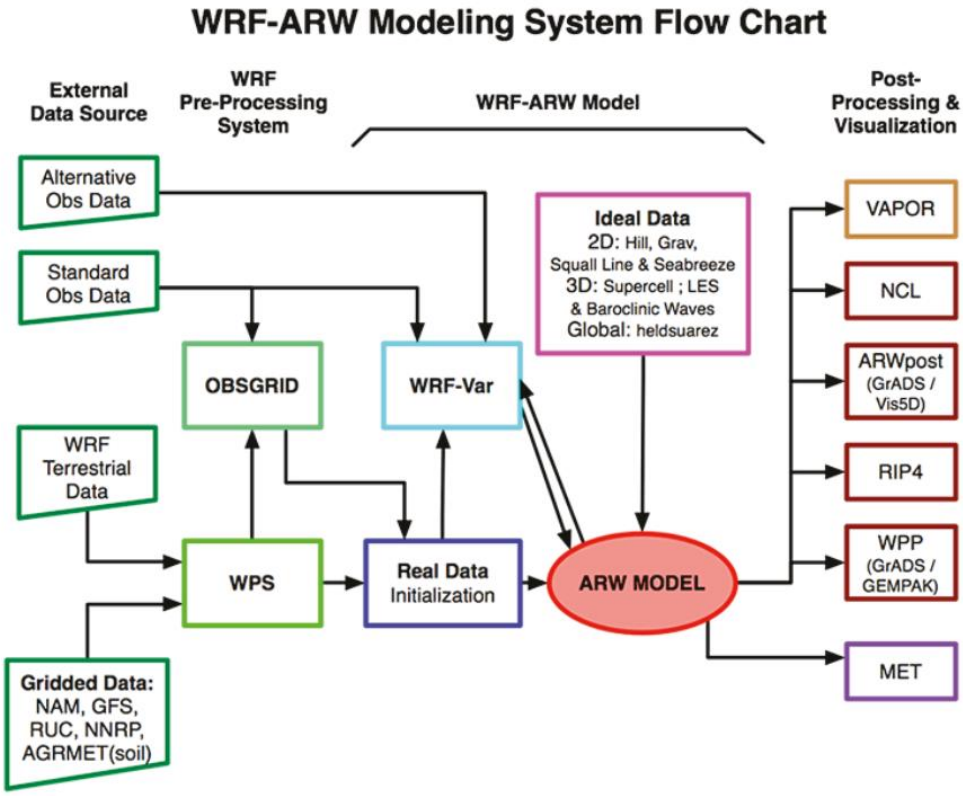


Figure 2.8 Organigramme de WRF/ARW tiré de (W. Wang et al., 2012)

Comme expliqué plus haut, une des spécificités des MRC est la simulation du climat sur une aire géographique limitée de la surface du globe terrestre. Pour cela, la désagrégation du signal climatique se fait par descente d'échelle. Le modèle utilise des emboîtements successifs de domaines dont la résolution entre chaque domaine varie d'un facteur 2 à 5. Cette descente d'échelle implique des échanges d'informations entre les deux domaines. On parle de « one-way » ou « two-way » nesting. Dans le premier cas, les informations échangées ne se font que dans un seul sens, du domaine à large maille vers le domaine à maille plus fine ; les conditions aux limites des domaines emboîtés sont déterminées par le domaine parent. À contrario, si les domaines imbriqués (domaines enfants) peuvent influencer les domaines parents en retour, alors il s'agit d'un « two-way » nesting.

Dans cette thèse, les simulations réalisées sont produites à partir du cœur ARW, car il est, en effet, adapté pour des volumes conséquents de données en mode real. La thèse nécessitant de désagréger des données à très haute résolution, le noyau ARW est configuré en « one-way » nesting et quatre domaines imbriqués de 25km X 25km, de 5km X 5km et enfin de 1 km X 1 km sont simulés pour affiner à une échelle très fine la résolution de la région voulue.

2.4.4 Préparation des données avec WPS

Les simulations WRF commencent avec le système de prétraitement de WRF appelé WRF Preprocessing System (WPS). Le WPS, et plus précisément le WPS 4.0.3 utilisé dans le cadre de cette thèse, est composé de trois étapes successives et indispensables à la préparation

des données, « geogrid », « ungrib » et « metgrid ». Elles permettent de définir la période et l'aire géographique, les données d'entrée statistiques et dynamiques à utiliser ainsi qu'à effectuer l'interpolation des données à la maille de simulation. Enfin, les champs de saisie sont placés sur les niveaux verticaux du modèle et des conditions aux limites latérales sont générées. WRF est alors prêt à fonctionner.

2.4.5 Protocole expérimental de simulation

- **Domaines d'étude**

La simulation est réalisée avec la version 4.1 du modèle WRF en mode non-hydrostatique sur le domaine 3 centré au point de coordonnées géographiques 21° 20 S, 55° 29 E qui couvre l'île de La Réunion dont la superficie est de 2512 km² et le domaine 4 centré au point de coordonnées géographiques 20° 20 S, 57° 33 E qui, elle, couvre l'île Maurice dont la superficie est de 2040 km². Ils sont désagrégés à partir du modèle « ERA5 3-hourly model data » de 37 niveaux verticaux de pression de 1000 à 1hPa et ont une résolution spatiale horizontale de 1 km. La résolution verticale est configurée avec une forte densité des niveaux à proximité du sol. La résolution temporelle des sorties du modèle WRF est configurée au pas de temps horaire afin d'avoir les plus d'informations possibles sur les variables tout en limitant le volume de fichiers en sortie. Ainsi, les données « instantanées » ou « accumulées », selon les variables, sont recueillies toutes les heures. Les simulations du modèle WRF sont faites sur une période de deux ans qui s'étend du 1er Janvier 2017 au 31 Décembre 2018.

La résolution spatiale la plus fine est de 1 km × 1 km et est le résultat d'une descente d'échelle qui a mis en jeu 4 domaines imbriqués. En effet, la désagrégation avec le modèle WRF exige le respect d'un ratio de 3 ou 5 entre la taille des pixels des grilles de deux domaines consécutifs imbriqués. De ce fait, dans ce travail, afin d'étudier à une résolution très fine l'île de La Réunion, quatre domaines imbriqués de résolution horizontale respective de 25, 5 et 1 km, soit un ratio de 5, ont été utilisés. La figure 2.9 présente l'étendue des domaines imbriqués.

- **Le domaine 1** possède une résolution horizontale d'environ 0.25° × 0.22° (soit environ 25 km × 25 km) et couvre une aire géographique qui comprend l'île de La Réunion, l'île Maurice et Madagascar. La largeur du domaine est importante afin d'intégrer une grande partie de l'Océan indien dans les calculs. En prenant en compte un si large domaine 1, nous obtenons une représentation le plus conforme possible des systèmes météorologiques au niveau des domaines 3 et 4, c'est-à-dire l'île de La Réunion et l'île Maurice.
- **Le domaine 2**, a quant à lui une résolution spatiale d'environ 0.048° × 0.045° (soit environ 5 km × 5 km) et couvre La Réunion et l'île Maurice. Une grande distance entre le domaine 1 et le domaine 2 est nécessaire afin que la dynamique de l'échelle large développe des traits du climat d'échelle plus fine.
- **Le domaine 3**, avec une résolution horizontale de 1 km x 1 km, couvre La Réunion.
- **Le domaine 4**, comme le domaine 3, a une résolution horizontale de 1 km × 1 km et

couvre l'île Maurice mais ces données, simulées et disponibles, ne sont pas utilisées dans cette thèse mais peuvent faire l'objet d'une autre étude (voir la section Perspectives).

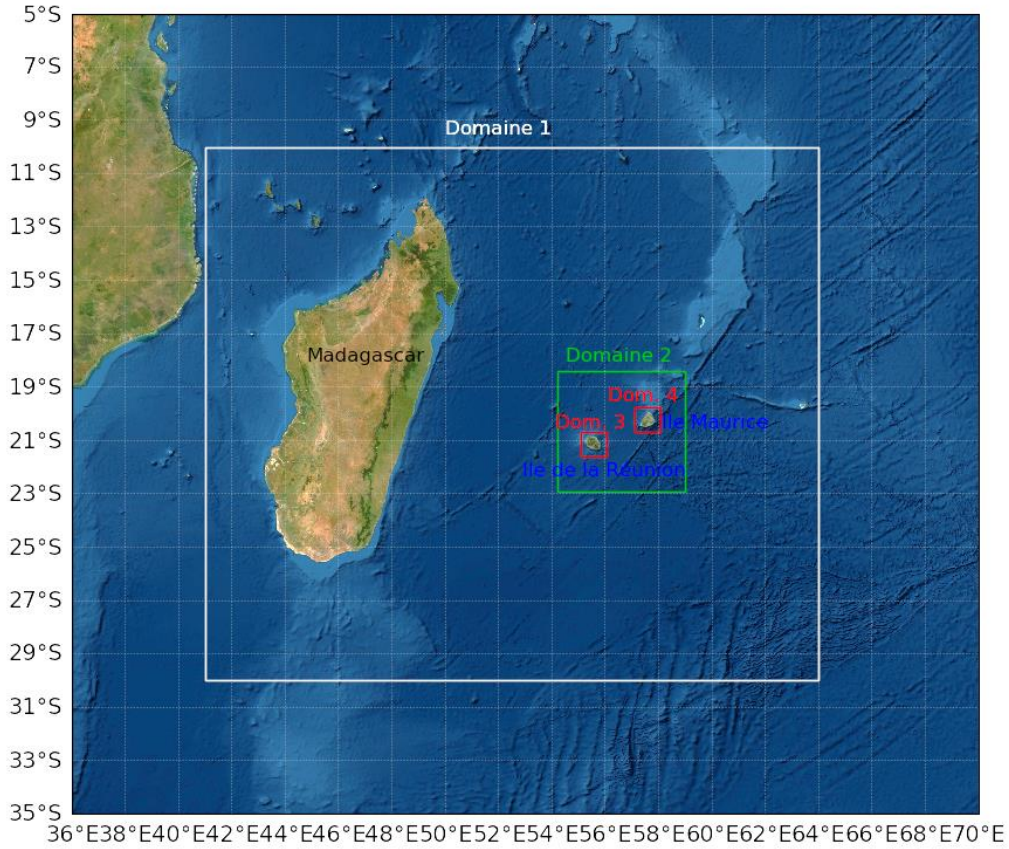


Figure 2.9 Position des domaines imbriqués dans l'Océan Indien.

- **Paramétrisation physique**

Le modèle WRF propose plusieurs configurations pour les schémas physiques implémentés. La physique du modèle comprend la convection cumulée, la microphysique des processus nuageux et des précipitations, le rayonnement à courtes et grandes longueurs d'ondes (SW et LW), la turbulence et la diffusion, les modules de couche limite planétaire (PBL), la couche de surface et la couche de sol. Ces multiples options de physique permettent de traiter avec souplesse les processus physiques Il existe potentiellement plusieurs milliers de combinaisons différentes pour l'ensemble des options physiques de base à disposition. Le choix s'avère donc une étape cruciale pour le bon déroulement de la simulation numérique du climat. Toutefois, cette thèse n'a pas vocation à trouver le meilleur compromis concernant les options physiques. La paramétrisation des principaux processus physiques implémentés s'est basée sur plusieurs recherches effectuées dans la zone d'Afrique austral et du sud-ouest de l'Océan Indien (Crétat et al., 2011, 2012; Crétat & Pohl, 2012). Les principaux processus physiques et leurs options retenues sont énumérés dans le tableau 2.2.

Tableau 2.2 Tableau récapitulatif des options sélectionnées pour les principaux schémas physiques

Principaux processus physiques	Option choisie	Référence
Radiation de courte longueur d'onde	Dudhia Shortwave Scheme	(Dudhia, 1988)
Radiation de grande longueur d'onde	RRTM Longwave Scheme	(Mlawer et al., 1997)
Microphysique des nuages	WRF Single moment 6-class Graupel Scheme	(Hong et al., 2006)
Schéma de surface terrestre	Unified Noah Land Surface Model	(Tewari et al., 2004)
Physique des cumulus	Kain–Fritsch Scheme (Domaine 1 et 2) et pas de cumulus (Domaine 3 et 4)	(Kain, 2004)
Couche limite planétaire	Yonsei University Scheme	(Hong et al., 2006)

2.4.6 Données en sorties de simulation et ressources informatiques

Afin de mieux comprendre les performances du modèle dans cette région particulière qu'est le Sud-Ouest de l'océan indien (SOOI), un ensemble de simulations WRF forcées par des réanalyses ERA-5 avec une configuration physique précise du modèle.

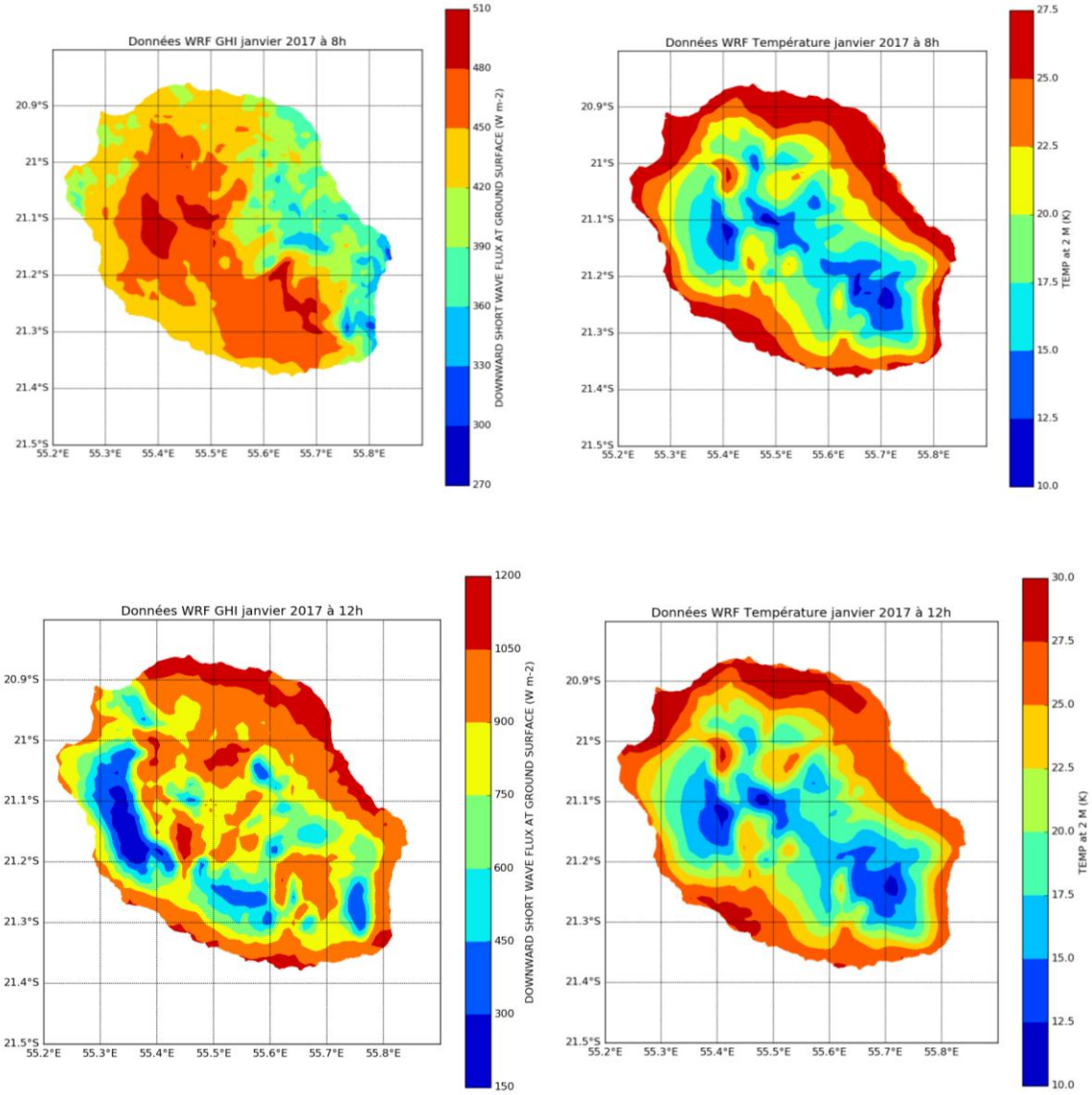
Les simulations ont été lancées sur le nouveau Centre de Calcul de l'Université de la Réunion (CCUR) et ont duré au total de 2700 heures sur 32 CPUs en parallèle (soit 86400 heure CPU). Ces simulations ont produit environ 1 Téra octet (To) de données pour le domaine 3 et 4. Le volume et le coût de calcul très conséquents ont conduit au choix de conserver les données en sortie à un pas de temps horaire pour chacune des variables considérées sur les années étudiées c'est-à-dire 2017 et 2018. Les données obtenues en sortie sont présentées dans le paragraphe suivant.

2.4.7 Données simulées WRF

Plusieurs variables climatiques sont analysées : le rayonnement solaire descendant en surface (SWDOWN), la température de l'air en surface, plus précisément à deux mètres du sol (T2), l'humidité relative (QVAPOR) et la vitesse du vent avec notamment les trois composantes x/y/z de la vitesse du vent (U/V/W). Ces données sont regroupées en détail dans le tableau 2.3. Le modèle WRF a été considéré dans cette thèse comme un « générateur de données météorologiques » sur toute l'île de La Réunion avec une très fine résolution.

Ces données simulées et recueillies sur les années 2017 et 2018 constitueront la base des données utilisée dans la suite de la thèse. Mais avant cela, une étude de ces données est

nécessaire et la confrontation de ces données avec d'autres jeux de données météorologiques est indispensable. Il s'agit de données météorologiques horaires. Des données plus précises (de l'ordre la dizaine de minutes) avaient été envisagées, toutefois, la taille des données générées a été jugée trop grande et le format horaire a été retenu. Notons que la taille des données horaires correspond à environ 110 Gigaoctets (Go) de données par année simulée. La figure 2.10 représente les sorties des données météorologiques WRF moyennées mensuelles (ici la température à deux mètres du sol et l'irradiance solaire) sur toute l'île de la Réunion à différentes heures.



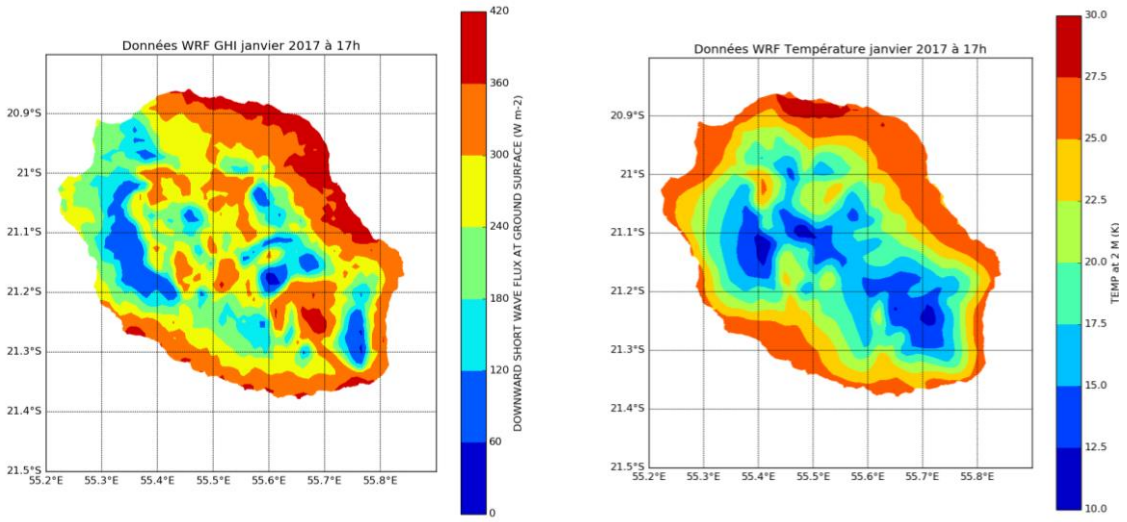


Figure 2.10 Cartes en moyenne mensuelle de l'île de la Réunion représentant les données WRF d'irradiance solaire et de température (Janvier 2017) à 8h, 12h et 17h

Tableau 2.3 Liste des variables récupérées au pas de temps horaire en sortie de simulation

Nom de la variable dans WRF	Nom usuel de la variable (en anglais)	Nom usuel de la variable (en français)	Unités
SWDOWN	Shortwave incoming radiation at surface	Rayonnement solaire descendant en surface	$W.m^{-2}$
U/V	x/y wind component	Composante x/y de la vitesse du vent	$m.s^{-1}$
T2	Temperature at 2M	Température à 2 mètres	K
QVAPOR	Water vapor mixing ratio	Humidité spécifique	$kg.kg^{-1}$

Dans la prochaine section, le rayonnement solaire, la température au sol et la vitesse du vent (U/V/W) produits par les simulations sur les années 2017 et 2018 sont, tout d'abord, analysés, comparés puis validés par rapport aux mesures au sol (Météo-France), qui couvrent l'essentiel de l'île de La Réunion et, afin d'examiner plus en détail la performance des données de rayonnement solaire. Les données de sortie du modèle WRF sont aussi validées par rapport aux données de rayonnement solaire SARAH 2.1 en points de grille.

2.5 Comparaison et analyse des données météorologiques

Dans cette section, la comparaison et la validation des données météorologiques WRF sont faites. Pour cela, les données au sol Météo-France sur les 16 stations précédemment décrites et les données d'irradiance solaire SARAH-2.1 sont utilisées. Pour rendre compte de la pertinence des données, la validation va utiliser plusieurs mesures et scores statistiques pour évaluer la qualité des données d'irradiance solaire, de température et vitesse de vent. Outre le biais et le RMSE communément utilisés, sont utilisés l'erreur absolue moyenne ainsi que le

coefficient de Pearson obtenus à partir des mesures au sol ou de SARAH-2.1 et des données WRF. Le biais et le RMSE ne fournissent pas à eux seuls une information suffisante sur la qualité climatique d'un jeu de données. Pour chaque jeu de données d'irradiance solaire, le nombre de mois (ou de jours) qui dépassent la précision cible pour caractériser la qualité des enregistrements de données est fourni. Les mesures de qualité appliquées sont décrites dans les chapitres suivants.

Ainsi, la variable "*y*" décrit le jeu de données à valider (ici, WRF) et "*o*" désigne le jeu de données de référence (par exemple, stations Météo-France). Le pas de temps individuel est marqué par "*k*" et "*n*" est le nombre total de pas de temps.

2.5.1 Les mesures statistiques

- **Biais**

Le biais (également appelé erreur moyenne) est défini comme la différence moyenne entre la moyenne de deux jeux de données, résultant de la moyenne arithmétique de la différence sur les éléments des jeux de données. Il indique si le jeu de données surestime ou sous-estime en moyenne le jeu de données de référence.

$$Biais = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k) = \bar{y} - \bar{o} \tag{10}$$

- **Erreur absolue moyenne**

Contrairement au biais, l'erreur absolue moyenne (Mean Absolute Error, MAE) est la moyenne arithmétique des valeurs absolues des différences entre chaque élément (toutes les paires) de la série chronologique. Il s'agit donc d'une mesure pertinente de l'"erreur" moyenne d'un jeu de données.

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|y_k - o_k| \tag{11}$$

- **Erreur quadratique moyenne (Root Mean Squared Error, RMSE)**

Le RMSE est l'écart type des résidus (erreurs). Les résidus sont une mesure de l'éloignement des points de données par rapport à la ligne de régression ; le RMSE est une mesure de la dispersion de ces résidus. En d'autres termes, elle indique à quel point les données sont concentrées autour de la ligne de meilleur ajustement. L'erreur quadratique moyenne est couramment utilisée en climatologie, en prévision et en analyse de régression pour vérifier les résultats expérimentaux.

$$RMSE = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2 \tag{12}$$

- **Le coefficient de Pearson**

Le coefficient de Pearson est un indice reflétant une relation linéaire entre deux variables continues. Le coefficient de corrélation varie entre -1 et +1, 0 reflétant une relation nulle entre les deux variables, une valeur négative (corrélation négative) signifiant que lorsqu'une des variables augmente, l'autre diminue ; tandis qu'une valeur positive (corrélation positive) indique que les deux variables varient ensemble dans le même sens

$$r = \frac{\sum_{k=1}^{n}(y_k - \bar{y})(o_k - \bar{o})}{\sqrt{\sum_{k=1}^{n}(y_k - \bar{y})^2}\ \sqrt{\sum_{k=1}^{n}(o_k - \bar{o})^2}} \quad (13)$$

2.5.2 Validation des données WRF avec Météo-France et SARAH-2.1

Avant de pouvoir utiliser les données simulées de WRF, il faut pouvoir les valider. (Ma et al., 2023) ont simulé le climat actuel du plateau tibétain et ont validé ces données observations de stations météorologiques au sol en concluant que WRF capturait avec succès le modèle spatial et temporel de la température moyenne et des précipitations. Sur l'île de La Réunion, il n'y a pas eu à proprement parler de validation des données climatiques simulées par WRF par des observations au sol mais Diagne et al. (2014) ont utilisé le modèle WRF pour faire de la prévision solaire à La Réunion et De Meij et al. (2018) ont étudié la sensibilité de plusieurs schémas microphysiques et dynamiques sur les valeurs calculées du GHI dans le modèle WRF. Afin d'évaluer et de valider les données météorologiques WRF, comme ont été validées les données GHI satellitaires SARAH-2.1 (Pfeifroth et al., 2019), à savoir les données d'irradiance solaire, de température au sol et de vitesse de vent, par rapport aux données au sol Météo-France et les données satellitaires SARAH-2.1, qui ont été présentées plus tôt dans le chapitre. Il faut préciser que la comparaison a été faite aux seize points de grille de la carte correspondant à la longitude et latitude des différentes stations Météo-France (MF) dont les données sur les années 2017 et 2018 étaient en notre possession. Quant à la validation des données d'irradiance solaire vis-à-vis des données SARAH-2.1, elle se fait à chaque point de grille (5 km × 5 km) sur la carte réunionnaise. Nous rappelons que seule l'irradiance solaire en 2017 est validée par les données SARAH-2.1 car ces données satellitaires s'étendent seulement jusqu'en 2017, le reste se faisant grâce aux données au sol MF.

- **Irradiance solaire**

Moyenne mensuelle

Les résultats de la validation de la moyenne mensuelle du jeu de données WRF d'irradiance solaire sont résumés dans le tableau 2.4 pour 2017 et le tableau 2.5 pour 2018. Cela montre que le biais est de l'ordre de 8 W/m^2 et que l'erreur absolue moyenne (MAE) est de l'ordre de 20.94 W/m^2 (par rapport à MF) et 23.95 W/m^2 (par rapport à SARAH-2.1) pour 2017 et 20.85 W/m^2 pour 2018, ce qui représente une erreur d'environ 10%, erreur non négligeable, mais qui prouve que sur les données simulées restent cohérentes sur les deux

années mensuellement. Il existe une forte corrélation (respectivement 89% et 79%) entre les données WRF, SARAH-2.1 et MF.

Tableau 2.4 Résultats de la comparaison entre les données mensuelles d'irradiance solaire entre les données WRF et SARAH-2.1 et Météo-France sur la période de 2017. Sont inclus le nombre de mois analysés, le biais, l'erreur absolue moyenne, l'erreur quadratique moyenne et le coefficient de Pearson ainsi que le pourcentage de mois excédent la précision cible qui est de 13 W/m^2 (définie dans (Pfeifroth et al., 2019)).

Irradiance solaire	N_{mois}	*Biais [W/m^2]*	*MAE [W/m^2]*	*RMSE [W/m^2]*	*r*	$Frac_{mon} >$ *précision cible [%]*
SARAH-2.1	12	8.30	23.95	31.29	0.89	25 (> 13 W/m^2)
Météo-France	12	7.95	20.94	26.35	0.79	33 (> 13 W/m^2)

Tableau 2.5 Résultats de la comparaison entre les données mensuelles d'irradiance solaire entre les données WRF et Météo-France sur la période de 2018.

Irradiance solaire	N_{mois}	*Biais [W/m^2]*	*MAE [W/m^2]*	*RMSE [W/m^2]*	*r*	$Frac_{mon} >$ *précision cible [%]*
Météo-France	12	6.38	20.85	25.55	0.79	16 (> 13 W/m^2)

Les figure 2.11 et 2.12 représentent le biais entre les données WRF et les données Météo-France par mois (toutes stations confondues) sur la période 2017/2018 et par station (tous mois confondus). Ainsi, nous pouvons juger de la pertinence des données d'un point temporel et spatial. Les diagrammes boîte-moustache représentent l'intervalle entre les percentiles de 25 % et 75 % (1er et 3^{e} quartiles) par les boîtes colorées ; les moustaches s'étendent jusqu'à 1,5 fois l'écart interquartile. Les erreurs proviennent essentiellement au mois de mars et en été plutôt qu'en hiver. Cela peut s'expliquer par de plus hautes valeurs d'irradiance solaire sur cette période. Il faut noter que pour toutes les stations, les données sont jugées homogènes et cohérentes avec les sorties de WRF.

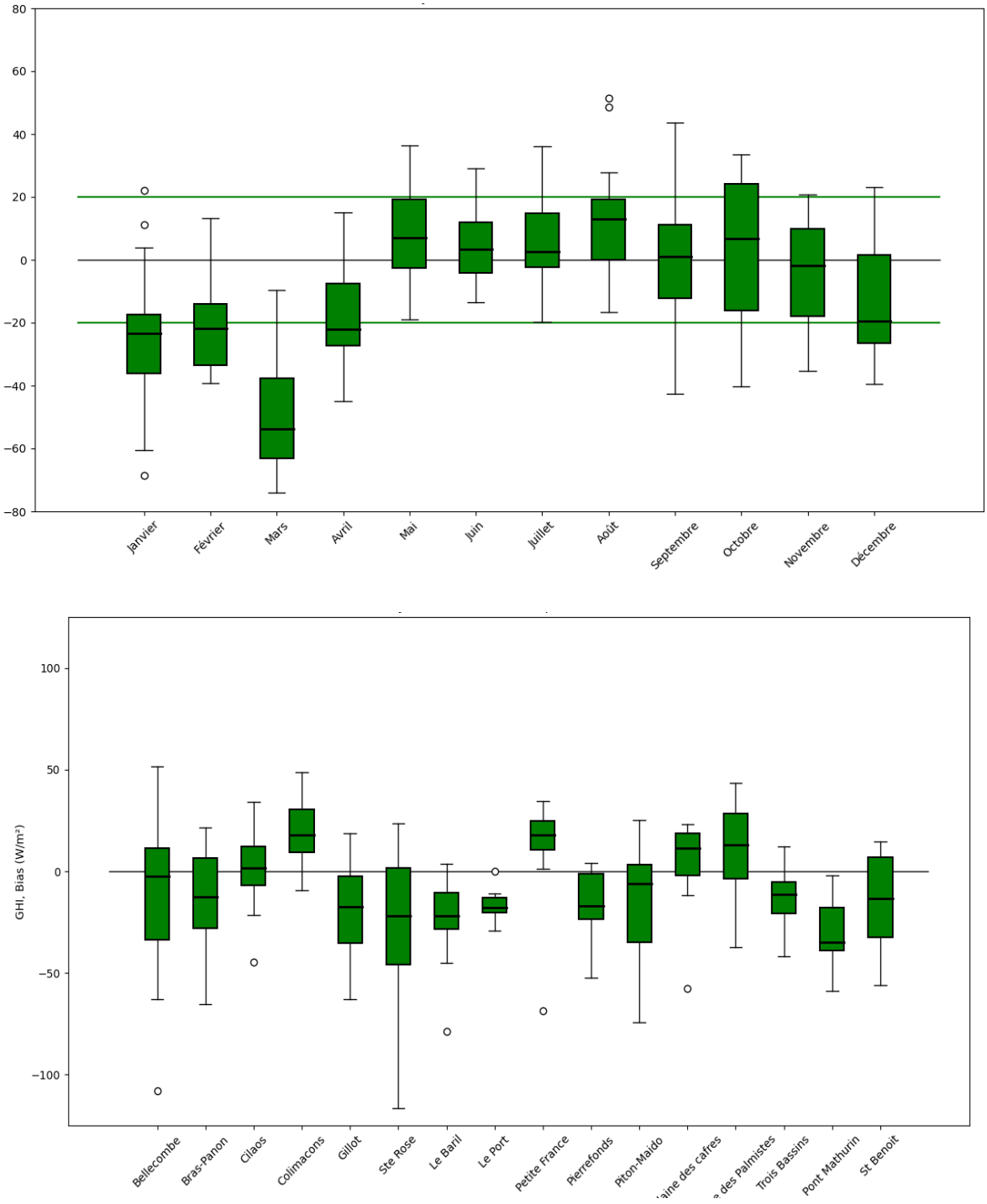


Figure 2.11 Biais entre les données mensuelles de GHI en 2017 par mois (haut) et par station (bas) entre les données simulées WRF et les données au sol Météo-France

Les résultats entre les données WRF et Météo-France en 2018 sont sensiblement les mêmes qu'en 2017 et sont visibles dans le tableau 2.5.

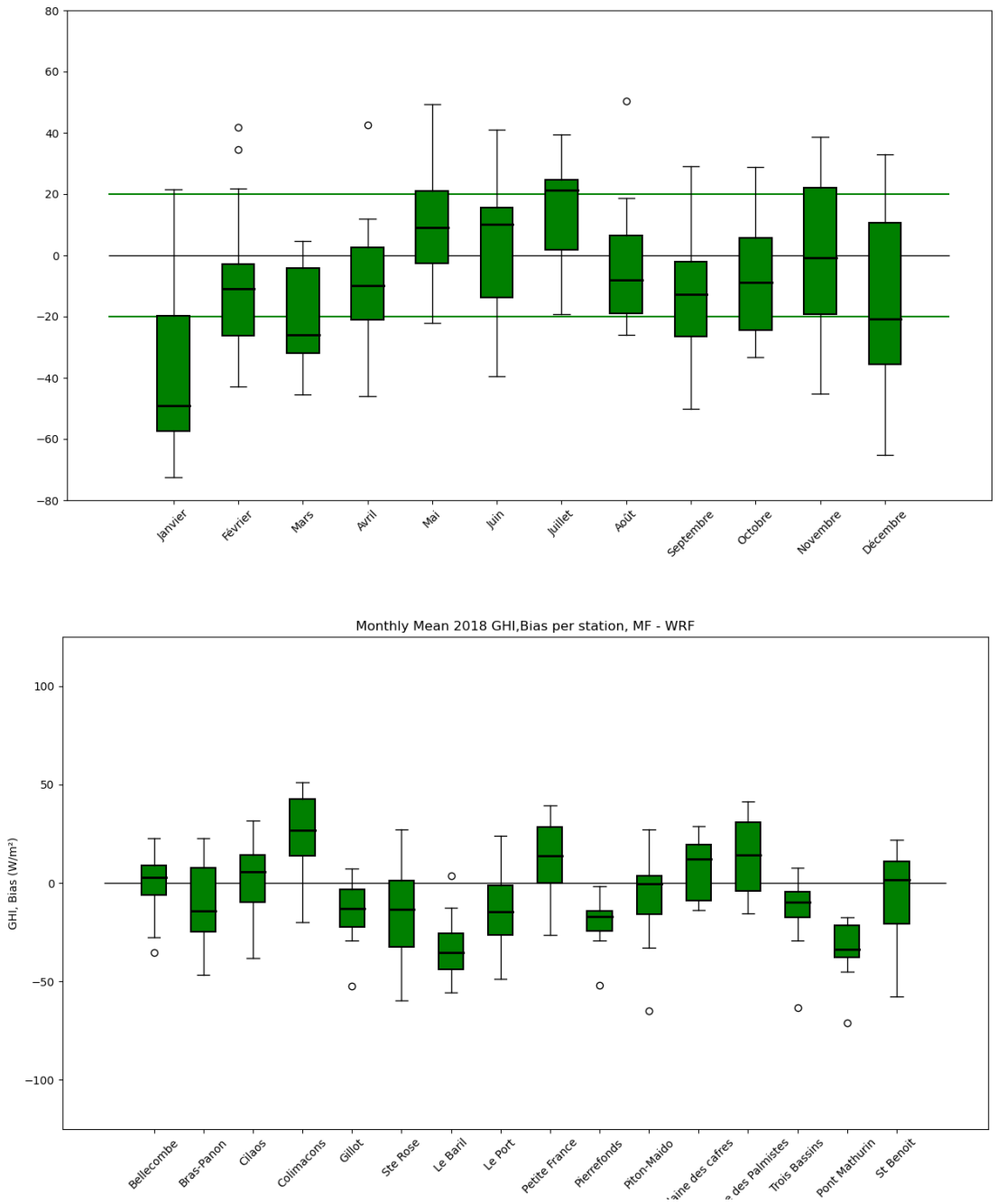


Figure 2.12 Biais entre les données mensuelles de GHI en 2018 par mois (haut) et par station (bas) entre les données simulées par WRF et les données au sol Météo-France

Moyenne journalière

Le tableau 2.6 rassemble les résultats de la validation des moyennes journalières des données d'irradiance solaire du modèle WRF par rapport au jeu de données SARAH-2.1 et des données au sol Météo-France pour l'année 2017. Comme précédemment, concernant l'année 2018, la comparaison et la validation n'ont pu se faire qu'avec les données à notre disposition, c'est-à-dire les données observables au sol Météo-France. Les résultats montrent que le biais est de 9 W/m^2 et que l'erreur absolue moyenne (MAE) est de l'ordre de 49.21 W/m^2 (par rapport à MF) et 59.57 W/m^2 (par rapport à SARAH-2.1) pour 2017 et 50.49 W/m^2 pour 2018,

ce qui représente une erreur entre 20% et 25%, erreur non négligeable. Cette plus grande différence s'explique par le plus grand nombre de données prises en compte pour cette comparaison, mais qui prouve que sur les données simulées restent cohérentes mensuellement sur les deux années. Il faut aussi noter une corrélation de plus de 55% (respectivement 79% et 57%) entre les données WRF et SARAH-2.1 et MF en 2017 et 53% avec les données MF en 2018.

Tableau 2.6 Résultats de la comparaison entre les données journalières d'irradiance solaire entre les données WRF et Météo-France sur la période de 2017. Sont inclus le nombre de jours analysés, le biais, l'erreur absolue moyenne, l'erreur quadratique moyenne et le coefficient de Pearson ainsi que le pourcentage de mois excédent la précision cible qui est de 13 W/m^2.

Irradiance solaire	N_{jour}	*Biais [W/m^2]*	*MAE [W/m^2]*	*RMSE [W/m^2]*	*r*	*$Frac_{jour}$ > précision cible [%]*
SARAH 2.1	365	8.6	59.57	62.95	0.79	21 (> 13 W/m^2)
Météo-France	365	9.09	49.21	66.49	0.57	30 (> 13 W/m^2)

Tableau 2.7 Résultats de la comparaison entre les données journalières d'irradiance solaire entre les données WRF et Météo-France sur la période de 2018.

Irradiance solaire	N_{jour}	*Biais [W/m^2]*	*MAE [W/m^2]*	*RMSE [W/m^2]*	*r*	*$Frac_{jour}$ > précision cible [%]*
Météo-France	365	6.61	50.47	66.36	0.53	25 (> 13 W/m^2)

La figure 2.13 représente le biais entre les données WRF et les données Météo-France sur la période 2017 et par station. Nous obtenons plus de valeurs aberrantes (ou « *outliers* ») sur les données journalières de manière générale, car il existe un plus grand nombre de données à comparer et donc une plus grande possibilité de voir apparaitre ce type de valeurs. De plus, pour toutes les stations, les données sont jugées homogènes et cohérentes avec les sorties WRF.

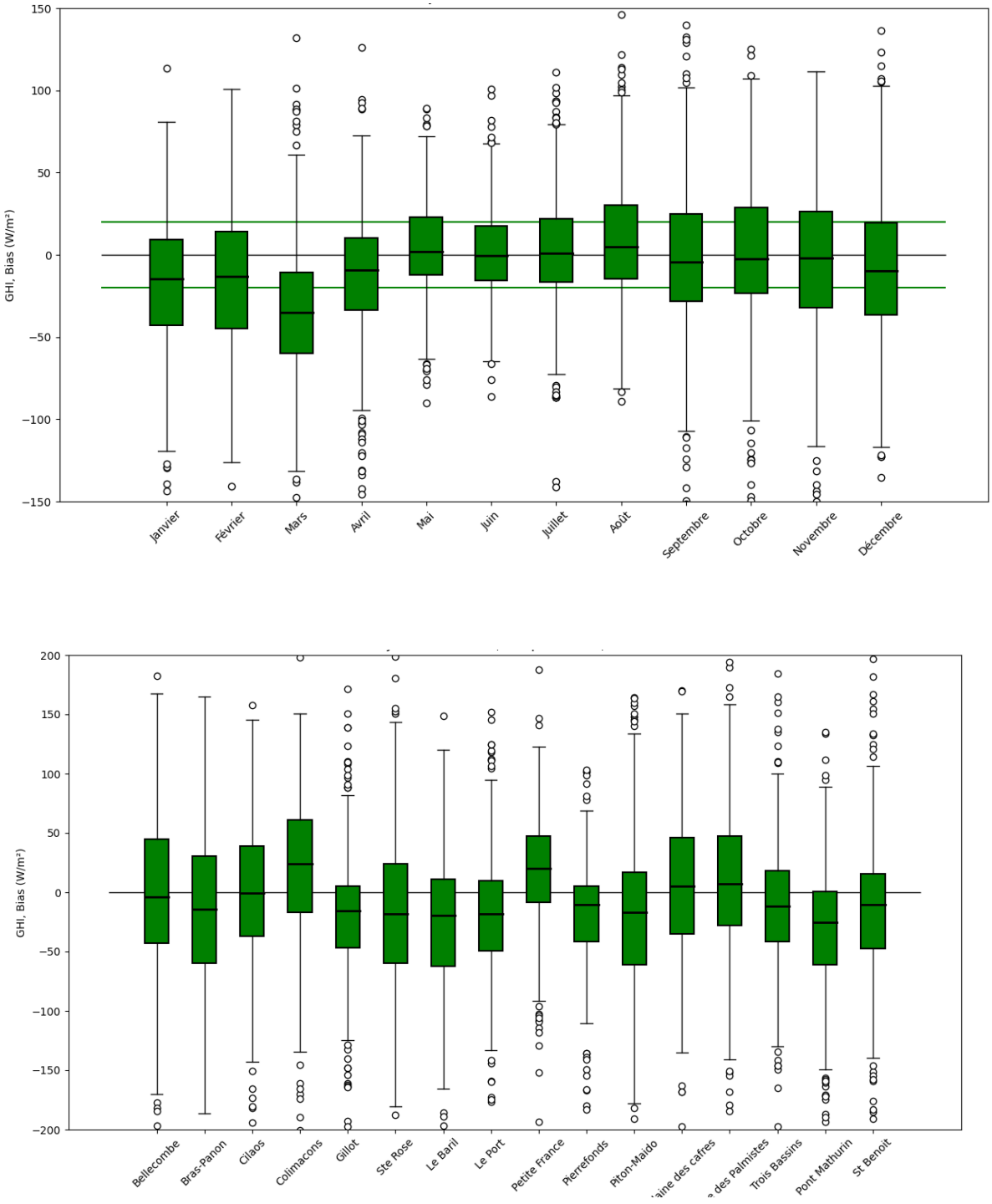


Figure 2.13 Biais entre les données journalières de GHI en 2017 par mois (haut) et par station (bas) entre les données simulées WRF et les données au sol Météo-France

Les résultats entre les données WRF et Météo-France en 2018 sont sensiblement les mêmes que ceux de 2017 et sont visibles dans le tableau 2.7.

- **Température**

Moyenne mensuelle

Les résultats de la validation de la moyenne mensuelle du jeu de données WRF des températures ont été résumés dans le tableau 2.8 pour 2017 et le tableau 2.9 pour 2018. Ils montrent que le biais est de l'ordre de 0.85 °C et 0.78°C respectivement pour 2017 et 2018 et que l'erreur absolue moyenne (MAE) est de l'ordre de 0.94 °C par rapport à MF pour 2017 et 0.82 °C pour 2018, ce qui représente une erreur d'environ 5%, erreur que nous pouvons considérer comme négligeable et qui prouve que sur les données simulées restent cohérentes sur les deux années mensuellement. Il est bon de noter la forte corrélation (respectivement 91% et 97%) entre les données WRF et MF.

Tableau 2.8 Résultats de la comparaison entre les données mensuelles de température entre les données WRF et Météo-France sur la période de 2017. Sont inclus le nombre de mois analysés, le biais, l'erreur absolue moyenne, l'erreur quadratique moyenne et le coefficient de Pearson.

Température	N_{mon}	*Biais [°C]*	*MAE [°C]*	*RMSE [°C]*	*r*
Météo-France	12	0.85	0.94	1.09	0.91

Tableau 2.9 Résultats de la comparaison entre les données mensuelles de température entre les données WRF et Météo-France sur la période de 2018.

Température	N_{mois}	*Biais [°C]*	*MAE [°C]*	*RMSE [°C]*	*r*
Météo-France	12	0.78	0.82	0.96	0.97

Les figures 2.14 à 2.16 affichent le biais entre les données mensuelles et journalières WRF et les données Météo-France (toutes stations confondues) sur la période 2017/2018 et par station (tous mois confondus). Ce qui est observable dans ces figures est l'homogénéité des boites à moustaches et que les données WRF ont tendance à légèrement surévaluer la température par rapport aux données réelles au sol de MF d'où ce décalage quasi permanent vers le haut des boites d'1° C en moyenne. Autre observation vis-à-vis des stations est le nombre élevé de valeurs aberrantes sur les années 2017 et 2018 sur la station de Bras-Panon. Au-delà ce celle-ci, les données sont jugées très cohérentes avec les sorties WRF.

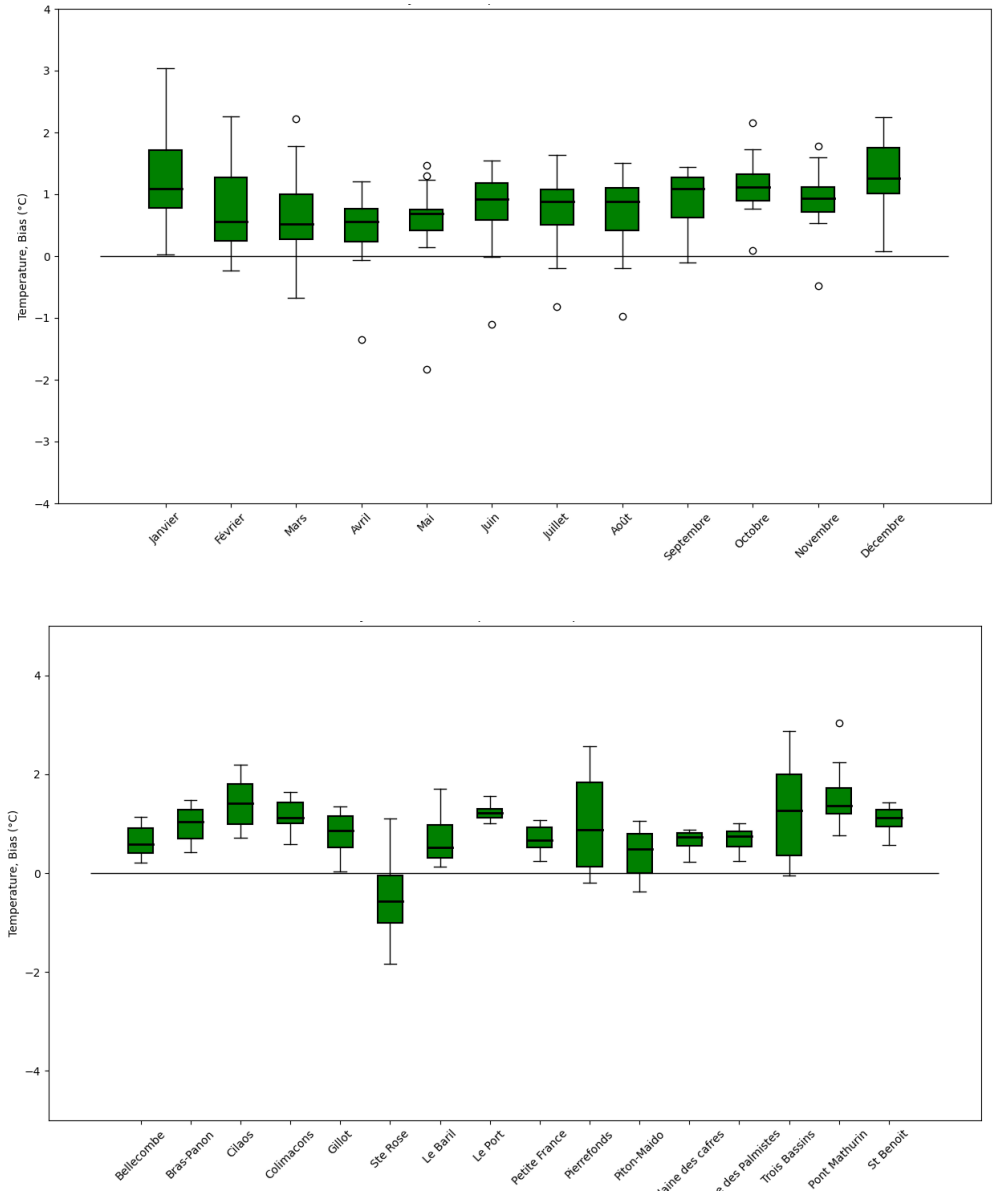


Figure 2.14 Biais entre les données mensuelles de température en 2017 par mois (haut) et par station (bas) entre les données simulées WRF et les données au sol Météo-France

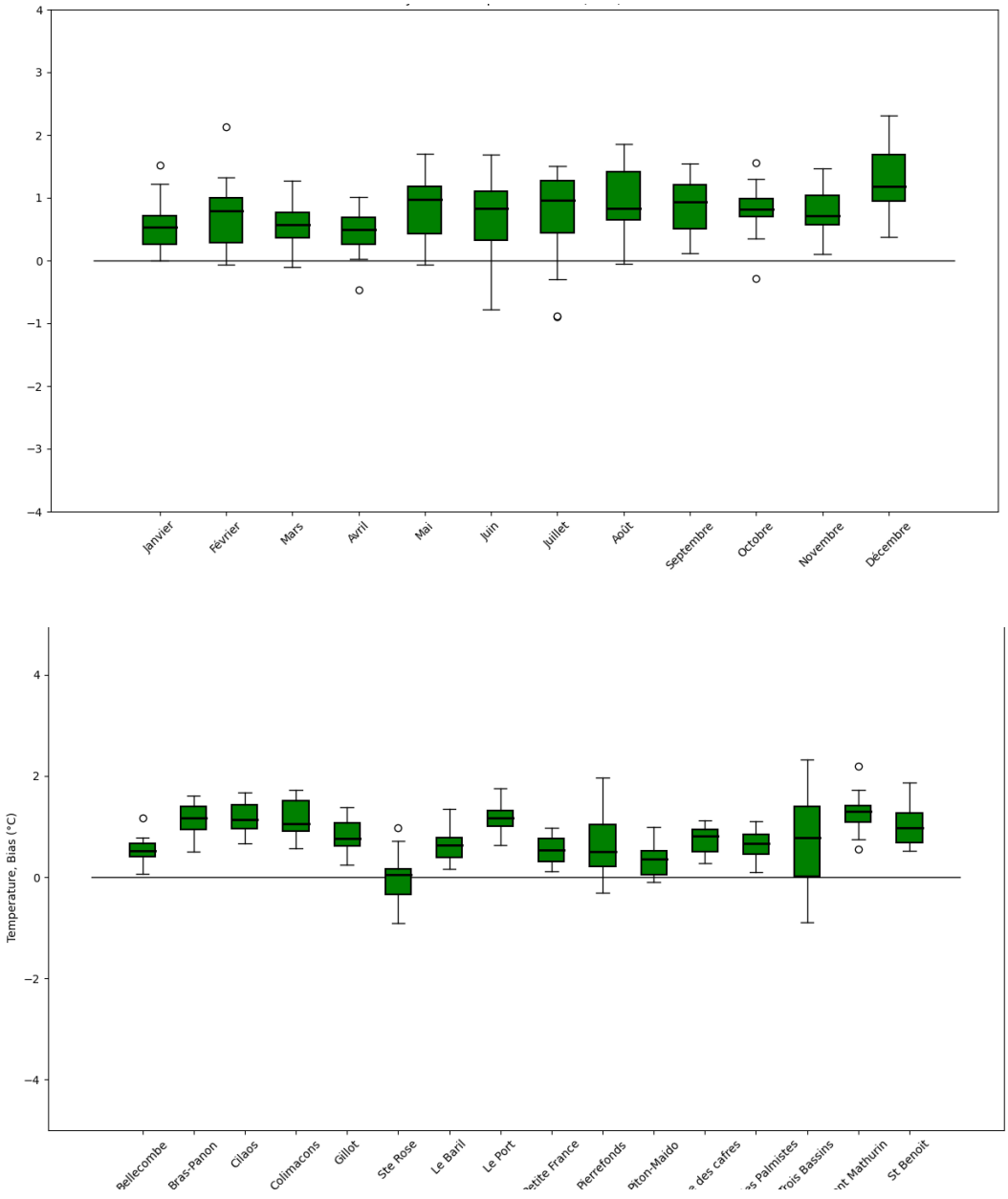


Figure 2.15 Biais entre les données mensuelles de température en 2018 par mois (haut) et par station (bas) entre les données simulées WRF et les données au sol Météo-France

Moyenne journalière

Les tableaux 2.10 et 2.11 représentent les résultats de la validation des moyennes journalières des données de température du modèle WRF par rapport au jeu de données au sol Météo-France pour les années 2017 et 2018. Ces résultats montrent que le biais est de l'ordre de 0.96 °C et que l'erreur absolue moyenne (MAE) est de l'ordre de 3.49 °C en 2017 et 3.28°C en 2018 soit respectivement une erreur de l'ordre de 17% et 15% respectivement. Bien qu'une erreur plus importante soit à noter comparativement aux données mensuelles, cela s'explique par le plus grand jeu de données à notre disposition lors de la comparaison. Nous notons toutefois une forte corrélation entre le jeu de données WRF et MF (respectivement 78% et 82%).

Tableau 2.10 Résultats de la comparaison entre les données journalières de température entre les données WRF et Météo-France sur la période de 2017. Sont inclus le nombre de jours analysés, le biais, l'erreur absolue moyenne, l'erreur quadratique moyenne et le coefficient de Pearson.

Température	N_{jour}	*Biais [°C]*	*MAE [°C]*	*RMSE [°C]*	*r*
Météo-France	365	0.96	3.49	4.61	0.78

Tableau 2.11 Résultats de la comparaison entre les données journalières de température entre les données WRF et Météo-France sur la période de 2018.

Température	N_{jour}	*Biais [°C]*	*MAE [°C]*	*RMSE [°C]*	*r*
Météo-France	365	0.90	3.28	4.33	0.82

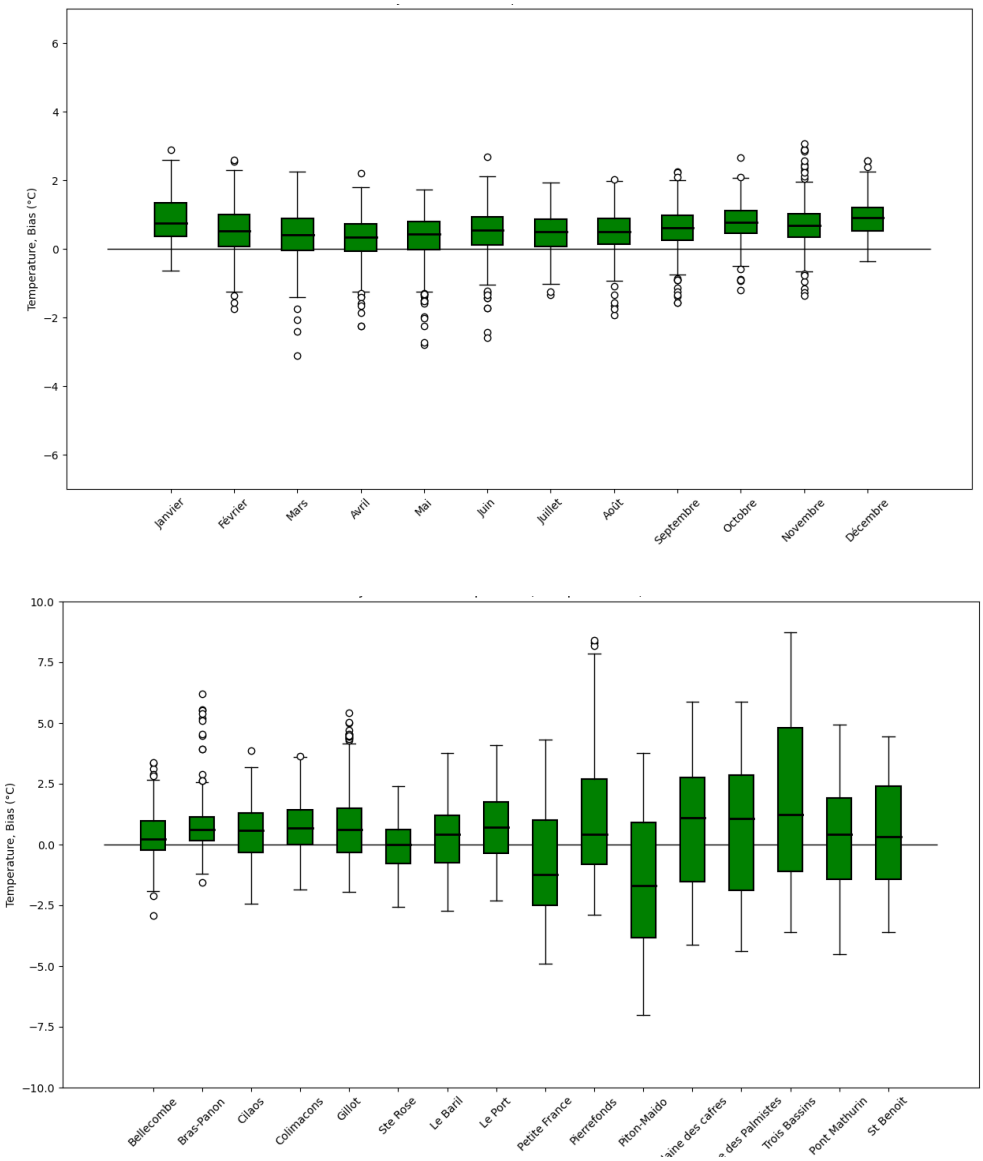


Figure 2.16 Biais entre les données journalières en température en 2017 par mois (haut) et par station (bas) entre les données simulées WRF et les données au sol Météo-France

- **Vitesse de vent**

Moyenne mensuelle

Les résultats de la validation de la moyenne mensuelle du jeu de données WRF des vitesses de vent ont été résumés dans le tableau 2.12 pour 2017 et 2.13 pour 2018. Ils montrent que le biais est de l'ordre de 0.39 $m.s^{-1}$ et 0.61 $m.s^{-1}$ respectivement pour 2017 et 2018 et que l'erreur absolue moyenne (MAE) est de de 0.99 $m.s^{-1}$ par rapport à MF pour 2017 et 1.06 $m.s^{-1}$ pour 2018, ce qui représente une erreur moyenne d'environ 35%. Contrairement aux deux autres paramètres, la précision sur les sorties de vitesses de vent par simulations WRF donne beaucoup moins satisfaction vis-à-vis des données au sol Météo-France. Il reste toutefois une assez forte corrélation (respectivement 64% et 77%) entre les données WRF et MF.

Tableau 2.12 Résultats de la comparaison entre les données mensuelles de vitesse de vent entre les données WRF et Météo-France sur la période de 2017. Sont inclus le nombre de mois analysés, le biais, l'erreur absolue moyenne, l'erreur quadratique moyenne et le coefficient de Pearson.

Vitesse du vent	N_{mois}	*Biais* $[m.s^{-1}]$	*MAE* $[m.s^{-1}]$	*RMSE* $[m.s^{-1}]$	*r*
Météo-France	12	0.39	0.99	1.27	0.64

Tableau 2.13 Résultats de la comparaison entre les données mensuelles de vitesse de vent entre les données WRF et Météo-France sur la période de 2018.

Vitesse du vent	N_{mois}	*Biais* $[m.s^{-1}]$	*MAE* $[m.s^{-1}]$	*RMSE* $[m.s^{-1}]$	*r*
Météo-France	12	0.61	1.06	1.36	0.77

Les figures 2.17 à 2.20 représentent le biais entre les données mensuelles et journalières WRF et les données Météo-France (toutes stations confondues) sur la période 2017/2018 et par station (tous mois confondus). Les boites à moustaches sont centrées autour de 0. Notons qu'il y a une légère surévaluation de la vitesse du vent visible en 2017 et 2018 entre octobre et février. Il faut noter que pour certaines stations, les données sont jugées homogènes et cohérentes avec les sorties WRF. Cependant, pour le Port, Petite France et Trois Bassins, les valeurs des vitesses de vent simulées sont surestimées, contrairement au Baril, les valeurs sont nettement sous-estimées.

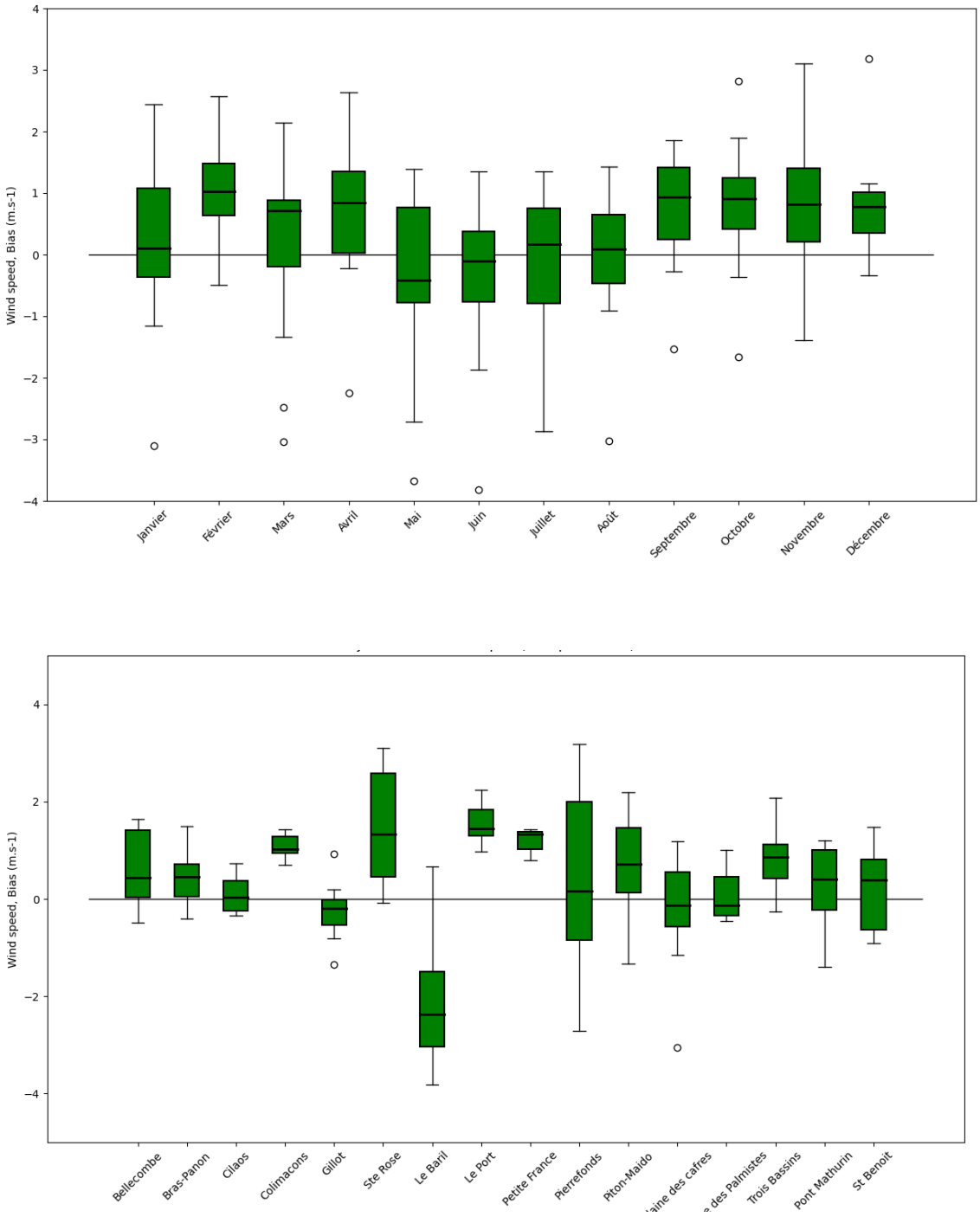


Figure 2.17 Biais entre les données mensuelles de vitesse de vent en 2017 par mois (haut) et par station (bas) entre les données simulées WRF et les données au sol Météo-France

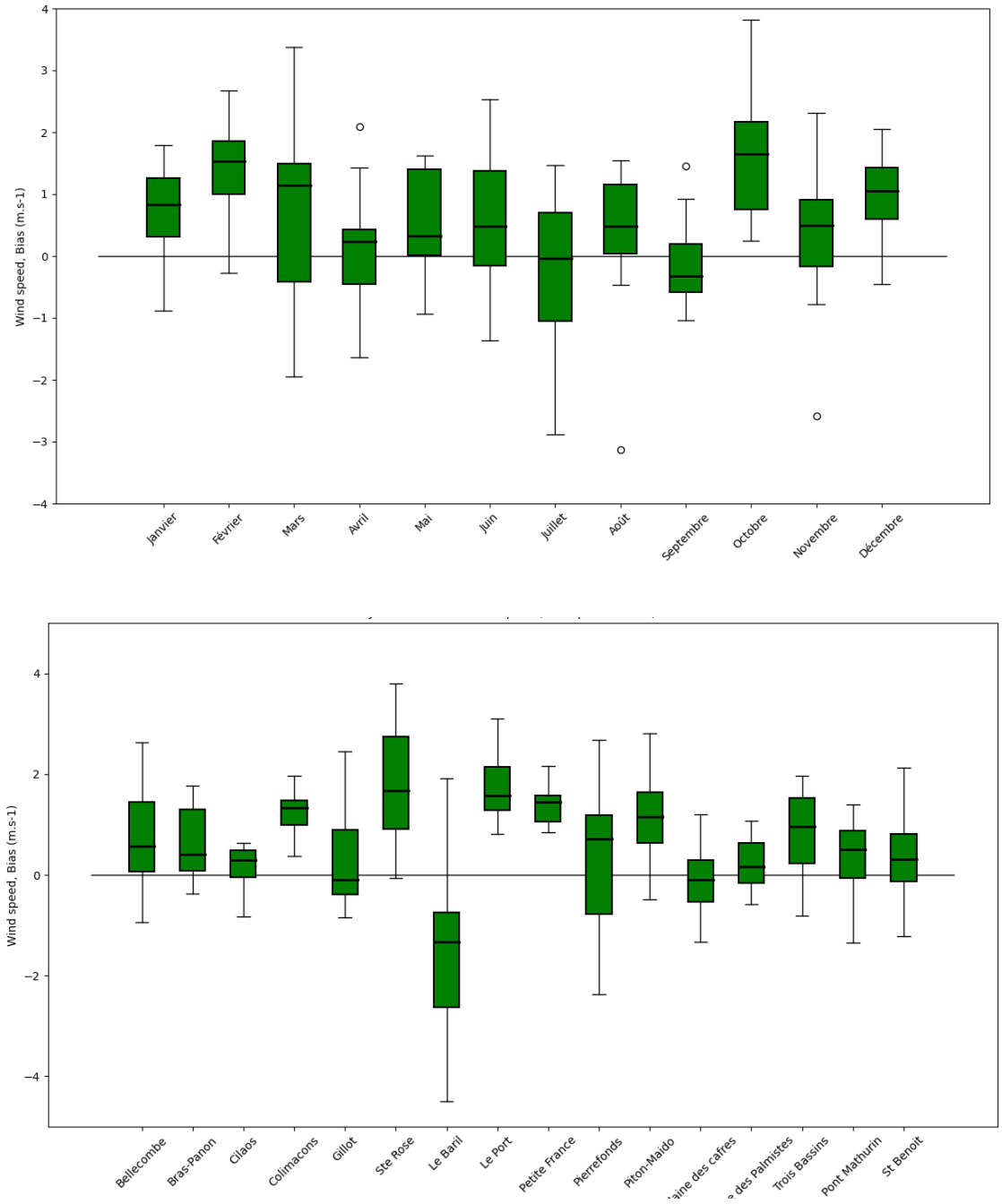


Figure 2.18 Biais entre les données mensuelles de vitesse de vent en 2018 par mois (haut) et par station (bas) entre les données simulées WRF et les données au sol Météo-France

Moyenne journalière

Les résultats de la validation des moyennes journalières des données de vitesse de vent du modèle WRF par rapport au jeu de données au sol Météo-France pour l'année 2017 et 2018 sont regroupés dans les tableaux 2.14 et 2.15. Ces résultats montrent que le biais est de l'ordre de 0.4 $m.s^{-1}$ (0.44 $m.s^{-1}$ en 2017 et 0.42 $m.s^{-1}$ en 2018) et que l'erreur absolue moyenne (MAE) est de l'ordre de 1.95 $m.s^{-1}$ en 2017 et 2.08 $m.s^{-1}$ en 2018, soit respectivement une erreur de l'ordre de 67% à 70%. A ce niveau d'erreur, il faut clairement énoncer que les données de

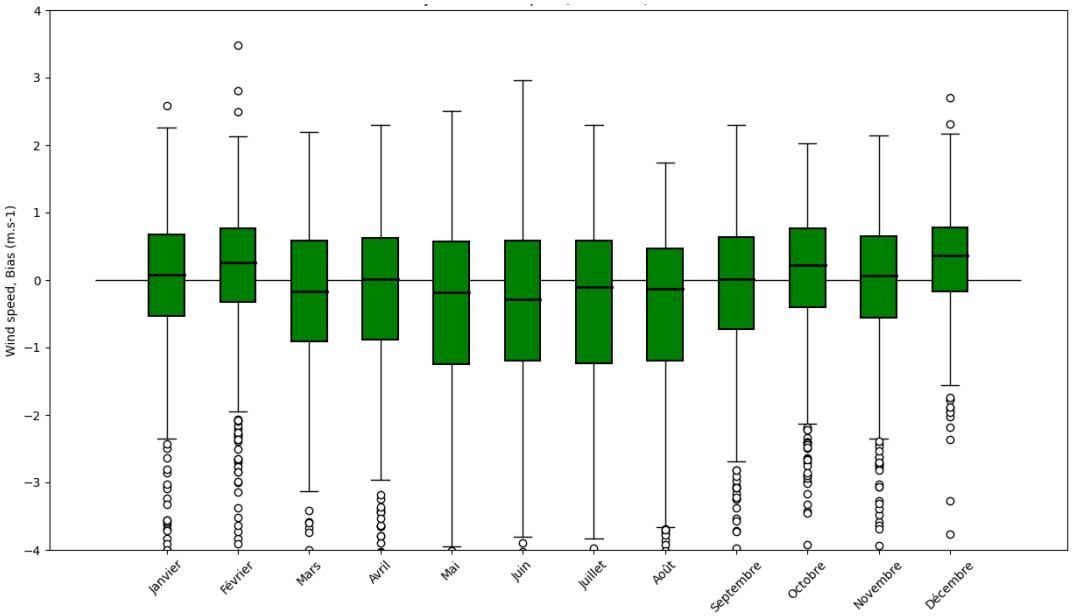
vitesse de vent simulées par WRF même s'il peut être constaté (comme illustré dans la figure 32) qu'elles sont dans la plupart des cas centrées sur l'axe des 0 comportent des erreurs et sont éloignées d'un point de vue statistique de données au sol Météo-France. A noter le nombre plus importants de manière générale d'« *outliers* » ou valeurs aberrantes sur les données journalières à notre disposition lors de la comparaison. Nous notons toutefois une corrélation non négligeable entre le jeu de données WRF et MF (respectivement 52% et 78%).

Tableau 2.14 Résultats de la comparaison entre les données journalières de vitesse de vent entre les données WRF et Météo-France sur la période de 2017. Sont inclus le nombre de mois analysés, le biais, l'erreur absolue moyenne, l'erreur quadratique moyenne et le coefficient de Pearson.

Vitesse du vent	N_{jour}	*Biais* $[m.s^{-1}]$	*MAE* $[m.s^{-1}]$	*RMSE* $[m.s^{-1}]$	*r*
Météo-France	365	0.44	1.95	2.77	0.52

Tableau 2.15 Résultats de la comparaison entre les données journalières de vitesse de vent entre les données WRF et Météo-France sur la période de 2018.

Vitesse du vent	N_{jour}	*Biais* $[m.s^{-1}]$	*MAE* $[m.s^{-1}]$	*RMSE* $[m.s^{-1}]$	*r*
Météo-France	365	0.42	2.08	2.97	0.78



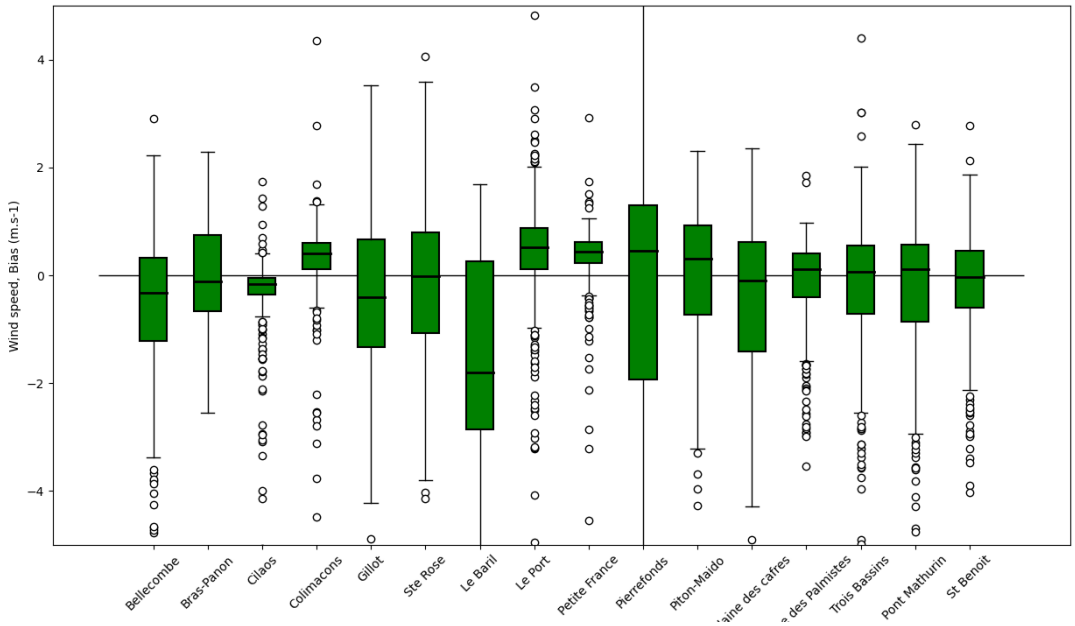
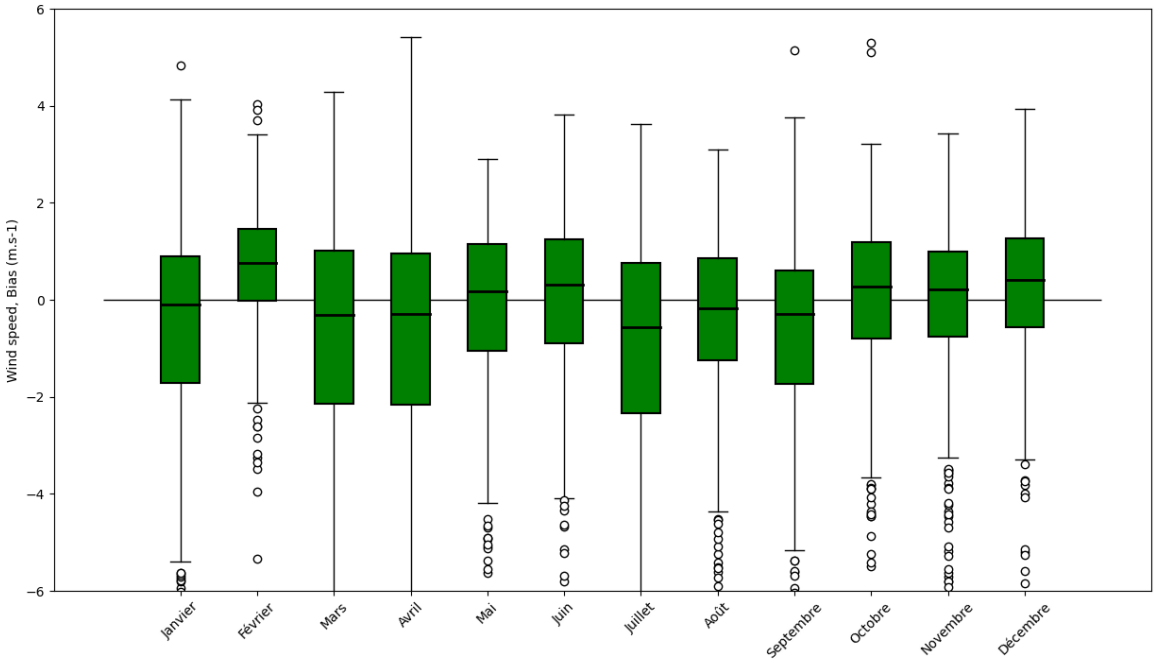


Figure 2.19 Biais entre les données journalières de vitesse de vent en 2017 par mois (haut) et par station (bas) entre les données simulées WRF et les données au sol Météo-France



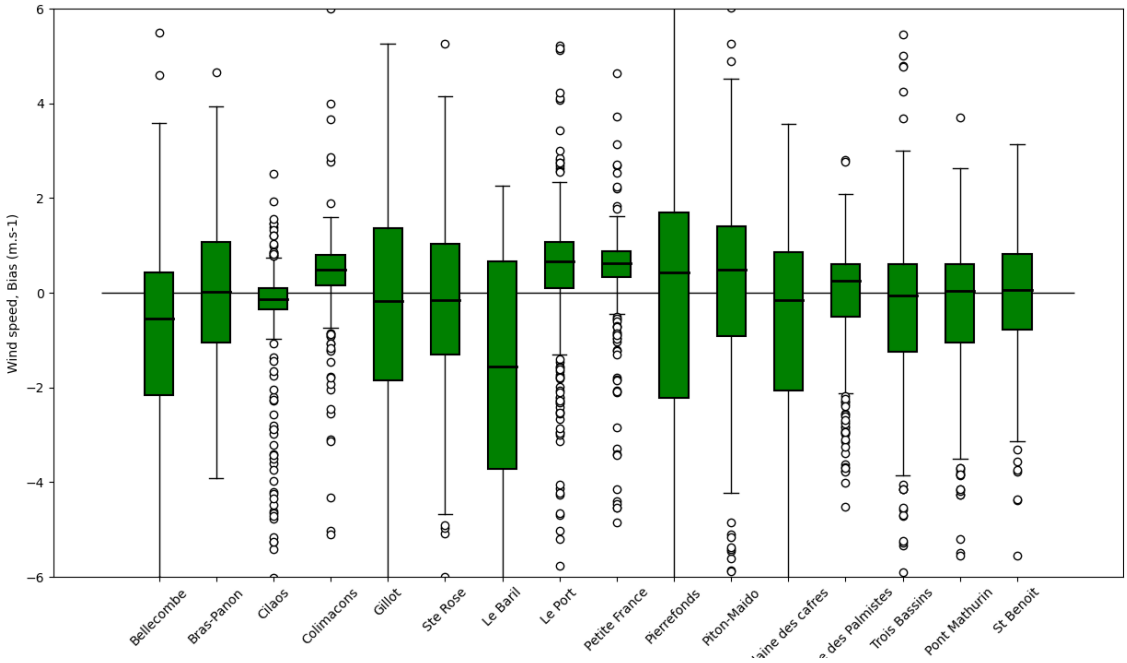


Figure 2.20 Biais entre les données journalières de vitesse de vent en 2018 par mois (haut) et par station (bas) entre les données simulées WRF et les données au sol Météo-France

Discussion

Ce paragraphe traite de la discussion sur l'étude de la validation des données climatiques de sortie WRF, à savoir les données d'irradiance solaire, de température et de vitesse de vent, par les données d'observations directes au sol de Météo-France de haute qualité pour 2017/2018 et avec les données climatiques satellitaires CM SAF SARAH-2.1 pour l'irradiance solaire en 2017.

Tout d'abord, nous avons analysé les différentes sources d'incertitude qui peuvent interférer dans la validation des données. La principale source d'erreur dans la validation de données climatiques journalières (et mensuelles) utilisées ici provient de l'intégration des valeurs horaires dans les moyennes journalières (puis mensuelles). Un autre facteur externe pouvant influencer la validation est la qualité des capteurs utilisés au sol par MF (pyranomètres, sonde thermométriques et anémomètres). Les relevés au sol avec une résolution très fine, comme ceux du BSRN, sont nécessaires pour valider les données de sortie horaires et cette étude se concentre donc uniquement sur la validation des moyennes journalières et mensuelles. Toutefois, seule la station météorologique de l'aéroport de Gillot sur l'île de La Réunion répond à ces critères sur la période 2017/2018.

Concernant le GHI, les données mensuelles montrent une forte corrélation entre les données de WRF et les autres bases de données, avec moins de 25% des valeurs mensuelles excédent la précision-cible de 13 W/m², gage de la qualité des données de simulation WRF. Les erreurs apparaissent en Mars. Après vérification, elles sont dues à la présence de certaines valeurs aberrantes dans les données au sol de MF. Au-delà de cela, les données sont cohérentes spatialement sur toutes les stations météorologiques et sur les 2 ans de validation. D'un point

de vue journalier, 25% ou moins des données journalières excèdent la précision-cible avec une forte corrélation des données, surtout avec le produit satellitaire griddé SARAH-2.1.

Concernant la température, les données mensuelles de WRF présentent une forte corrélation avec les données MF (91% et 97%) avec un biais de moins de 1°C. Les mêmes conclusions peuvent être tirées avec les données journalières avec une forte corrélation (78% et 82%). Toutefois, d'un point de vue temporel et spatial, les données de WRF sont surévaluées.

Concernant la vitesse du vent, pour les données mensuelles et journalières, un biais de moins de 0,5 $m.s^{-1}$ est observé et une corrélation moins forte est notée (entre 52% et 78%).

Nous pouvons avancer, que ce soit d'un point de vue spatial ou temporel, que les données climatiques générées par WRF, que ce soit le GHI, la température et la vitesse du vent, ont été validées par les données au sol et satellitaires. Ces données peuvent donc être utilisées dans la suite de ces travaux de thèse.

Conclusion

Le modèle WRF est historiquement un modèle largement utilisé au sein du laboratoire LE²P-Energy^{LAB} (Morel et al., 2014; Pohl et al., 2016) ainsi que, très récemment, dans le projet SWIO-Energy afin de modéliser à très haute résolution un territoire comme l'île de La Réunion dont la topographie est complexe et marquée. De plus, il est libre d'accès et possède une large communauté scientifique qui ne cesse de croître.

Le but de ce chapitre était de prendre en compte les limites des différentes bases de données à notre disposition et d'expliquer le choix d'utiliser le modèle WRF comme un « générateur de données » climatiques. Toutefois, il nous fallait être sûr de la pertinence de ces données avant de les utiliser comme données d'entrée pour nos modèles.

Ces données climatiques de sortie WRF, ainsi simulées, ont pu ainsi être analysées et validées sur les années 2017 et 2018 en les comparant aux données d'observations directes au sol de Météo-France de haute qualité et avec les données climatiques satellitaires CM SAF SARAH-2.1 pour l'irradiance solaire en 2017.

Le chapitre suivant décrit les modèles qui ont été choisis et utilisés pour ces travaux de thèse. Il explicite le choix des modèles : dans un premier temps, un modèle dit « statistique » est considéré ; il s'agit de l'approche de Johansen par cointégration. Cette approche populaire en Économétrie appliquée va être discutée. Puis les modèles réseaux de neurones vont être introduits. Leurs fonctionnements seront présentés, dont un modèle de réseaux de neurones récurrents va être explicité en particulier : le modèle à mémoire court et long terme (LSTM). Enfin, un état de l'art sur les modèles LSTM dans la prévision de production photovoltaïque sera réalisé afin de selectionner le modèle de prévision le plus pertinent pour notre problématique.

CHAPITRE 3

METHODES ET OUTILS

Introduction

Les prévisions ou estimations de l'évolution à court, moyen ou long terme d'une variable ou d'un phénomène servent de base à la prise de décisions et à l'élaboration de stratégies adéquates. L'état de l'art présenté dans le chapitre 1 mettait en avant les différents modèles adaptés à la prévision de production PV. Toutefois, dans ces travaux de thèse, le choix s'est porté sur deux modèles uniques statistiques qui nécessitent une moindre puissance de calcul et une plus simple implémentation. Ces modèles de prévision nécessitent des données historiques sous forme de séries temporelles en entrée (générées par le modèle WRF) et ne nécessite pas une prise en compte des paramètres internes à une centrale PV (que nous ne possédons pas).

La première partie du chapitre annonce l'approche du principe de cointégration et le modèle à correction d'erreur de Johansen (Fanchette et al., 2020). Cette méthode, essentiellement utilisée en Économétrie pour l'analyse de séries temporelles, n'avait jusqu'alors jamais été employé dans le domaine de la Physique énergétique. Puis, la seconde partie présente le deuxième modèle retenu, à savoir, le modèle Long Short Term Memory (LSTM), réseaux de neurones récurrents de plus en plus employés dans le domaine de la prévision de production PV (R. Huang et al., 2022; Muzaffar & Afshari, 2019; Qing & Niu, 2018; F. Wang et al., 2020b).

3.1 Modèle statistique de Johansen

La cointégration est une propriété statistique des séries temporelles qui permet de détecter la relation de long terme entre deux ou plusieurs séries temporelles. Ainsi, l'étude de la cointégration permet de tester l'existence d'une relation stable de long terme entre deux variables non stationnaires, en incluant des variables retards et des variables exogènes. Il existe plusieurs tests de la cointégration, le plus général étant celui de Johansen. Quel que soit le test retenu, il n'a de signification que sur des séries non stationnaires longues. Par conséquent, l'analyse de la cointégration permet d'identifier clairement la relation véritable entre deux variables, en recherchant l'existence d'un vecteur de cointégration et en éliminant son effet le cas échéant. Le modèle de cointégration de Johansen utilise deux tests statistiques : la statistique de la trace et celle de la valeur propre maximale. Les distributions asymptotiques de ces statistiques sont non standard. En ce qui concerne ces travaux de thèse, cette méthode statistique vise à créer un modèle de prévision à moyen terme de la production journalière et horaire d'électricité photovoltaïque pour les prochains jours, afin de répondre aux besoins des gestionnaires de réseaux électriques, des négociants et des producteurs en énergie. Un modèle paramétré et régressif qui est proposé est mieux construit pour faire de la prévision à court et moyen terme.

3.1.1 Analyse de séries temporelles

Afin de décrire le modèle statistique de Johansen, il est nécessaire de s'attarder sur plusieurs notions indispensables à la compréhension de la conception du modèle. Ce dernier s'appuie sur l'analyse de séries temporelles. Une série temporelle est un ensemble d'observations, ici des données météorologiques, sur les valeurs que prend une variable à

différents moments. Les données de séries temporelles sont collectées à intervalles réguliers, par exemple quotidiennement, hebdomadairement, mensuellement ou annuellement (Gujarati, 2004). Toutefois, la plupart des travaux empiriques basés sur des séries temporelles supposent que la série sous-jacente est stationnaire. La définition précise de la stationnarité d'une série temporelle sera abordée dans le prochain paragraphe, mais de manière générale, une série temporelle est stationnaire si sa moyenne et sa variance ne varient pas systématiquement dans le temps. En général, une série temporelle observée comporte quatre composantes principales. Ces composantes sont la tendance, le cycle, la saisonnalité, et l'irrégularité.

- La tendance : si les données d'une série temporelle ont une tendance générale, cela signifie que la série temporelle a une tendance générale à l'augmentation, à la diminution ou à la stagnation sur une longue période de temps.
- Le cycle : Cette composante des séries temporelles explique les variations à moyen terme des séries, variations à moyen terme de la série, causées par des circonstances, mais qui se répètent par cycles. La durée du cycle varie selon le type de série temporelle ; généralement, il s'étend sur une période plus longue, généralement deux ans ou plus.
- La saisonnalité : La variation qui se produit dans une série temporelle de données en raison du changement de saison à saison est appelée saisonnière. Le modèle saisonnier est celui qui se produit de manière répétée dans la série temporelle, mais à une période très spécifique de l'année.
- L'irrégularité : Ce sont les variations aléatoires d'une série de données, qui sont causées par des influences imprévisibles. Ces variations ne se répètent pas selon un schéma particulier. C'est cette partie qui fait généralement d'une série temporelle un processus stochastique.

Le point de départ de l'analyse des séries temporelles est le test de la propriété de stationnarité des données de ces séries temporelles. Dans les modèles de régression simples, par exemple les modèles des moindres carrés ordinaires (MCO), il est fortement supposé que la série de données soit stationnaire. Cependant, il existe une grande possibilité, dans la plupart des cas, d'une série non stationnaire dans le cas des variables économiques. En effet, il est établi dans la littérature (Gujarati, 2004; Ramenah et al., 2017) sur les séries temporelles qu'il peut y avoir une régression fallacieuse dans le cas de séries de données non stationnaires. Par conséquent, le test de stationnarité est une étape indispensable pour un modèle de prévision.

3.1.1.1 Stationnarité des séries temporelles

Les données étudiées dans cette thèse sont pour la plupart de données météorologiques ou des paramètres météorologiques dépendants (ex. la température extérieure, la vitesse du vent ou encore l'ensoleillement). Ce type de données est appelé données aléatoires ou stochastiques. Un processus aléatoire ou stochastique est un ensemble de variables aléatoires ordonnées dans le temps. Une série temporelle stochastique[1] est considérée comme étant stationnaire si sa moyenne et sa variance sont constantes dans le temps. Ou encore, si la covariance entre deux périodes de temps distinctes ne dépend que de la distance ou retard (« lag ») entre les deux périodes de temps et non du moment réel auquel la covariance est calculée (Gujarati, 2004). Dans la littérature sur les séries temporelles, un tel processus

[1] Le terme « stochastique » provient du mot grec « stokhos » qui signifie cible ou « bull's eye »

stochastique est connu comme processus faiblement stationnaire, ou stationnaire par covariance ou encore stationnaire du second ordre. Pour les besoins de ces travaux de thèse et dans la plupart des situations pratiques, cette définition de stationnarité est communément acceptée.

Pour expliquer mathématiquement la stationnarité, on considère Y_t une série temporelle stochastique avec comme propriétés :

Moyenne : $$E(Y_t) = \mu \quad (14)$$

Variance : $$var(Y_t) = E(Y_t - \mu)^2 = \sigma^2 \quad (15)$$

Covariance : $$\gamma_k = E[(Y_t - \mu)(Y_{t+k} - \mu)] \quad (16)$$

où E est l'Espérance (et donc la moyenne), γ_k, la covariance au lag k, est la covariance entre les valeurs de Y_t et Y_{t+k}, c'est-à-dire entre les valeurs de Y séparées de k périodes entre elles. Si k = 0, nous obtenons γ_0, qui est la variance de Y ($= \sigma^2$) et si k = 1, $\gamma_1$est la covariance entre deux valeurs successives de Y.

Pour résumer, si une série temporelle est stationnaire, sa moyenne, sa variance et son auto covariance (à différents lags) restent les mêmes, quel que soit le moment mesuré. Elle est pour ainsi dire invariable dans le temps. Une telle série temporelle aura tendance à revenir à sa moyenne (appelée « mean reversion » ou retour à la moyenne) et les fluctuations autour de cette moyenne (mesurées par sa variance) auront une amplitude globalement constante (Cuthbertson et al., 1992). Si une série temporelle n'est pas stationnaire au sens où elle vient d'être définie, cette série temporelle est non-stationnaire. En d'autres termes, elle aura une moyenne qui varie dans le temps ou une variance qui varie dans le temps ou les deux.

Les séries temporelles stationnaires se révèlent être très importantes. En effet, si la série temporelle étudiée est non stationnaire, il n'est possible d'étudier son comportement que sur la période de temps considérée. Chaque ensemble de données de cette série temporelle correspondra donc à un épisode particulier. Ainsi, il est impossible de la généraliser à d'autres périodes de temps. Par conséquent, pour les besoins de prévision, comme cela est le cas dans ces travaux de thèse, de telles séries temporelles ne sont d'aucune utilité pratique. De plus, (Yule, 1926) met en avant le phénomène de régressions dites « fallacieuses » si une régression d'une série temporelle non stationnaire est faite sur une autre série temporelle non stationnaire et qui amène à conclure de manière erronée une relation statistique significative entre ces deux séries alors qu'à priori il ne devrait pas y en avoir (Granger et al., 2001; Mills, 2019).

3.1.1.2 Tests de stationnarité

La nature des processus stochastiques stationnaires est cruciale. Deux hypothèses importantes sont alors à vérifier :

- La stationnarité de la série temporelle
- Rendre stationnaire une série non-stationnaire

Dans ce paragraphe, la première hypothèse sera abordée. La deuxième hypothèse sera traitée dans le chapitre 4. Il existe plusieurs tests de stationnarités, mais seuls ceux qui seront utilisés dans la suite du manuscrit seront traités (Fanchette et al., 2019, 2020). Dans la suite, trois tests seront examinés : (i) l'analyse graphique, (ii) le test du corrélogramme, (iii) le test de racine unitaire.

i. L'analyse graphique

Avant de procéder à des tests formels, il est toujours conseillé de tracer la série temporelle étudiée en fonction du temps (Fanchette et al., 2019). Un tel tracé donne un premier indice sur la nature probable de la série temporelle. L'évolution de la courbe informe sur la variation de la tendance ou de la moyenne de la série pouvant suggérer la stationnarité ou non de la série.

ii. La fonction d'autocorrélation (ACF) et corrélogramme

Un test simple de stationnarité est basé sur la fonction d'autocorrélation (ACF). La fonction d'autocorrélation des erreurs décrit la corrélation entre les erreurs de prédiction et les décalages temporels. Dans un modèle parfait, une seule valeur non nulle est obtenue à un décalage temporel de zéro (Hassan et al., 2021). La fonction d'autocorrélation (ACF) au décalage k, désignée par ρ_k est définie comme suit :

$$\rho_k = \frac{covariance\ au\ lag\ k}{variance} \tag{17}$$

où la covariance au lag k et la variance sont telles que définies précédemment.

Comme la covariance et la variance sont toutes deux mesurées dans les mêmes unités de mesure, ρ_k est un sans unité. Il se situe entre -1 et +1, comme tout coefficient de corrélation. Tracer ρ_k en fonction de k permet d'obtenir le graphique connu sous le nom de corrélogramme.

Compte tenu du fait qu'un échantillon de la série temporelle est étudié, la fonction d'autocorrélation d'échantillon est calculée (SAFC), $\widehat{\rho_k}$. Pour cela, la covariance d'échantillon au lag k $\widehat{\gamma_k}$ et la variance de l'échantillon $\widehat{\gamma_0}$ sont nécessaires et sont définies comme suit :

$$\widehat{\gamma_k} = \frac{\Sigma(Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{n} \tag{18}$$

$$\widehat{\gamma_0} = \frac{\Sigma(Y_t - \bar{Y})^2}{n} \tag{19}$$

où n est la taille de l'échantillon et $\bar{Y}$ est la moyenne de l'échantillon.

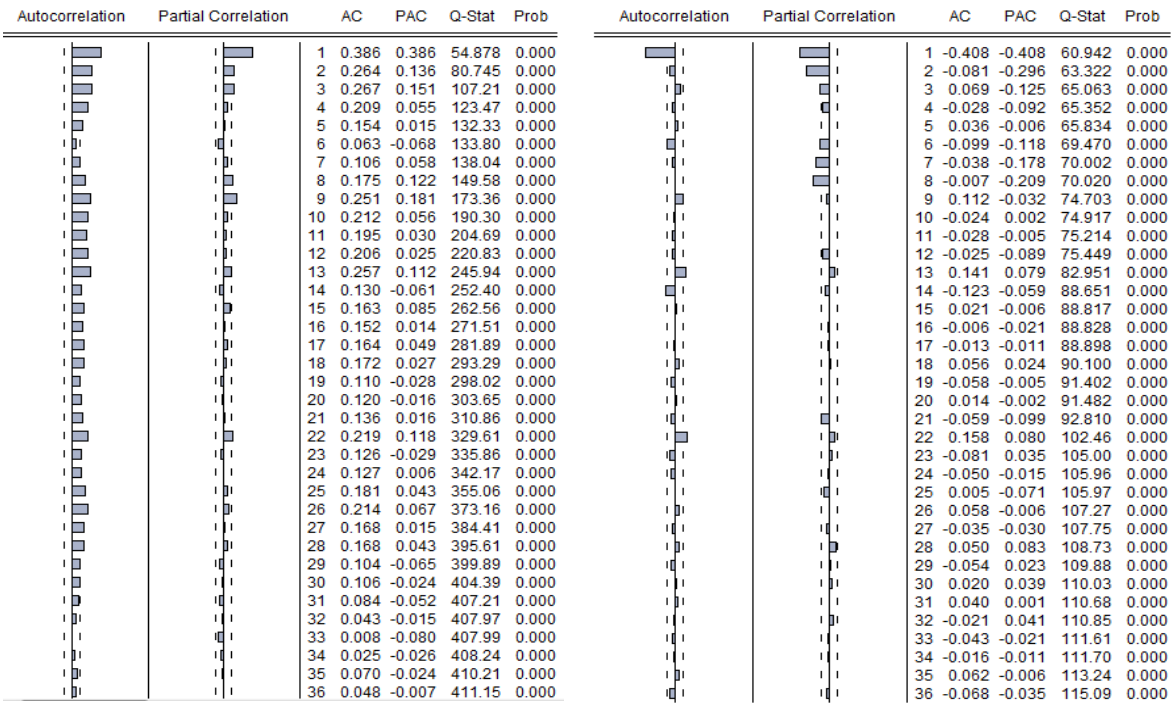
Par conséquent, la fonction d'autocorrélation de l'échantillon à lag k est :

$$\widehat{\rho_k} = \frac{\widehat{\gamma_k}}{\widehat{\gamma_0}} \tag{20}$$

Il s'agit du rapport entre la covariance de l'échantillon (à lag k) et la variance de l'échantillon. Un tracé de $\widehat{\rho_k}$ en fonction de k donne ainsi le corrélogramme de l'échantillon.

L'intérêt du corrélogramme est qu'il permet de différencier une série temporelle stationnaire d'une série non stationnaire.

La figure 3.1 représente deux corrélogrammes extraits de (Fanchette et al., 2020). Sur le corrélogramme de gauche, la colonne intitulée AC, qui est la fonction d'autocorrélation de l'échantillon, est le premier diagramme de gauche, intitulé autocorrélation. La ligne verticale pleine dans ce diagramme représente l'axe du zéro, les observations à gauche de la ligne sont des valeurs positives et celles à droite de la ligne sont des valeurs négatives. Comme le montre ce diagramme, pour un processus non-stationnaire, le coefficient d'autocorrélation commence à une valeur très élevée et diminue très lentement vers zéro à mesure que le lag augmente. De plus les autocorrélations à divers lags sont toujours du même côté de la ligne verticale représentant le zéro. C'est la structure classique d'un corrélogramme d'une série temporelle non-stationnaire. Sur le corrélogramme de droite, la ligne verticale pleine a toujours la même signification. Comme le montre clairement ce diagramme, pour un processus stochastique stationnaire, les autocorrélations à différents retards se situent autour de la ligne verticale et donc de zéro. Il s'agit de la représentation classique d'un corrélogramme d'une série temporelle stationnaire.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.386	0.386	54.878	0.000
		2	0.264	0.136	80.745	0.000
		3	0.267	0.151	107.21	0.000
		4	0.209	0.055	123.47	0.000
		5	0.154	0.015	132.33	0.000
		6	0.063	-0.068	133.80	0.000
		7	0.106	0.058	138.04	0.000
		8	0.175	0.122	149.58	0.000
		9	0.251	0.181	173.36	0.000
		10	0.212	0.056	190.30	0.000
		11	0.195	0.030	204.69	0.000
		12	0.206	0.025	220.83	0.000
		13	0.257	0.112	245.94	0.000
		14	0.130	-0.061	252.40	0.000
		15	0.163	0.085	262.56	0.000
		16	0.152	0.014	271.51	0.000
		17	0.164	0.049	281.89	0.000
		18	0.172	0.027	293.29	0.000
		19	0.110	-0.028	298.02	0.000
		20	0.120	-0.016	303.65	0.000
		21	0.136	0.016	310.86	0.000
		22	0.219	0.118	329.61	0.000
		23	0.126	-0.029	335.86	0.000
		24	0.127	0.006	342.17	0.000
		25	0.181	0.043	355.06	0.000
		26	0.214	0.067	373.16	0.000
		27	0.168	0.015	384.41	0.000
		28	0.168	0.043	395.61	0.000
		29	0.104	-0.065	399.89	0.000
		30	0.106	-0.024	404.39	0.000
		31	0.084	-0.052	407.21	0.000
		32	0.043	-0.015	407.97	0.000
		33	0.008	-0.080	407.99	0.000
		34	0.025	-0.026	408.24	0.000
		35	0.070	-0.024	410.21	0.000
		36	0.048	-0.007	411.15	0.000

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.408	-0.408	60.942	0.000
		2	-0.081	-0.296	63.322	0.000
		3	0.069	-0.125	65.063	0.000
		4	-0.028	-0.092	65.352	0.000
		5	0.036	-0.006	65.834	0.000
		6	-0.099	-0.118	69.470	0.000
		7	-0.038	-0.178	70.002	0.000
		8	-0.007	-0.209	70.020	0.000
		9	0.112	-0.032	74.703	0.000
		10	-0.024	0.002	74.917	0.000
		11	-0.028	-0.005	75.214	0.000
		12	-0.025	-0.089	75.449	0.000
		13	0.141	0.079	82.951	0.000
		14	-0.123	-0.059	88.651	0.000
		15	0.021	-0.006	88.817	0.000
		16	-0.006	-0.021	88.828	0.000
		17	-0.013	-0.011	88.898	0.000
		18	0.056	0.024	90.100	0.000
		19	-0.058	-0.005	91.402	0.000
		20	0.014	-0.002	91.482	0.000
		21	-0.059	-0.099	92.810	0.000
		22	0.158	0.080	102.46	0.000
		23	-0.081	0.035	105.00	0.000
		24	-0.050	-0.015	105.96	0.000
		25	0.005	-0.071	105.97	0.000
		26	0.058	-0.006	107.27	0.000
		27	-0.035	-0.030	107.75	0.000
		28	0.050	0.083	108.73	0.000
		29	-0.054	0.023	109.88	0.000
		30	0.020	0.039	110.03	0.000
		31	0.040	0.001	110.68	0.000
		32	-0.021	0.041	110.85	0.000
		33	-0.043	-0.021	111.61	0.000
		34	-0.016	-0.011	111.70	0.000
		35	0.062	-0.006	113.24	0.000
		36	-0.068	-0.035	115.09	0.000

Figure 3.1 Exemple d'un corrélogramme non-stationnaire (à gauche) et d'un corrélogramme stationnaire (à droite)

AC = autocorrelation, PAC = Autocorrelation partiel, Q-Stat = Q statistique, Prob = probabilité.

iii. Le test de racine unitaire

Le test de stationnarité (ou de non-stationnarité) est le test de racine unitaire. Nous allons d'abord l'expliquer puis l'illustrer et enfin considérer certaines limites de ce test. Le point de départ est le processus de racine unitaire (stochastique). On considère l'équation suivante :

$$-1 \leq \rho \leq 1 \qquad Y_t = \rho Y_{t-1} + u_t \qquad (21)$$

où u_t est le terme d'erreur de bruit blanc.

Si $\rho = 1$, c'est-à-dire dans le cas de la racine unitaire, (21) devient un modèle de marche aléatoire sans dérive qui est un processus statistique non stationnaire. Par conséquent, régresser Y_t sur la valeur retardée d'une période Y_{t-1} et vérifier si le ρ estimé est statistiquement égal à 1, reviendrait à prouver que Y_t est non stationnaire. C'est le principe du test de stationnarité. L'équation (21) peut être réécrite ainsi :

$$Y_t - Y_{t-1} = \rho Y_{t-1} - Y_{t-1} + u_t \qquad (22a)$$

$$Y_t - Y_{t-1} = (\rho - 1)\, Y_{t-1} + u_t \qquad (22b)$$

$$\Delta Y_t = \delta Y_{t-1} + u_t \qquad (23)$$

où $\delta = (\rho - 1)$ est l'opérateur de première différence.

En pratique, au lieu d'estimer (21), il est primordial d'estimer (23) et de tester l'hypothèse (nulle, $\delta = 0$). Si $\delta = 0$, alors $\rho = 1$, c'est-à-dire que nous avons une racine unitaire, ce qui signifie que la série temporelle considérée est non stationnaire.

Avant de procéder à l'estimation de (23), il est bon de noter que si $\delta = 0$, alors (23) deviendra :

$$\Delta Y_t = Y_t - Y_{t-1} = u_t \qquad (24)$$

Comme u_t est un terme d'erreur à bruit blanc, il est de ce fait stationnaire et par conséquent, cela signifie que les premières différences d'une série temporelle à marche aléatoire sont stationnaires.

Pour estimer (23), il faut maintenant prendre les premières différences de Y_t et de les régresser sur Y_{t-1} et de voir si le coefficient de pente estimé dans cette régression ($= \hat{\delta}$) est nul ou non. S'il est nul, Y_t est non stationnaire. Mais s'il est négatif, alors Y_t est stationnaire.

Parmi ces tests unitaires, les tests de Dickey-Fuller (DF), Dickey-Fuller Augmenté (ADF) et de Phillips-Perron permettent de rendre compte de la stationnarité ou non d'une série temporelle. Les deux tests qui seront abordés dans la section suivante sont ceux qui ont été utilisés dans les travaux de thèse, il s'agit des tests DF et ADF. Il faut noter que le test de Philips-Perron, qui lui permet d'intégrer en complément l'hétéroscédasticité des erreurs, ne sera pas abordé. Pour rappel, on parle d'hétéroscédasticité lorsque les variances des résidus des variables examinées sont différentes.

○ Le test de Dickey-Fuller

Dans la littérature, la statistique ou le test de tau (τ) est connu sous le nom de test de Dickey-Fuller, en l'honneur de ses découvreurs (Dickey & Fuller, 1979). Dickey et Fuller ont démontré que sous l'hypothèse nulle que δ = 0, la valeur t estimée du coefficient de Y_{t-1} dans (23) suit la statistique τ (tau). Ces auteurs ont calculé les valeurs critiques de la statistique tau sur la base de simulations de Monte-Carlo. Le tableau étant restreint, Mackinnon a préparé des tableaux plus complets, qui sont maintenant intégrés dans plusieurs logiciels économétriques tels qu'Eviews (cf. chapitre 4). Il est intéressant de noter que si l'hypothèse que δ = 0 est rejetée (par exemple dans le cas où la série temporelle est stationnaire), le test *t* habituel Student est appliqué. La procédure réelle de mise en œuvre du test DF implique plusieurs décisions, un processus stochastique peut ne pas avoir de dérive ou peut avoir une dérive ou à la fois une tendance stochastique et déterministe. Face aux différentes possibilités, le test de DF peut être estimé sous trois formes, c'est-à-dire sous trois hypothèses nulles différentes.

- Y_t est un processus aléatoire sans dérive

$$\Delta Y_t = \delta Y_{t-1} + u_t \quad (23)$$

- Y_t est un processus aléatoire avec dérive

$$\Delta Y_t = \delta Y_{t-1} + \beta_1 + u_t \quad (25)$$

- Y_t est un processus aléatoire avec dérive et une tendance stochastique

$$\Delta Y_t = \delta Y_{t-1} + \beta_1 + \beta_2 t + u_t \quad (26)$$

où t est le temps ou la tendance variable, β_1 la dérive et β_2 la tendance.

Dans tous les cas, *l'hypothèse nulle* est $\delta = 0$, c'est-à-dire qu'il y a une racine unitaire et que la série temporelle est donc non-stationnaire. L'hypothèse alternative est que $\delta < 0$, la série est alors stationnaire. Si l'hypothèse nulle est rejetée, cela signifie que Y_t est une série temporelle stationnaire avec une moyenne nulle dans le cas de (23), Y_t est stationnaire avec une moyenne non nulle dans le cas de (25) et Y_t est stationnaire autour d'une tendance déterministe dans le cas de (26).

○ Le test augmenté de Dickey-Fuller

En effectuant le test DF comme dans (23), (25), ou (26), le terme d'erreur u_t est supposé non corrélé. Mais, dans le cas où les u_t sont corrélés, Dickey et Fuller ont développé un test, connu sous le nom de test Dickey-Fuller Augmenté (ADF). Ce test est effectué en "augmentant" les trois équations précédentes en ajoutant les valeurs retardées de la variable dépendante Y_t pour plus de précisions. Le test ADF consiste ici à d'estimer la régression suivante :

$$\Delta Y_t = \beta_1 + \beta_2 t + \delta Y_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta Y_{t-i} + \epsilon_t \quad (27)$$

où ε est un terme d'erreur de bruit blanc pur et où $\Delta Y_{t-1} = Y_{t-1} - Y_{t-2}$, $\Delta Y_{t-2} = Y_{t-2} - Y_{t-3}$, etc.

Le nombre de termes de différence retardés à inclure est souvent déterminé de manière empirique, l'idée étant d'inclure un nombre suffisant de termes pour que le terme d'erreur dans (27) soit non corrélé en série. Dans le test ADF, nous testons toujours si $\delta = 0$ et ce test suit la même distribution asymptotique que la statistique DF, donc les mêmes valeurs critiques peuvent être utilisées.

3.1.2 Causalité de Granger

Bien que l'analyse de régression traite de la dépendance d'une variable par rapport à d'autres variables, elle n'implique pas nécessairement une causalité. En d'autres termes, l'existence d'une relation entre les variables ne prouve pas la causalité ou la direction de l'influence. Mais dans les régressions impliquant des données de séries chronologiques, la situation peut être quelque peu différente, car « le temps ne revient pas en arrière. En d'autres termes, si un événement A se produit avant un événement B, il est possible que A soit la cause de B. Cependant, il n'est pas possible que B soit la cause de A. En d'autres termes, les événements du passé peuvent provoquer des événements qui se produisent aujourd'hui. Les événements futurs ne le peuvent pas » (Koop, 2005).

C'est l'idée qui sous-tend le test de causalité dit de Granger (Granger, 1969). Mais il convient de noter que la question de la causalité est profondément « philosophique » et donne lieu à toutes sortes de controverses, allant d'un extrême où la causalité signifie que « tout cause tout », à l'autre extrême niant l'existence d'une quelconque causalité entre deux évènements.

Ainsi cette causalité est plus souvent appelée « causalité prédictive » (Diebold, 2001). Il avance que l'énoncé "Y_i cause Y_j" n'est qu'un raccourci pour l'énoncé plus précis, mais plus long, " Y_i contient des informations utiles pour prédire Y_j (au sens des moindres carrés linéaires), en plus de l'histoire passée des autres variables du système". Pour faire plus simple, Yi cause en sens de Granger Y_j.

3.1.3 Choix des facteurs environnementaux influençant la production PV

Modéliser statistiquement la réalité revient à décrire un phénomène via une équation mathématique. Cela implique donc deux notions, à savoir, la notion de variables explicatives et dépendantes. Une variable dépendante ou variable à expliquer est celle qui est, comme son nom l'indique, la variable à décrire, à expliquer, à prédire. Dans le cas de ces travaux de thèse, la variable dépendante qui est étudiée sera bien évidemment la puissance photovoltaïque (PV).

Les variables explicatives sont également appelées variables indépendantes, car elles sont utilisées dans le but de prédire justement cette variable dépendante. Selon le modèle employé, les variables dépendantes et indépendantes peuvent être uniques ou multiples, qualitatives ou quantitatives.

La question qui se pose est le choix des facteurs environnementaux qui doivent être étudiés et qui influencent (ou expliquent) la prévision d'une variable dépendante, ici la production PV. Ces variables seront ainsi les variables explicatives de la variable de puissance PV. Tous les facteurs environnementaux pour lesquelles des données sont accessibles n'influencent pas ou trop peu la puissance PV pour être considérés comme des variables explicatives utiles. Le test de Granger permet de choisir, de ce fait, les facteurs environnementaux qui vont être pertinents pour la suite de ces travaux.

Ainsi, le choix a été fait de garder plusieurs variables explicatives utiles. Ces dernières sont :

- Rayonnement solaire ($W.m^{-2}$)
- Température du module (°C)
- Vitesse du vent ($m.s^{-1}$)
- Humidité spécifique ($kg_{eau}/kg_{air\ humide}$)

Les données de ces paramètres seront celles recueillies au niveau des stations météorologiques au sol ainsi que celles simulées par le modèle numérique WRF (discutées dans le chapitre 3 et reprises dans les chapitres suivants). D'autres paramètres météorologiques tels que la pression atmosphérique (en bar) ne causant pas au sens de Granger la production PV n'ont donc pas été pris en compte dans la suite de ces recherches ainsi que le coefficient de salissure des modules PV (telle que la possibilité de la poussière volcanique sur l'île de La Réunion).

3.1.4 Modèle de prévision de production PV basé sur la méthode de cointégration de Johansen

3.1.4.1 Propriété de la cointégration

La méthode de cointégration dans l'analyse de régression est basée sur une hypothèse d'incréments stationnaires avec un décalage temporel fixe appelé I(d). Cette terminologie et cette notation ont été établies dans les précédents paragraphes. Le développement de la technique de cointégration est basé sur l'intégration I(d) pour déduire des relations d'équilibre à court et à long terme entre des variables non stationnaires via l'analyse de régression (Hamisultane, 2002). Il convient de souligner ici que la régression d'une série temporelle non stationnaire (sur une autre série temporelle non stationnaire) peut produire une régression fallacieuse. Une façon de s'en prémunir est de déterminer si les séries temporelles sont cointégrées. Une combinaison de deux ou plusieurs séries non stationnaires individuelles non stationnaires peut donner lieu à une série stationnaire. Le concept et les conditions de la cointégration peuvent être résumés comme suit : La cointégration de deux ou plusieurs séries temporelles suggère qu'il existe une relation à long terme, ou d'équilibre, entre-elle. Les deux conditions de cointégration sont, premièrement, que ces séries doivent être intégrées d'ordre d I(d). Deuxièmement, une combinaison linéaire de ces séries permet de réduire la série à un ordre d'intégration inférieur. Afin de réconcilier le comportement à court terme avec son comportement à long terme, un mécanisme de correction d'erreur a été développé par Engle & Granger (EG) (Engle & Granger, 1991).

3.1.4.2 Modèle de correction d'erreur

Si deux séries sont cointégrées Y_t et X_t (telle que $Y_t - \hat{\alpha}X_t - \beta \sim > I(0))$, il est possible d'estimer le modèle à correction d'erreur (MCE) suivant :

Avec $\delta < 0$ $$\Delta Y_t = \gamma \Delta X_t + \delta(Y_{t-1} - \alpha X_{t-1} - \beta) + \nu_t \quad (28)$$

Le paramètre δ doit être négatif pour qu'il y ait un retour de Y_t à sa valeur d'équilibre sur le long terme, $\alpha X_{t-1} + \beta$. En effet, il existe une force de rappel vers l'équilibre de long terme qu'à condition que $\delta < 0$.

Ainsi, les MCE permettent de modéliser en même temps les dynamiques de court terme (variables en différences premières) et de long terme (variables en niveau).

La dynamique de court terme peut s'écrire :

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 X_t + \alpha_3 X_{t-1} + \nu_t \quad (29)$$

La dynamique de long terme, comme la cointégration des deux séries suggère une relation d'équilibre à long terme, peut s'exprimer de la façon suivante :

$$Y_t = \alpha X_t + \beta + \epsilon_t \quad (30)$$

Cette dernière (30) est déduite de (29) en utilisant sur le long terme $Y_{t-1} = Y_t$ et $X_{t-1} = X_t$.

Ainsi l'estimation du MCE avec une seule variable explicative peut être résumée à deux étapes clé :

Etape 1 : Estimation par la méthode des Moindres Carré Ordinaires (MCO) de la relation de long terme.

Etape 2 : Estimation par les MCO de la relation du modèle dynamique à court terme.

L'inconvénient de la méthode d'EG est qu'elle ne permet pas de distinguer plusieurs relations de cointégration. Pour un nombre N de variables explicatives N>2, il ne peut y avoir que (N-1) relations de cointégration. La méthode d'EG ne permet d'obtenir qu'une relation de cointégration. C'est pour cela, que pour pallier à cette difficulté, Johansen (Johansen, 1988) a proposé une approche multivariée de la cointégration dans le cas où il y avait plusieurs variables explicatives N>2.

3.1.4.3 Cointégration de Johansen à modèle de correction d'erreurs vectoriel (MCEV)

Le modèle de cointégration de Johansen à modèle de correction d'erreurs vectoriel (MCEV) est utilisé dans les cas rares où une cointégration entre plusieurs variables est nécessaire. Utilisé généralement en Économétrie (Cuthbertson et al., 1992), ce modèle a été transposé à un autre domaine d'expertise : de l'estimation et de la prévision PV.

Il convient de rappeler que si des séries non stationnaires sont intégrées du premier ordre I(1) et se révèlent être cointégrées, un modèle de correction d'erreurs vectoriel (MCEV) peut

être utilisé afin de permettre l'examen de la dynamique à court terme et à long terme des séries de cointégration. La méthode ci-après est appliquée dans le chapitre 4 à des données météorologiques réelles (cf. section 4.5).

En considérant un vecteur Y_t contenant N variables non stationnaires I(1), il est possible de représenter le Vecteur d'Auto Régression (VAR) de Y_t :

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \epsilon_t \quad (31)$$

où le rang de la matrice est respectivement, Y_t (N, 1), A_1 (N, N), Y_{t-1} (N, 1), …, A_p (N, N), Y_{t-p} (N, 1), ϵ_t (N, 1) et p le retard

Dans le cas d'un modèle VAR(2) composé de 4 variables ($Y_{1t}, Y_{2t}, Y_{3t}\ et\ Y_{4t}$) de Y_t avec p = 2 retards, s'écrit de la façon suivante :

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + \epsilon_t \quad (32)$$

La forme matricielle de (32) s'écrit alors :

$$\begin{pmatrix} Y_{1t} \\ Y_{2t} \\ Y_{3t} \\ Y_{4t} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \begin{pmatrix} Y_{1t-1} \\ Y_{2t-1} \\ Y_{3t-1} \\ Y_{4t-1} \end{pmatrix} + \begin{pmatrix} a_{15} & a_{16} & a_{17} & a_{18} \\ a_{25} & a_{26} & a_{27} & a_{28} \\ a_{35} & a_{36} & a_{37} & a_{38} \\ a_{45} & a_{46} & a_{47} & a_{48} \end{pmatrix} \begin{pmatrix} Y_{1t-2} \\ Y_{2t-2} \\ Y_{3t-2} \\ Y_{4t-2} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \\ \epsilon_{4t} \end{pmatrix} \quad (33)$$

Le système d'équation devient le suivant :

$$\begin{aligned} Y_{1t} &= a_{11}Y_{1t-1} + a_{12}Y_{2t-1} + a_{13}Y_{3t-1} + \cdots + a_{17}Y_{3t-2} + a_{18}Y_{4t-2} + \epsilon_{1t} \\ Y_{2t} &= a_{21}Y_{1t-1} + a_{22}Y_{2t-1} + a_{23}Y_{3t-1} + \cdots + a_{27}Y_{3t-2} + a_{28}Y_{4t-2} + \epsilon_{2t} \\ Y_{3t} &= a_{31}Y_{1t-1} + a_{32}Y_{2t-1} + a_{33}Y_{3t-1} + \cdots + a_{37}Y_{3t-2} + a_{38}Y_{4t-2} + \epsilon_{3t} \\ Y_{4t} &= a_{41}Y_{1t-1} + a_{42}Y_{2t-1} + a_{43}Y_{3t-1} + \cdots + a_{47}Y_{3t-2} + a_{48}Y_{4t-2} + \epsilon_{4t} \end{aligned} \quad (34)$$

Ce modèle VAR (2) de première différence peut être réécrit sous la forme d'un modèle à correction d'erreurs vectoriel (MCEV) et en fonction de Y_{t-1} comme suit :

$$Y_t - Y_{t-1} = A_1 Y_{t-1} - Y_{t-1} + A_2 Y_{t-2} + A_2 Y_{t-1} - A_2 Y_{t-1} + \epsilon_t \quad (35)$$

$$\Delta Y_t = (A_1 - I) Y_{t-1} - A_2 (Y_{t-1} - Y_{t-2}) + A_2 Y_{t-1} + \epsilon_t \quad (36)$$

$$\Delta Y_t = -A_2 \Delta Y_{t-1} + (A_1 + A_2 - I) Y_{t-1} + \epsilon_t \quad (37)$$

$$\Delta Y_t = -A_2 \Delta Y_{t-1} + \Pi Y_{t-1} + \epsilon_t \quad (38)$$

où $\Pi = A_1 + A_2 - I$ et I est la matrice unitaire

Si nous posons $\Pi = \alpha \beta'$ avec α une matrice (N, r) avec r < N contenant les vitesses d'ajustement pour chaque vecteur de cointégration et β' une matrice (r, N) comprenant r

relations de cointégration (avec 0 < r < N) afin de mettre en évidence un modèle MCEV. Soit r = 4, alors :

$$\begin{pmatrix} \Delta Y_{1t} \\ \Delta Y_{2t} \\ \Delta Y_{3t} \\ \Delta Y_{4t} \end{pmatrix} = -\begin{pmatrix} a_{15} & a_{16} & a_{17} & a_{18} \\ a_{25} & a_{26} & a_{27} & a_{28} \\ a_{35} & a_{36} & a_{37} & a_{38} \\ a_{45} & a_{46} & a_{47} & a_{48} \end{pmatrix}\begin{pmatrix} \Delta Y_{1t-1} \\ \Delta Y_{2t-1} \\ \Delta Y_{3t-1} \\ \Delta Y_{4t-1} \end{pmatrix} + \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \\ \alpha_{31} & \alpha_{32} \\ \alpha_{41} & \alpha_{42} \end{pmatrix}\begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} \\ \beta_{21} & \beta_{22} & \beta_{23} & \beta_{24} \end{pmatrix}\begin{pmatrix} Y_{1t-2} \\ Y_{2t-2} \\ Y_{3t-2} \\ Y_{4t-2} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \\ \epsilon_{4t} \end{pmatrix} \quad (39)$$

Pour estimer un modèle MCEV, le rang de la matrice doit être égal à r, ce qui signifie que Π a r valeurs propres non nulles et donc β' aussi.

Le test et la stratégie d'estimation de Johansen qui est un test de maximum de vraisemblance permet d'estimer tous les vecteurs de cointégration pour N variables, qui ont toutes des racines unitaires et il y a au plus N-1 vecteurs de cointégration. Le test de Johansen fournit des estimations de tous les vecteurs de cointégration si la relation de cointégration existe, et un test de rang est utile. Ainsi, trois cas peuvent se présenter, si :

- Rang (Π) = 0, alors r = 0 signifie qu'il n'y a pas de relation de cointégration et que le MCEV ne peut pas être appliqué,
- Rang (Π) = r, signifie que les variables sont cointégrées et que le nombre de relations de cointégration est égal à r. Le modèle MCEV peut être estimé.
- Rang (Π) = N, ce qui signifie qu'il n'y a pas de relation de cointégration.

La procédure de Johansen est basée sur les tests de valeur d'Eigen maximale et de la trace qui sont effectués sur la base du modèle à correction d'erreurs. Pour les deux tests statistiques, le test initial de Johansen est un test d'hypothèse nulle d'absence de cointégration contre l'alternative de cointégration.

Le premier test de valeur d'Eigen maximale consiste à déterminer si le rang de la matrice est nul, et l'hypothèse nulle est rang (Π) = 0 alors que l'hypothèse alternative est rang (Π) = 1.

Le deuxième test de la trace consiste à déterminer si le rang de la matrice est r0, l'hypothèse nulle est rang (Π) = r_0 et l'hypothèse alternative est que r_0 < rang (Π) ≤ r, où r est le nombre maximum de vecteurs de cointégration possibles.

Une description plus détaillée est disponible dans les ouvrages de spécialité (Johansen, 1988, 1991, 2015). Cependant, dans la suite de cette étude, nous résumons l'ordre de test de Johansen qui sera appliqué en cinq étapes :

Étape 1 : Réalisation de tests de stationnarité des séries (corrélogramme et ADF) pour déterminer s'il existe ou non une relation de cointégration.
Étape 2 : Si l'étape 1 est vraie, cela signifie que les séries sont du même ordre d'intégration et que la cointégration est probable, donc le modèle MCEV peut être estimé. Détermination de la longueur du lag en utilisant les critères d'Akaike et de Schwarz.
Étape 3 : Mise en œuvre du test de Johansen pour déterminer le nombre de relations de cointégration.
Étape 4 : Identification des relations de cointégration ou des relations à long terme entre les variables.
Étape 5 : Estimation du modèle MCEV par la méthode du maximum de vraisemblance, tests de validation par diagnostic visuel ou corrélogramme, et vérification que les résidus du modèle sont des bruits blancs.

La méthode statistique de Johansen qui est un modèle fortement utilisé en Économétrie a été adaptée de manière originale à de la prévision météorologique et PV. Le processus est basé sur la cointégration de variables explicatives (les paramètres météorologiques et physiques) qui influent et améliorent de manière pertinente une variable expliquée, la production PV, et sa prévision. Toutefois, bien que la méthode soit novatrice (cf. chapitre 4), son point faible réside dans son temps de calcul élevé face à la masse de données importantes en entrée nécessaires à la prévision du modèle. Il devient alors nécessaire de faire appel un autre outil capable de traiter de manière efficace ces données sous forme de séries temporelles. Un modèle neuronal est ainsi choisi : le modèle LSTM.

3.2 Modèle neuronal

Afin d'augmenter en précision le modèle de prévision de Johansen, il serait intéressant d'étudier des méthodes d'intelligences artificielles, par exemple des réseaux de neurones (RN). Toutefois, le choix du bon modèle et de la structure neuronale adéquate passe par plusieurs étapes essentielles :

- La sélection des variables d'entrées / de sorties et le traitement des données nécessaires au préalable à l'utilisation de ces données c'est-à-dire le nettoyage de valeurs aberrantes, la normalisation, etc. constitue la première étape.
- La deuxième étape concerne la détermination d'un type de réseau de neurones en adéquation à l'application recherchée. Cette structure s'accompagne en général d'un algorithme d'optimisation des hyperparamètres (paramètres de la structure) afin d'améliorer les performances du modèle.
- La dernière étape, indispensable, consiste à ajuster les poids neuronaux. Un algorithme d'apprentissage associe, pour cela, le modèle identifié préalablement aux données d'entrées et de sorties.

Par la suite, on présentera le modèle RN choisi en fonction des données utilisées en entrée et en sortie de l'application que nous voulons en faire c'est-à-dire, ici, la prévision PV à partir de données météorologiques et paramètres physiques sous forme de séries temporelles.

3.2.1 Présentation des réseaux de neurones

Le modèle RN ou Artificial Neural Networks (ANN) est inspiré directement par les neurones biologiques. Dans un réseau de neurones biologiques, plusieurs neurones travaillent ensemble, reçoivent des signaux d'entrée, traitent l'information et déclenchent un signal de sortie. La figure 3.2 représente la structure d'un neurone artificiel. Ce neurone est une unité de calcul qui prend certaines entrées et un terme d'interception dont la valeur est généralement 1. Chaque entrée et le terme d'interception sont multipliés par un poids et ajoutés ultérieurement. Le poids du terme d'interception est appelé biais. Ensuite, une fonction de transfert, également appelée fonction d'activation, est appliquée au résultat. Sans cette fonction de transfert, ce neurone ne serait qu'une fonction linéaire. Par conséquent, la fonction de transfert donne aux réseaux neuronaux la capacité de résoudre des problèmes non linéaires (Royo, 2019).

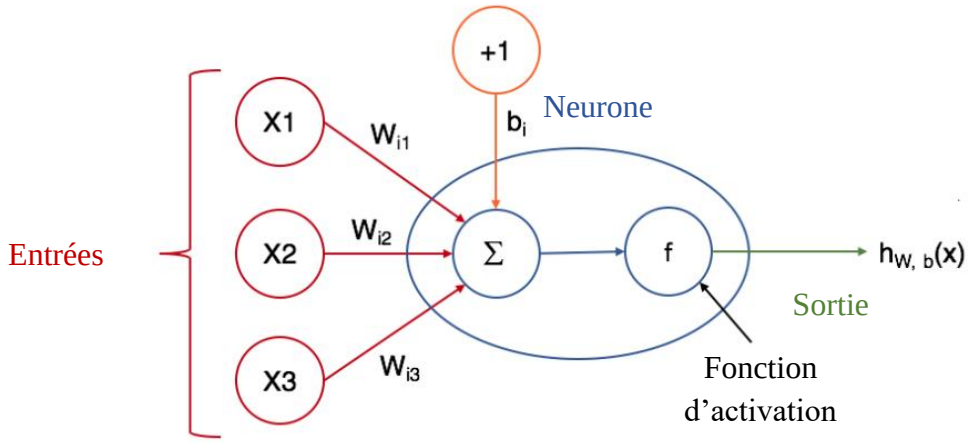


Figure 3.2 Structure d'un neurone artificiel

Ce neurone a pour formule de sortie :

$$h_{w,b}(X) = f(w_{i1}x_1 + w_{i2}x_2 + w_{i3}x_3 + b_i) \quad (40)$$

où $h_{w,b}$ est la sortie du neurone, X le vecteur d'entrée, x_j l'entrée j, f la fonction d'activation du neurone i, w_{ij} le poids j du neurone i et b_i le biais du neurone i.

L'objectif des modèles RN n'est pas de reproduire parfaitement un neurone biologique, mais d'imiter ses principales fonctions que sont le transport et le traitement de l'information. Le traitement est réalisé par une fonction d'activation f(x) où i représente le neurone tandis que le transport est effectué par l'intermédiaire d'interconnexions pondérées par des coefficients w. La structuration de plusieurs neurones forme un réseau de neurones. Ce dernier est composé de plusieurs éléments principaux :

- Les neurones qui prennent une donnée d'entrée pour produire une donnée de sortie. Un certain nombre de neurones sont groupés en couches (ou layers). Tous les neurones de la même couche remplissent un type de fonction similaire. Il existe trois types de couches comme illustré sur la figure 3.3:
  - Couche d'entrée (input layer)
  - Couche cachée (hidden layer)
  - Couche de sortie (output layer)
- Les poids et biais sont des variables du modèle qui sont mis à jour de manière continue afin d'améliorer la précision du RN. Un poids est appliqué à l'entrée de chacun des neurones permettant de calculer une donnée de sortie. Les biais, quant à eux, sont également des valeurs numériques qui sont ajoutées une fois que les poids sont appliqués aux valeurs d'entrée. Poids et biais jouent le rôle d'auto-apprentissage dans les RN.
- La fonction d'activation normalise la donnée de sortie avant d'être transmise aux neurones suivants. Lorsque les neurones calculent la somme pondérée des valeurs d'entrée et du biais, elles sont transmises à la fonction d'activation, qui vérifie si la valeur calculée est supérieure au seuil requis. Si cela est avéré, la fonction d'activation est activée et la valeur de sortie est calculée. Parmi ces fonctions d'activation, les plus

courantes sont la sigmoïde, la tangente hyperbolique et l'unité linéaire rectifiée (ReLU (Le et al., 2015)).

- Couche d'entrée Couche cachée Couche de sortie

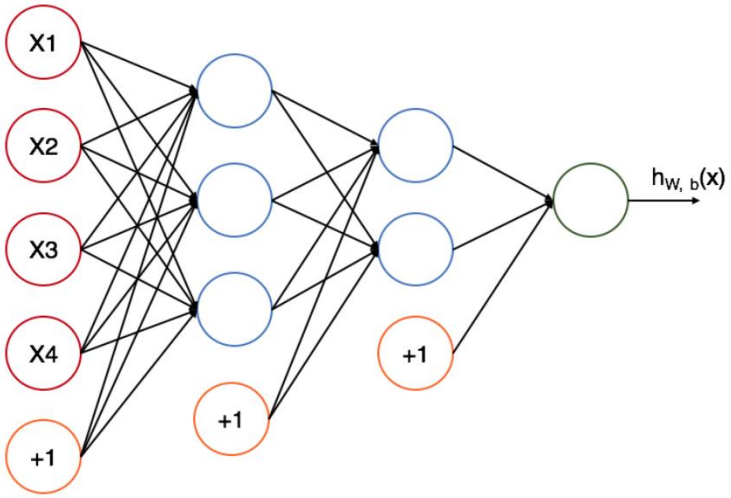


Figure 3.3 Structure d'un réseau de neurones. Source :(Lin-Kwong-Chon, 2020)

Un parallèle entre le monde biologique et le monde algorithmique peut être fait. Les dendrites d'un neurone qui sont les récepteurs de l'information. Elles se développent selon une structure arborescente qui définit la connectivité propre du neurone et du réseau de neurones et constituent la pondération des interconnexions en entrée. Le corps cellulaire du neurone ou soma est un élément post-synaptique qui réalise la sommation de toutes les entrées x(t). L'axone est une fibre nerveuse en sortie du neurone qui renferme la fonction d'activation $f_n(x)$ qui transmet ainsi l'information au neurone suivant dès que le franchissement d'un seuil de sommation est atteint (Lin-Kwong-Chon, 2020).

3.2.2 Réseaux de neurones récurrents

Il existe de nombreux modèles neuronaux. On compte parmi les plus couramment utilisés le modèle Perceptron multicouche (MLP) (Hornik, 1989), le modèle convolutif (CNN) (Fukushima, 1980), le modèle à réservoir (RC) (Jaeger, 2000) ou encore le modèle à capsules (CapsNet) (Goldani et al., 2021) dont l'intérêt et l'architecture sont adaptés à l'application recherché. Toutefois, le réseau neuronal récurrent (RNN) est une autre classe de réseaux neuronaux artificiels spécialement conçue pour modéliser les données séquentielles ou les séries chronologiques, tant pour la classification (Buber and Diri, 2019; Gill and Khehra, 2021; Smirnov and Nguifo, 2018) que pour la prédiction (Barman & Boruah, 2018; Y. Liu et al., 2020). Les données de séries temporelles contiennent des informations temporelles intrinsèques, qui ne peuvent pas être capturées par un réseau neuronal simple (Hüsken & Stagge, 2003). Contrairement à un réseau neuronal simple, les RNN sont renforcés par une arête de pas de temps supplémentaire qui introduit une notion de temps dans le réseau neuronal (Yu et al., 2018). Les bords nommés récurrents connectent les étapes adjacentes pour former un cycle d'auto-connexions d'un neurone à lui-même (Rahman et al., 2018). Ces boucles autoconnectées représentent les différentes étapes temporelles. La figure 3.4 présente l'architecture de base d'une unité récurrente. Un vecteur d'état caché (ht) fixé à la valeur zéro

à l'étape initiale est connecté à chaque unité cachée. Il a la même longueur que le nombre d'entrées et stocke les informations utiles observées dans le passé. L'état caché à l'instance t utilise les connexions de rétroaction pour rappeler le vecteur d'état caché à l'instance t-1 temps. De cette façon, le vecteur d'état caché à l'instance temporelle précédente ainsi que l'entrée actuelle (x_t) sont utilisés pour calculer l'état caché à l'instance temporelle t. Par conséquent, la sortie finale (y_t) est influencée à la fois par l'entrée actuelle et par les informations stockées précédemment (Kumari & Toshniwal, 2021).

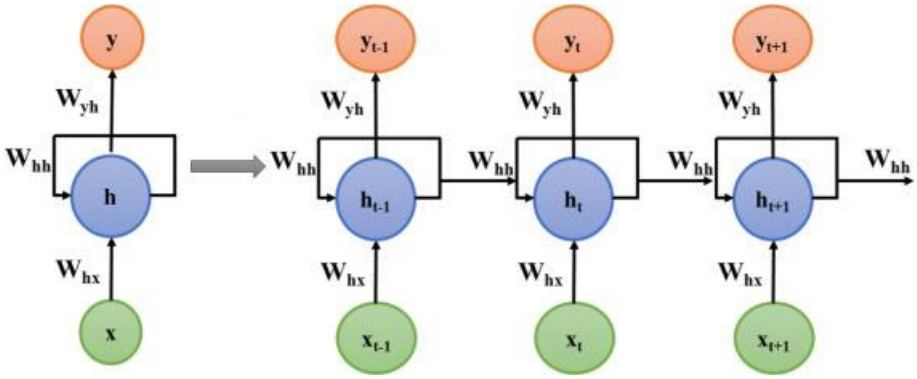


Figure 3.4 L'architecture « pliée » et « dépliée » d'un réseau de neurones récurrents (Kumari & Toshniwal, 2021)

Les équations suivantes représentent mathématiquement le processus des RNN :

$$h_t = f(w_{hx}x_t + w_{hh}h_{t-1} + b_h) \quad (41)$$

$$\widehat{y_t} = f(w_{yh}h_t + b_y) \quad (42)$$

où f représente la fonction d'activation, w_{hx} et w_{yh} représentent la matrice poids entre la couche d'entrée et la couche cachée et la couche cachée et elle-même à des pas de temps antérieurs, respectivement, b_h et b_y représentent le vecteur biais.

Ainsi, les neurones récurrents possèdent, en plus de leurs entrées, une information temporelle de leurs états précédents. Cette caractéristique leur confère une aptitude équivalente à une mémoire volatile qui est mise à jour à chaque présentation de nouvelles entrées. Ces neurones sont soumis à d'importantes pertes d'informations temporelles. L'intégration de l'état précédent n'affecte que l'état futur, ce qui implique la perte progressive des informations anciennes au cours du temps. Comparés aux neurones récurrents, les neurones de mémoire (« Long Short-Term Memory cells » ou neurone LSTM) possèdent en plus, des entrées et des sorties, des informations temporelles sur leurs états. Le calcul de ce neurone implique donc trois flux d'arrivées : les entrées, les états et les sorties. Ces flux ou portes (« gate ») sont gérés par des paramètres mémoriels, des coefficients avec une intensité comprise entre 0 et 1 qui permettent de réguler l'arrivée des informations.

3.2.3 Modèle à mémoire court et long terme

Le modèle à mémoire court et long terme (LSTM) est décrit comme une version améliorée et étendue des RNN qui a été appliquée avec succès aux difficultés de prévision de

séries temporelles (Qing & Niu, 2018). Il a été signalé que les réseaux RNN sont incapables de traiter les dépendances à long terme dans les données en raison de la disparition du gradient et du problème d'explosion du gradient (B. Hochreiter et al., 2009). Cependant, ce problème a été résolu avec l'introduction des réseaux LSTM introduits par Sepp Hochreiter et Jürgen Schmidhuber (S. Hochreiter & Schmidhuber, 1997). L'architecture des LSTM aborde le problème de la disparition du gradient en incorporant des "portes" autoconnectées dans les unités cachées, ce qui régulent le flux d'informations dans le réseau. Sur de longues périodes de temps, ces portes permettent aux cellules LSTM de lire, écrire et oublier les informations de la mémoire. Cette caractéristique permet aux cellules de conserver les données importantes et d'« oublier » les informations inutiles. Un parallèle peut être fait avec la mémoire sélective humaine. La figure 3.5 montre l'architecture de base d'une cellule LSTM et la diffusion de l'information dans ce réseau. Elle contient les quatre portes suivantes : les portes d'oubli (« forget gate ») (f_t), porte d'entrée (i_t), porte de mise à jour (« update gate ») (g_t) et porte de sortie (o_t). Les portes reçoivent une entrée de la même sortie de l'unité LSTM obtenue au pas de temps précédent (h_{t-1}). Les portes reçoivent également des données d'entrée liées au pas de temps actuel (x_t). Les équations régissant ce neurone sont données ci-dessous.
La porte d'oubli est l'élément principal de l'architecture LSTM. Elle contrôle l'information qui doit être retirée de la cellule de mémoire. Elle applique une fonction sigmoïde (σ) qui prend une valeur comprise entre 0 (oublie tout) à 1 (conserve tout) sur la sortie de l'état précédent (h_{t-1}) et les données d'entrée (x_t) :

$$f_t = \sigma\big(w_f.[h_{t-1}, x_t] + b_f\big) \tag{43}$$

La porte d'entrée utilise également une fonction sigmoïde pour décider des valeurs à écrire, tandis que la porte de mise à jour utilise la fonction d'activation tangente hyperbolique ($tanh$) pour créer de nouvelles valeurs de cellules, comme indiqué ci-dessous :

$$i_t = \sigma(w_i.[h_{t-1}, x_t] + b_i) \tag{44}$$

$$g_t = \tanh\big(w_g.[h_{t-1}, x_t] + b_g\big) \tag{45}$$

L'état de cellule précédent (C_{t-1}) interagit avec les valeurs de la porte d'oubli et de la porte de mise à jour pour mettre à jour le nouvel état de la cellule (C_t) comme indiqué :

$$C_t = f_t * C_{t-1} + i_t * g_t \tag{46}$$

Enfin, la porte de sortie régule la sortie d'une cellule en utilisant une fonction σ. Elle est combinée avec l'état d'une cellule, qui est activée par la fonction d'activation ϕ (qui est souvent une fonction tangente hyperbolique ou plus rarement une fonction unité linéaire rectifiée ReLU) pour obtenir la sortie finale, h_t comme indiqué :

$$o_t = \sigma(w_o.[h_{t-1}, x_t] + b_o\ ) \tag{47}$$

$$h_t = o_t.\ \phi(C_t) \tag{48}$$

où les paramètres et hyperparamètres d'un tel modèle sont les suivants : σ et ϕ sont des fonctions d'activation, w_{xn} et w_{hn} sont les matrices poids associés au signal d'entrée et de sortie des cellules précédentes, b_n représentent les biais des différentes portes de régulations avec $n \epsilon (i, g, f, o)$.

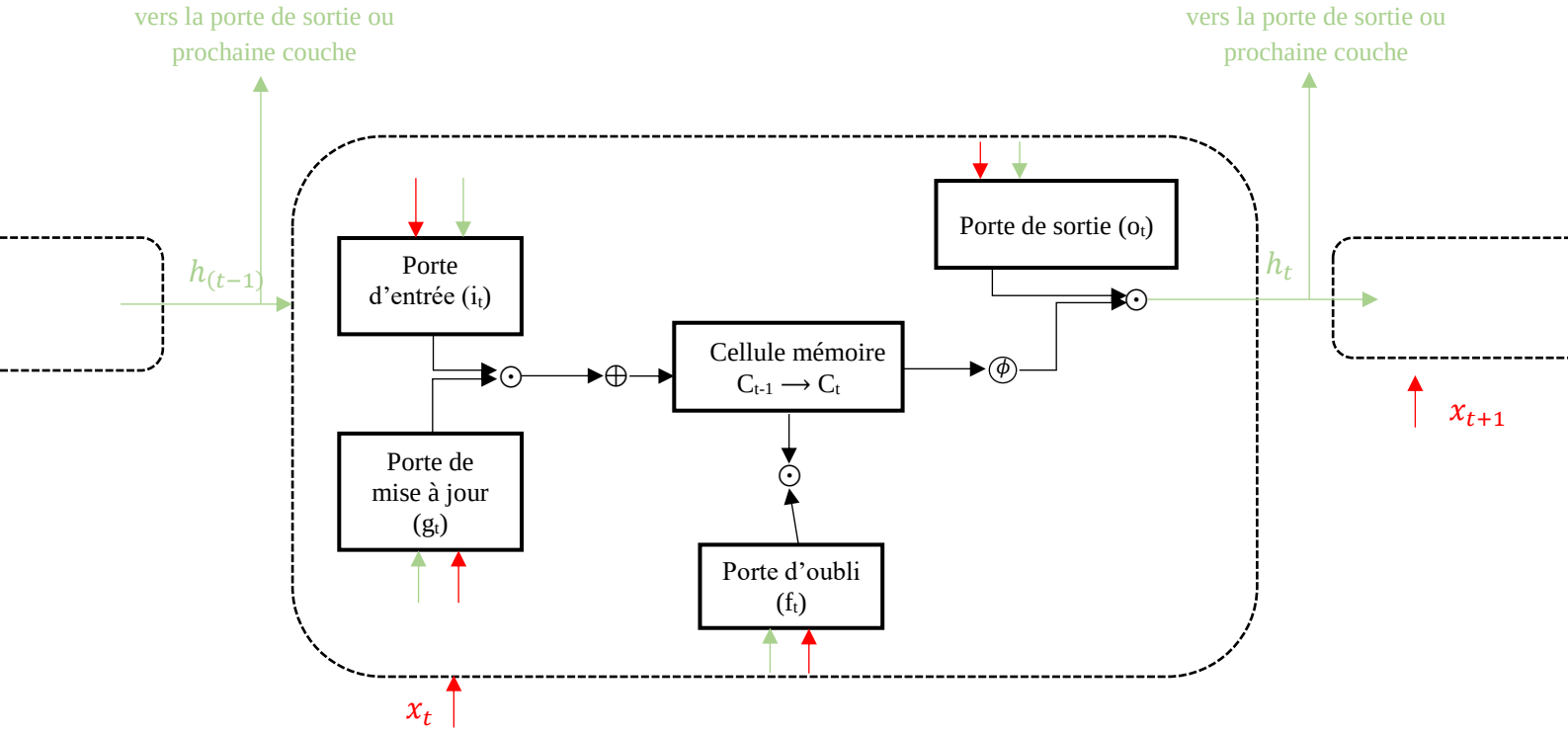


Figure 3.5 Une unité LSTM pour le pas de temps t

N.B. Les symboles ⊙ et ⊕ représentent respectivement la multiplication scalaire et la fonction somme.

Parmi les modèles neuronaux existants concernant la prévision dans la littérature, le choix s'est ainsi naturellement porté sur le modèle LSTM de par sa capacité à prévoir des séries temporelles (et dans notre cas, des séries temporelles de production PV) tout en traitant les dépendances à long termes des données, mais aussi sa capacité à traiter une masse importante de données tout en conservant les informations les plus utiles dans le processus d'apprentissage et de prévision via ses portes d'oubli. La prochaine section est un état de l'art concernant les modèles LSTM.

3.2.4 État de l'art sur les modèles LSTM dans la prévision de production photovoltaïque

Le choix de la modélisation de la prévision de la production PV qui est le cœur de ces travaux de thèse se porte désormais sur un modèle à mémoire court et long terme. Toutefois, dans la littérature, il existe un certain nombre de travaux concernant l'utilisation de modèles LSTM comme outil de prévision de production PV. Un état de l'art centré des RNN et LSTM est proposé dans cette section. Trois types de structures peuvent être distingués :

- Les modèles RNN dont la structure est proche des structures de type LSTM
    - Le modèle Gated Recurrent Unit (GRU)
    - Le modèle Bidirectionnel GRU (Bi-GRU)
    - Le modèle neuronal convolutif ou Convolutional Neural Network (CNN)
- Les variations du modèle LSTM
    - Le modèle Bidirectionnel LSTM (Bi-LSTM)
    - Le modèle LSTM multiplicatif (mLSTM)
- Les modèles hybrides
    - Le modèle hybride CNN-LSTM
    - Le modèle hybride LSTM-RNN
    - Le modèle LSTM basé sur le mécanisme d'attention (aLSTM)

Il faut savoir qu'il est difficile de comparer objectivement les performances des modèles entre eux. En effet, les critères d'erreurs employés dans les différentes études, les horizons temporels de prévision, le type de données d'entrée et les données météorologiques choisies ne sont pas équivalents. Malgré cela, les performances de chacun pourront être analysées et comparées aux différents besoins énoncés (dans le chapitre précédent).

- **Modèles dont la structure est proche des structures de type LSTM**

Le modèle Gated Recurrent Unit (GRU) a été proposé par (Cho et al., 2014) comme une architecture RNN plus simple que LSTM, ce qui facilite le calcul et la mise en œuvre. Le GRU est similaire au LSTM en termes de mémorisation des informations importantes et de capture des dépendances à long terme. La force de GRU est que le temps de calcul est plus efficace avec moins de complexité en raison de moins de paramètres que LSTM (F. Wang et al., 2018). Le GRU ne possède également que deux portes, à savoir une porte de réinitialisation (« reset gate ») et une porte de mise à jour (« update gate »). La porte de mise à jour agit exactement de la même manière que la porte d'oubli et d'entrée du LSTM, car elle décide des informations à stocker et celles à effacer. Pendant ce temps, la porte de réinitialisation décide de la quantité d'informations qui doivent être oubliées. Par conséquent, le temps de formation de GRU est plus rapide que celui de LSTM. De par son efficacité, il est de plus en plus utilisé dans la littérature.

Au même titre que le modèle GRU (Schuster & Paliwal, 1997; She & Jia, 2021), le modèle Bidirectionnel GRU (Bi-GRU) est une approche améliorée de GRU où l'architecture est identique à celle d'un Bi-LSTM, en ce sens qu'elle est constituée de deux couches cachées séparées et moins de paramètres sont nécessaires .

Le modèle neuronal convolutif ou Convolutional Neural Network (CNN) (Fukushima, 1980) est un algorithme d'apprentissage profond qui prend en compte les entrées spatiales. Identique aux autres réseaux neuronaux, les neurones CNN ont des poids et des biais ajustables. Cependant, le CNN est principalement utilisé pour traiter des données avec une topologie en grille, ce qui lui confère une caractéristique spécifique de son architecture (Lecun & Bengio, 1995). Le CNN est un réseau feedforward, car le flux d'information se produit dans une seule direction, c'est-à-dire, de leurs entrées vers leurs sorties (Rawat & Wang, 2017). Le modèle CNN utilise trois couches principales, à savoir la couche de convolution, la couche de mise en

commun des sorties (dite de « pooling ») et la couche entièrement connectée (dite « fully connected »). Les couches convolutives et de mise en commun sont utilisées pour réduire la complexité du calcul tandis que le couche entièrement connectée permet de produire un nouveau vecteur de sortie. Diverses techniques de mise en commun sont disponibles dans l'architecture de CNN. Cependant, la mise en commun maximale est surtout utilisée dans les couches CNN, où la fenêtre de mise en commun contient la valeur maximale de chaque élément (Rehman et al., 2019).

- **Les variations du modèle LSTM**

Le modèle LSTM unidirectionnel utilise les informations précédentes pour déduire les informations de suivi, tandis que le modèle Bi-LSTM utilise deux réseaux neuronaux LSTM pour prendre pleinement en compte les informations des données avant et après. Dans les séries chronologiques de données sur la vitesse du vent, compte tenu de la loi de changement des données avant et après, (Liang et al., 2021) ont utilisé la prévision bidirectionnelle de la vitesse du vent pour améliorer la précision de la prévision de la vitesse du vent. Le modèle Bi-LSTM comprend un calcul avant et un calcul arrière. La direction horizontale représente le flux bidirectionnel de la série chronologique, et la direction verticale représente le flux unidirectionnel de la couche d'entrée à la couche cachée et de la couche cachée à la couche de sortie. Le réseau neuronal Bi-LSTM utilise deux réseaux neuronaux LSTM, qui calculent le vecteur caché $\overrightarrow{h_t}$ dans la direction avant et calculent le vecteur caché $\overleftarrow{h_t}$ dans la direction arrière. En combinant les résultats de sortie de la séquence d'entrée avant et de la séquence d'entrée arrière, alors y_t est le résultat de sortie final :

$$\vec{h}_{\mathrm{t}} = \mathrm{LSTM}(\mathrm{x}_{\mathrm{t}}, \vec{h}_{t-1}) \tag{49}$$

$$\overleftarrow{\mathrm{h}}_{\mathrm{t}} = \mathrm{LSTM}(\mathrm{x}_{\mathrm{t}}, \overleftarrow{\mathrm{h}}_{t-1}) \tag{50}$$

$$y_t = g(W_{\vec{h}_{\mathrm{y}}} \vec{h}_{\mathrm{t}} + W_{\overleftarrow{\mathrm{h}}_{\mathrm{y}}} \overleftarrow{\mathrm{h}}_{\mathrm{t}} + b_y) \tag{51}$$

(Zhen et al., 2021) ont proposé un nouveau modèle de prédiction de la puissance PV à court terme basé sur le modèle amélioré de mémoire bidirectionnelle à long terme avec un algorithme génétique (GA-BiLSTM) pour améliorer les performances et de multiples séries de sorties PV sont prises de manière innovante comme entrées du modèle de prédiction.

Le modèle LSTM multiplicatif (mLSTM) (Krause et al., 2017) est une architecture améliorée de réseau neuronal récurrent pour la modélisation de séquences qui combine la transition « hidden to hidden » du réseau neuronal récurrent multiplicatif avec le cadre de contrôle de la mémoire à long terme (LSTM). Les architectures mRNN et LSTM peuvent être combinées en ajoutant des connexions de l'état intermédiaire du mRNN à chaque porte dans une unité LSTM.

Les avantages de cette architecture sont de combiner les transitions flexibles dépendant de l'entrée des mRNNs avec le long délai et le contrôle de l'information des LSTMs. Les unités de régulation des LSTMs pourraient faciliter le contournement des transitions complexes dans le résultat de la matrice des poids cachés. Les portes d'entrée et d'oubli sigmoïdes supplémentaires des unités LSTM permettent des fonctions de transition dépendantes de l'entrée encore plus flexibles que dans les mRNNs ordinaires.

- **Les modèles hybrides**

Le modèle hybride CNN-LSTM a été développé pour les problèmes de prédiction de séries temporelles visuelles et la génération de descriptions textuelles à partir de séquences d'images. L'architecture CNN-LSTM utilise des couches CNN pour l'extraction de caractéristiques sur les données d'entrée et se combine avec LSTM pour améliorer la prédiction de séquences. Plus précisément, le CNN extrait les caractéristiques des données d'entrée spatiales et les utilise dans l'architecture LSTM pour sortir la légende. Les applications de ce modèle hybride ont été utilisées pour résoudre de nombreux problèmes, tels que, la classification d'images (Gill & Khehra, 2021), les analyses de sentiments (Rehman et al., 2019) et les signaux de fréquence cardiaque . Des études ont montré des résultats prometteurs ; par exemple, (Qu et al., 2021) ont prédit la production PV à court terme grâce à la combinaison l'intensité future des précipitations dans une région locale sur une période relativement courte. Les expériences montrent que le réseau LSTM-CNN capture mieux les corrélations spatio-temporelles et surpasse systématiquement le réseau neuronal LSTM ainsi que d'autres méthodes statistiques de prévision telles que les modèles ARMA, ARIMA pour la prévision de la production PV.

Concernant le modèle LSTM-RNN, un avantage essentiel des réseaux neuronaux récurrents est leur capacité à utiliser des informations contextuelles lors du mappage entre les séquences d'entrée et de sortie. Malheureusement, pour les architectures RNN standard, la gamme de contextes auxquels il est possible d'accéder en pratique est assez limitée. Au cours du processus d'apprentissage, la sensibilité des RNN aux informations des étapes précédentes diminue. Le RNN peut même oublier les informations d'entrée les plus effacées. Le LSTM entre en jeu et permet d'améliorer les performances du modèle RNN traditionnel. Dans ce modèle LSTM-RNN, des blocs de mémoire sont appliqués et correspondent à un ensemble de sous-réseaux connectés de manière récurrente. Chaque bloc contient une ou plusieurs cellules de mémoire et trois unités multiplicatives appelées porte d'entrée, de sortie et d'oubli. Ces unités de porte peuvent fournir des analogues continus des opérations d'écriture, de lecture et de réinitialisation pour les cellules mémoire. Lorsque la porte d'entrée est activée, les données d'entrée sont enregistrées dans la cellule de mémoire, lorsque la porte de sortie est activée, les données sont transmises aux neurones suivants, lorsque la porte d'oubli est activée, l'information dans la cellule de mémoire est effacée. Par conséquent, la mémoire à long terme de la séquence de données d'entrée est réalisée en fonction de l'activation des différentes portes (Graves, 2012). (F. Wang et al., 2020b) a établi un modèle de prévision de prévision PV pour le jour suivant avec des performances excédents d'autres modèles tels que le modèle de

persistance, le modèle de réseau de neurones à propagation arrière (BPNN) et le modèle de machine à vecteur de support (SVM).

Le modèle LSTM basé sur le mécanisme d'attention (aLSTM) utilise l'attention qui dans le système de vision biologique permet à un animal de se concentrer sur des objets spécifiques pour les observer. Le mécanisme d'attention est un réseau de neurones qui simule les attentions du cerveau. Il a été tout d'abord été utilisé avec succès pour la traduction automatique (Choi et al., 2018) ou encore l'analyse vidéo.(W. Li et al., 2018). L'application du mécanisme d'attention au réseau neuronal profond permet au réseau neuronal de se concentrer de manière adaptative sur les caractéristiques d'entrée qui sont plus importantes pour la sortie actuelle, et d'atténuer l'interférence d'autres caractéristiques. (Zhou et al., 2019) ont proposé des structures d'un modèle basé sur l'attention en utilisant le vecteur de sortie de la couche cachée LSTM comme entrée du mécanisme d'attention qui recherche alors le poids d'attention. Ainsi, leurs études ont démontré une amélioration de la prévision de la production PV en utilisant ce modèle comparé au modèle de persistance, au modèle ARIMAX, au MLP ou encore au modèle simple LSTM. (T. Yang et al., 2019) utilisent, pour accroitre les performances de son modèle aLSTM, une optimisation bayésienne pour sélectionner les paramètres optimaux lors du processus de Machine Learning obtenant des résultats de prévision PV sur le lendemain supérieurs au modèle SVM, BPNN et LSTM simple.

3.2.5 Choix du modèle

Le principe de ce paragraphe s'appuie sur une série de critères permettant d'évaluer la structure neuronale la plus adaptée à notre problème.

- **Définition des critères de sélection**

Deux classes de critères sont mises en place dans cette étude. D'une part, des critères dits « objectifs » qui s'appuient sur la capacité du modèle à solutionner le problème le plus parfaitement possible et d'autre part, des critères dits « subjectifs » liés à l'utilisateur et qui dépend du niveau de connaissance de la solution.

- **Critères objectifs :**
  - L'originalité scientifique, le côté novateur de la recherche
  - L'absence de modèle physique dans la prévision
  - La place du réseau de neurones

- **Critères subjectifs**
  - Les performances de prévision
  - Les performances immédiates
  - Le minimum d'effort de calcul et de mise en place du modèle

- **Justification du choix**

Chaque critère évoqué revêt un aspect important quant au choix du modèle dans un ordre décroissant de priorité. Ainsi, ces travaux doivent, tout d'abord, s'inscrire dans une originalité scientifique. De plus, ce modèle de prévision doit utiliser une méthode statistique et non pas une méthode physique afin d'apporter une capacité de généricité essentielle à notre étude. La mise en place du RN est de ce fait important ici. Toutefois, le temps de calcul de la prévision est aussi de mon point de vue un critère clé. Un minimum d'effort de calcul doit être nécessaire à la prévision. De ce critère découle alors un autre : la facilité de mise en place de ce modèle. Le modèle doit alors être reproductible et facile à mettre en place d'où la volonté d'aller vers des modèles simples mais performants.

- **La structure du modèle choisi**

Dans le cadre de notre étude, l'approche retenue est celle du modèle LSTM ainsi que le Bi-LSTM (figure 3.6). En effet, la combinaison de ces modèles avec la méthode statistique de Johansen en amont de la prévision rend celle-ci novatrice et originale dans le cadre d'une prévision PV. De plus, la simplicité de mise en place des modèles LSTM et Bi-LSTM combinée à leur performance en matière de prévision est un critère primordial. La place du réseau de neurones était aussi importante et notamment des modèles LSTM, car ce type de RN permet l'analyse des larges jeux de données à notre disposition. Ces dernières sont essentiellement des données météorologiques et PV (discutées dans le chapitre précédent) sous forme de séries temporelles.

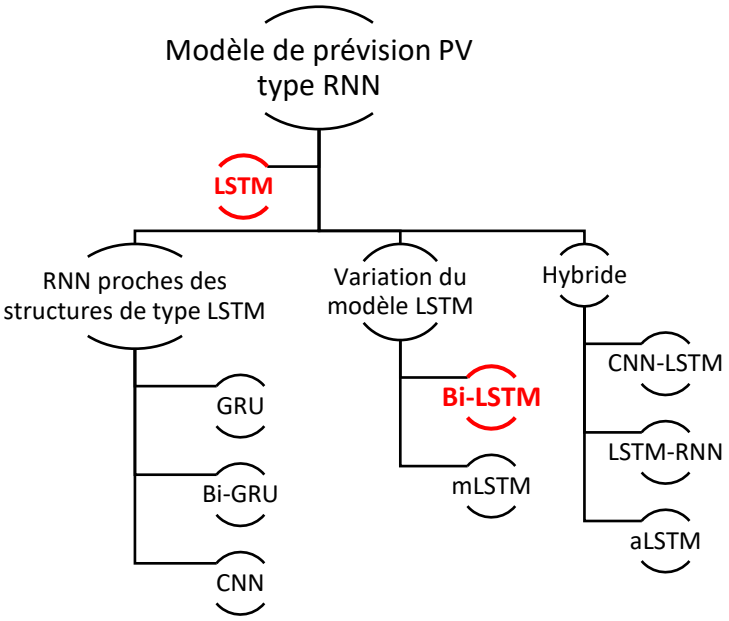


Figure 3.6 Choix de la structure du modèle de prévision adopté

Conclusion

Le modèle de prévision a été choisi selon plusieurs critères de sélection objectifs et subjectifs définis au fur et à mesure de la thèse. L'approche offre la possibilité d'obtenir un modèle de prévision PV optimal à notre objectif tout en octroyant une flexibilité nécessaire aux différents jeux de données. Ces derniers varient selon les diverses régions marquées par le relief de l'île.

Le chapitre suivant présente en détail l'approche de la cointégration par le modèle de correction d'erreurs vectoriel (MCEV) de Johansen pour proposer un modèle de conversion de la production d'énergie photovoltaïque à l'île de la Réunion. Pour cela, à partir d'un échantillon de données de moyennes journalières d'un an est extrait de sept années de mesures du bâtiment COREX situé à La Possession et à ces données sont appliquées la méthode MCEV de Johansen. Dans le chapitre est expliquée en premier lieu les séries temporelles et leurs propriétés, puis le test augmenté de Dickey Fuller (ADF test) utilisé sur les jeux de données permettant de mettre en avant la stationnarité des séries temporelles. Ensuite la cointégration de Johansen est mise en avant sur les différents paramètres météorologiques permettant de définir le modèle de prévision de la production PV du site et enfin les résultats expérimentaux sont explicités dans la dernière partie.

Chapitre 4

APPLICATION DE L'APPROCHE DE COINTEGRATION MCEV DE JOHANSEN POUR LA MODELISATION DE LA PREVISION DE LA PRODUCTION PHOTOVOLTAÏQUE A L'ILE DE LA REUNION

Avant-propos

Ce chapitre présente en détail l'approche de la cointégration par le modèle à correction d'erreurs vectoriel (MCEV) de Johansen pour proposer un modèle de prévision de la production d'énergie photovoltaïque à l'île de la Réunion. Pour cela, un échantillon de données de moyennes journalières d'un an est extrait de sept années de mesures du bâtiment COREX situé à La Possession sur la côte ouest de l'île et à ces données sont appliquées la méthode MCEV de Johansen qui auparavant été utilisée essentiellement en économétrie. Ce chapitre reprend très largement l'article nommé « Applying Johansen VECM cointegration approach to propose a forecast model of photovoltaic power output plant in Reunion Island » publié en mars 2020 dans la revue internationale à comité de lecture « AIMS Energy » (Fanchette et al., 2020). Les seules modifications apportées sont l'ajout de compléments permettant son intégration dans le manuscrit de thèse, ainsi que des modifications de certaines notations afin de les rendre cohérentes avec l'ensemble des travaux. Ainsi, dans ce chapitre est expliquée en premier lieu les séries temporelles et leurs propriétés, puis le test augmenté de Dickey Fuller (ADF test) utilisé sur les jeux de données permettant de mettre en avant la stationnarité des séries temporelles. Ensuite la cointégration de Johansen est mise en avant sur les différents paramètres météorologiques permettant de définir le modèle de prévision de la production PV du site et enfin les résultats expérimentaux sont explicités dans la dernière partie.

4.1 Introduction

To avoid an energy crisis created by the exhaustion of the fossil fuels, many countries have introduced renewable energy policies. For example, Reunion Island, a French overseas territory in the Indian Ocean, aims to achieve electrical autonomy by 2030 (Selosse et al., 2018). Among renewable energy sources (RES), solar energy is considered as a strong potential and availability on Reunion Island. That is why, the French government has made this insular zone an experimental territory for RES by implementing great powers photovoltaic (PV) plants more than 194 MW in 2019. However, due to the high variability of environmental factors (Laronde et al., 2010; Omubo-Pepple et al., 2009) on PV cell efficiency, the high penetration of PV in electric systems may threaten the stability and reliability of the electrical power grid for smart buildings or smart city applications. Indeed, one of Reunion island project is to build up active micro-grids or virtual PV power plants to feed in power to all devices and appliances of a smart building. Therefore, PV systems operating under real field conditions are of great importance for obtaining accurate prediction of their efficiency and power output. In this context, an accurate forecasting (Antonanzas et al., 2016; Sobri et al., 2018; Wan et al., 2015) of the PV power generation can improve system reliability and power quality, and reduce the impact of uncertainties on the grid. However, the accuracy of PV power output forecasts from inclined building mounted modules for optimal energy extraction (Al- Sabounchi, 1998) are not just based on climatic conditions (Amajama et al., 2020; Chandra et al., 2018; Kaldellis et al., 2014), because its fluctuation is due to several factors such as solar irradiance, temperature, wind speed, humidity and dust (Ketjoy & Konyu, 2014). Over the last decade, a large number of solar PV power generation forecasting techniques (Barbieri et al., 2017; Y. li et al., 2014; Raza et al., 2016) have been modeled. The state-of-the-art techniques to produce power forecasts for PV has been described and classified (Antonanzas et al., 2016) in three main approaches : physical, statistical and hybrid methods. (Sobri et al., 2018) also reviewed PV power output forecasting techniques but classified them into three different major methods: statistical-time series methods, physical methods and ensemble methods. Among the statistical-time series approach, classical regression methods, which have been studied, take advantage of the correlation nature of meteorological parameters using prediction models as input (AlSkaif et al., 2020; Zamo et al., 2014a, 2014b). However, regression between non-stationary series (Enders, 1995; Gujarati, 2004) may conclude on the existence of two variables even though there is no real relationship between them. The goal of this study is to parameterize, and to our knowledge for the first time in an insular zone, the more realistic relationship between PV plant power output and relevant meteorological factors such as incoming irradiation, cell temperature, wind speed and humidity using a powerful statistical technique.

The proposed method falls into the category of multiple linear regression methods. There have been some recent studies concerning PV power generation estimate thanks to a relationship between a dependent variable (PV power) and independent variables, called predictors.(Antonanzas et al., 2016) made a classification:

- Linear stationary models. Auto-Regressive methods (AR) (Bacher et al., 2011) models the PV power output as a linear combination of the lagged values of its predictors, simple Moving Average (MA) (Y. li et al., 2014), the Auto-Regressive Moving Average (ARMA) (Chu et al., 2015) models combining the two last methods, AR eXogenous (ARX) (Bacher et al., 2011) methods adds exogenous data to an AR model, ARMAX (Y. li et al., 2014) an

ARMA model with exogenous data and also the Vector AR (VAR) and Vector ARX (VARX) (Bessa et al., 2015).

- Linear non-stationary models. Auto-Regressive Integrated Moving Average (ARIMA) (Pedro & Coimbra, 2012) techniques model a stochastic process combining AR component to a MA component, the Seasonal ARIMA (SARIMA) (Bouzerdoum et al., 2013) which introduce a seasonal component and the coupled auto-regressive and dynamical system (CARDS) model (J. Huang et al., 2013).

Researches classified the forecasting of PV power production in different categories based on the needs of the PV production and transport actors. In general, PV power forecasting depends on the meteorological and solar irradiance data, the type of method used to forecast and the forecasting horizon. (Das et al., 2018; Kostylev & Pavlovski, n.d.; Zamo et al., 2014a) proposed a classification for these forecasting time horizon:

- Very short-term forecast horizon: a few second to one hour, used for electricity dispatch in real time and energy smoothing.
- Short-term forecast: for one hour, several hours up to a day ahead, to guarantee system commitment and scheduling
- Medium-term forecast: multiple days to months ahead, to ensure power system planning
- Long-term forecast: months to one to several years, to find and assess potentially resourceful sites.

This statistical method aims at creating a medium-term forecast model of the hourly production of PV electricity for the next days, in order to answer needs of electricity grid managers, energy traders and producers. Parametrized and regressive model such as the one proposed is best built for short and medium-term forecast horizon (M. Diagne et al., 2013).

In this paper, a linear relation analysis of time series data collected over a year is performed, and the dependent variable of PV power output P is investigated on explanatory variables such as solar irradiation, cell temperature, wind speed and humidity. These four variables are denoted respectively as G, T, swind and humi. For that, the stationarity of each previous cited time series is tested. Then, the Augmented Dickey Fuller (ADF) test is used to determine the method of regression estimation between the PV and all influencing variables. A former study (Ramenah et al., 2017) using a robust statistical technique has shown a relationship between P and T in a non-tropical zone. The statistical method used was the Engle & Granger (EG) cointegration technique. The disadvantage of the EG method is that it does not distinguish several cointegration relationships. For instance, the study of N variables simultaneously, with $N > 2$, may lead to up to N-1 cointegration relations, and the method of EG allows to obtain only one cointegration equation. To overcome this difficulty, an original statistical study of the Johansen technique (Andrei & Andrei, 2015; Marcinkiewicz, 2014) is proposed. The Johansen cointegration approach is suggested to determine the most appropriate PV power output-forecasting model. Even though the Johansen approach for cointegration has been a popular tool in applied economic (Katircioglu, 2009), it has never been applied to renewable energy technologies and even less to PV systems for forecasting.

For optimal energy extraction, the PV design for this study is a building mounted on the grid connected system where the modules that make up the PV plant are at a tilted angle of 21°, same as Reunion Island latitude. The polycrystalline PV cells of 180W each are equipped with solar irradiance, cell temperature and wind sensors. For this study, the sample of one year daily means data is retrieved among 7 years of measurements from the COREX building located at La Possession in the west coast of the island and all tests are performed under 64-bit Eviews

software 9 environment of HIS Global Incorporation. This paper is part of a European project[1], one of the main focuses of which is PV power output-forecasting in tropical island environments.

The rest of the paper is organized as follows. Section 4.2 explains the effect of environmental factors on PV systems. Section 4.3 describes time series and their properties, and statistical techniques used in this study to link PV power output and environmental parameters (explanatory variables). Section 4.4 deals with the principle of cointegration and vector error correcting model for the determination of the short and long run relationship between the explanatory variables. In section 4.5, the Johansen cointegration approach is detailed whereas Section 4.6 shows the application of this approach to experimental data. Section 4.7 presents obtained experimental results. Finally, a discussion is proposed in Section 4.8 and appropriate conclusions and future works are given in Section 4.9.

4.2 Effect of environmental factors on PV systems

4.2.1 Solar radiation effect

Light can be considered to consist of a stream of tiny particles of energy called photons which when fall on a PV cell convert photonic energy into electrical energy. The PV characteristic current—voltage (I—V) curve and power-voltage (P-V) is illustrated in Figure 4.1. The most relevant parameters used to evaluate the performance of solar cells are the short-circuit current (I_{SC}) and the open-circuit voltage (V_{OC}). The conversion efficiency (η) is determined from these parameters and is calculated as the ratio between the generated power at the maximum power point (P_{MPP}) and the incident solar irradiance (W/m²). As indicated in Figure 4.1, the greater is the power of the solar radiation, the greater is P_{MPP}. Therefore, solar irradiance is an environmental factor that is considered in this study.

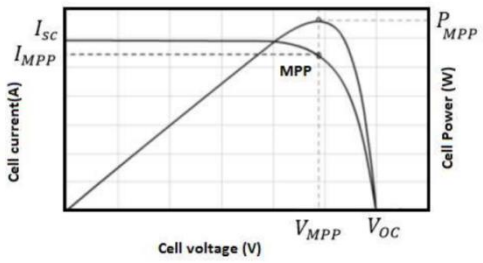


Figure 4.1 Power-Voltage and Current-Voltage curve of a solar cell.

4.2.2 Temperature effect

Solar irradiance is the biggest environmental factor for solar cells that convert light into electricity. PV modules generate electrical power proportionally of the solar radiation while considering the PV module performance is sharply sensitive to cell temperature. Solar irradiance and cell temperature are two factors, which affect the performance of a PV cell. The PV cell temperature affects negatively its voltage and positively its current.

[1] French acronym for Supervision, Dynamic Management and Optimization of Urban Micro grids for Island Electricity Self-sufficiency.

Whereas manufacturers only provide PV characteristics under laboratory Standard Testing Conditions (STC), in real conditions, the PV cell temperature T also has great influence (Dubey et al., 2013; Skoplaki & Palyvos, 2009) on the power output P. In tropical zone, cell temperatures can reach or even exceed 70 ℃ compared to 25 ℃ STC conditions. T is required to calculate the power loss, usually between −0.35 and −0.5%/℃. This means that every 10 ℃ in excess results in a decrease in P between 3.5 and 5%. As P changes with temperature fluctuations, this parameter must be taken into account to optimize the annual yield and to analyze electrical grid temporal stability of their supply.

Although T is one of the most important factors that affect the performance of the PV modules, additional physical conditions such as humidity, wind and dust have drastic impact on P.

4.2.3 Humidity effect

Humidity is the amount of steam present in air. Figure 4.2(a) shows the low percentage of reflected light due to the glazing cover when most of the incident photonic energy is converted into electrical energy.

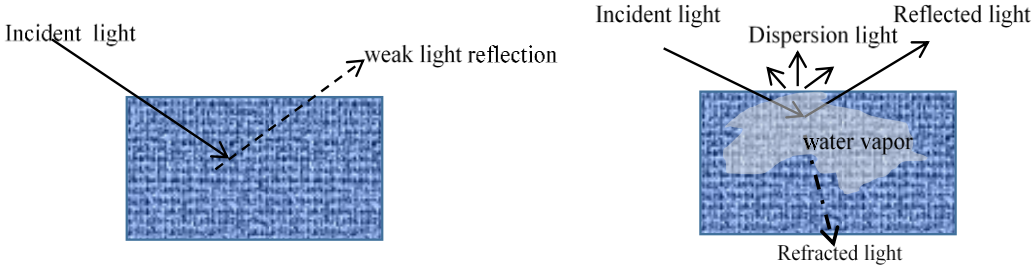


Figure 4.2 (a) Strong PV power output; (b) Weak PV power output due to humidity.

Small steam particles in the atmosphere cause light diffraction or scattering, and thus altering direct PV irradiation. Figure 4.2(b) shows light refraction due to steam that greatly reduces light intensity to the solar cells leading to reducing power output. Additionally, reflection and dispersion of the striking light to these water molecules that act as a prism is subjected to more losses of the total energy which is not subjected to conversion by the PV module.

Tropical regions such as Reunion Island in the Indian Ocean are frequently humid areas. This environmental factor creates obstacles and much sharper drop in irradiance levels resulting in a PV efficiency decrease mainly on open circuit voltage and short circuit current.

4.2.4 Wind effect

As the wind cools well by ventilation the solar modules, this factor reduces the temperature impact. Higher is the wind speed, better is the conversion efficiency. Consequently, the impact of the wind effect is opposite to solar irradiation and temperature effects. As indicated by manufacturers, PV modules that are cooled by 1 ℃ should increase the efficiency up to 0.05% with increasing percentage over time. It has also been shown that adaptive cooling mechanism (Luo et al., 2017) for PV modules can reduce thermal losses below 5% compared to uncool PV systems. Wind speed also has other effects on PV modules such as an increase in wind speed improves short circuit current and open circuit voltage. Such

experiment has been conducted to show the effect of the wind speed on the output of modules. The wind effect on photonic particles is similar to that on propagating electromagnetic radio waves (Joseph & Oku, 2016). Collisions between particles of the air and photonic particles result in a change of direction of the latter in an opposite direction. Thereby, air temperature, humidity and wind are three environmental factors considered in this study.

4.2.5 Dust effect

The effect of dust particles deposits on PV modules has the effect of decreasing the electrical energy output by reducing the amount of absorbed solar radiation. The quantity of dust on PV modules has been studied (Qasem et al., 2014) where a decrease in electrical energy output has been observed varying between 3% and 11% depending on this quantity. Although Piton de la Fournaise Volcano in Reunion Island is still under activity generating the dust, this factor is not considered in this study yet.

4.3 Time series & properties

4.3.1 Time series

A time series is a set of observations on the values that a variable takes at different times. Time series data are collected at regular time intervals such as for instance daily, weekly or annually (Jalil & Rao, 2019). A time series is stationary if its mean and variance are constant over time, and the value of the covariance between two time periods depends only on the distance or gap or lag between the two time periods and independently of the actual time (Granger & Weiss, 1983). If a time series is stationary, that is it does not require any differencing. In this case, it is said to be integrated of order zero and denoted I(0). If a time series is not stationary in the sense just defined, it is called a non-stationary time series. Usually, if a non-stationary time series has to be differenced d times to make it stationary, that time series is integrated of order d noted as I(d). For example, if a time series has to be differenced twice, that is taking the first difference of the first derivatives to make it stationary, it is called second order integrated time series denoted as I(2). Although the interest is in stationary time series, non-stationary time series are often encountered. Let's explain this process through a random walk model and finally define conditions for stationarity. Considering y_t as a variable following a random walk where y_t is regressed at time t on its value lagged one period, as given in Eq 52:

$$ry_t = y_{t-1} + \varepsilon_t \tag{52}$$

where ε_t is a white noise error term with the mean of zero and the variance $\sigma 2$. From Eq 52, by proceeding by recurrence Eq 53 can be obtained as follows:

$$y_1 = y_0 + \varepsilon_1$$

$$y_2 = y_1 + \varepsilon_2 = y_1 + \varepsilon_1 + \varepsilon_2 \tag{53}$$

$$y_t = y_0 + \sum_{i=1}^{t} \varepsilon_i$$

where ε_i is given as NID (0, σ2) and NID represents normally and independently distributed with a mean value of zero and a constant variance. This process is non-stationary as shown in Eq 54 where Var stands for variance.

$$Var\,(y_t) = Var\left(\sum_{i=1}^{t} \varepsilon_i\right) = \sum_{i=1}^{t} Var(\varepsilon_i) = \sum_{i=1}^{t} \sigma_\varepsilon^2 = t\,\sigma_\varepsilon^2 \quad (54)$$

From Eq 54 is deduced that the variance of y_t process is a time function. Considering that t increases, its variance increases indefinitely, and thus violating a condition of stationarity.

A time Series is stationary if it has the following conditions:

- constant mean for all time t,
- Var (y_t) is a finite constant independent of t,
- Cov (y_t, y_{t-1}) is a finite function which is independent of t.

It is also interested to express Eq 51 in a differential form as given in Eq 55:

$$y_t - y_{t-1} = \Delta y_t = \varepsilon_t \quad (55)$$

If y_t is non-stationary, its first derivatives is stationary and correspond to the first derivatives of a random walk-time series which are stationary.

4.3.2 Properties of time series

Regression analysis of time series is used to discover or to verify the predicted relationships and properties of integrated series have to be verified. Regression of a non-stationary time series on another non-stationary time series can produce a spurious regression (Mills, 2019). To avoid the spurious regression problem from such regression, we must transform non-stationary time series to make them stationary. Several statistical tests have to be executed to determine if a time series is stationary. Unit root test has become one of the most widely used methods for testing the stationarity of a time series. To explain the idea behind the unit root test, a general form of Eq 56 is used as follows:

$$y_t = \beta\, y_{t-1} + \varepsilon_t \quad (56)$$

which can be transformed as Eq 57

$$y_t - y_{t-1} = \beta\, y_{t-1} - y_{t-1} + \varepsilon_t \Rightarrow \Delta y_t = (\beta - 1) y_{t-1} + \varepsilon_t \quad (57)$$

where Δ is the first difference operator. The unit root test is a hypothesis test with the following hypothesis: if β = 1, there is a unit root and the time series is non-stationary which refers to the null hypothesis H_0. The alternative hypothesis H_1 is that β < 1 and the time series is stationary.

4.3.2.1 Residual diagnostics

The serial correlation in residual from estimated equation test is based on the hypothesis testing. There many residual tests, first order, second order or squared residuals. The following sections describe only some tests and their outcome interpretations that are used in this study to determine the most perfect model between PV output and climate parameters in a tropical zone. The goal of this section is to give the guideline about serial correlation. Only if a model is free from serial correlation or heteroscedasticity, then it can be used for forecasting.

4.3.2.2 Augmented dicker fuller

Dickey Fuller (DF) test is the simplest approach to test for a unit root. In case of autocorrelation problem of εt, DF developed a test called Augmented Dickey Fuller (ADF) test (Davidson & MacKinnon, 2009). As an example, the outcome of the ADF test applied to the variable P, using the Eviews software, is represented in Table 4.1. If the null hypothesis is true, which expresses itself P has a unit root, such series is a non-stationary one. The ADF test is based on t-statistic approach and probability approach as represented in Table 4.1, respectively. DF tabulated critical values are chosen significance level at 1%, 5% and 10%, respectively. To check the unit root test, the calculated t-statistic with its corresponding probability, which is indicated as statistic ADF test in Table 4.1, has to be compared to the critical values, and mainly at 5% level. According to the guideline, if the test statistic value is greater than the 5% level critical value as well as for probability value, the null hypothesis cannot be rejected. Therefore, power series are non-stationary series. In Table 4.1, the ADF t-statistic value which is −1.000357 greater than −1.941740 value at 5% level as well as for the probability value at 28.44%.

Table 4.1 Example of the outcome of an ADF test

Null Hypothesis: POWER has a unit root			
Exogenous: None			
Lag Length: 8 (Automatic—based on SIC, maxlag = 16)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		−1.000357	0.2844
Test critical values	1% level	−2.571643	
	5% level	−1.941740	
	10% level	−1.616087	
* MacKinnon (1996) one-sided p-values			

The same test was repeated for the first difference of P. It can be deduced from the corresponding outcome given in Table 4.2 where D(P) or I(1) series is stationary at first difference.

Table 4.2 Example of the outcome of an ADF test for stationary series.

Null Hypothesis: D(POWER) has a unit root

Exogenous: None

Lag Length: 8 (Automatic - based on SIC, maxlag = 16)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		−12.51960	0.2844
Test critical values	1% level	−2.571643	
	5% level	−1.941740	
	10% level	−1.616087	

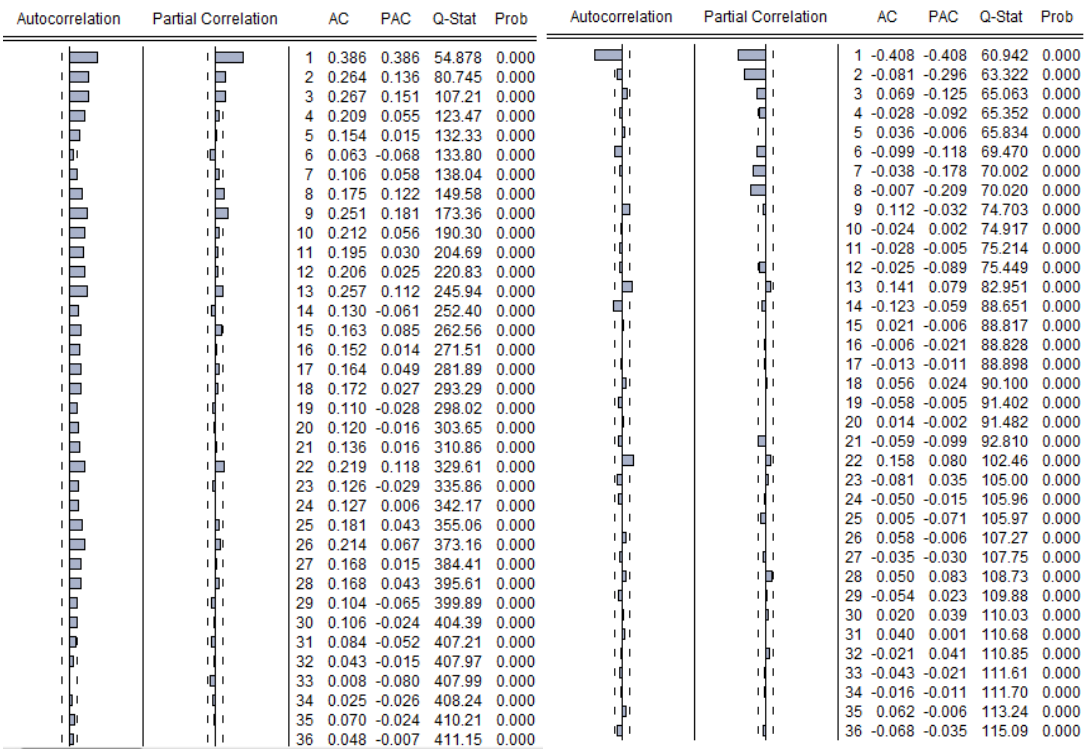
* MacKinnon (1996) one-sided p-values

4.3.2.3 Correlogram

Another visual diagnosis tool identified as the first step for the stationary test can be done by computing the autocorrelation function (ACF), and the partial autocorrelation function (PACF). This graphical tool is known as correlogram or Ljung Box (LB) statistics(Hassani & Yeganegi, 2019). The following figures show correlograms of a time series. Autocorrelation and partial autocorrelation functions characterize the pattern of temporal dependence in the series. Autocorrelation and partial autocorrelation give the impression that the residuals are purely random. Correlogram is simply plots of ACFs and PACFs against the lag length given as the arithmetical progression 1 to 36 in each above table. The solid vertical line in this diagram represents the zero axis and observations between spikes above or below the line are positive and negative values, respectively. For stationary time series, the correlogram tapers off quickly, whereas it dies off gradually for non-stationary time series. For example, Figure 4.3(a) represents a stationary time series correlogram, similar to a purely white noise process and the autocorrelations at various lags hover around zero. Figure 4.3(b) represents a typical non-stationary series, as the autocorrelation coefficient starts at quite a high value and declines very slowly toward zero as the lag lengthens. Finally, we can simply point out the statistical significance of the autocorrelation coefficients given by the Eq 58, where k is the lag number:

$$\rho k \cong \text{NID}\ (0,\ 1/N) \tag{58}$$

The sample autocorrelation coefficients are normally distributed with zero mean and a variance equal to one over the sample size N. We conclude this visual diagnostic from Figure 4.3 by specifying that, if the time series is not stationary at level, it has to be differenced once or more times to achieve stationarity. Furthermore, correlograms of both autocorrelation and partial autocorrelation must indicate that residuals are purely random.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.386	0.386	54.878	0.000
		2	0.264	0.136	80.745	0.000
		3	0.267	0.151	107.21	0.000
		4	0.209	0.055	123.47	0.000
		5	0.154	0.015	132.33	0.000
		6	0.063	-0.068	133.80	0.000
		7	0.106	0.058	138.04	0.000
		8	0.175	0.122	149.58	0.000
		9	0.251	0.181	173.36	0.000
		10	0.212	0.056	190.30	0.000
		11	0.195	0.030	204.69	0.000
		12	0.206	0.025	220.83	0.000
		13	0.257	0.112	245.94	0.000
		14	0.130	-0.061	252.40	0.000
		15	0.163	0.085	262.56	0.000
		16	0.152	0.014	271.51	0.000
		17	0.164	0.049	281.89	0.000
		18	0.172	0.027	293.29	0.000
		19	0.110	-0.028	298.02	0.000
		20	0.120	-0.016	303.65	0.000
		21	0.136	0.016	310.86	0.000
		22	0.219	0.118	329.61	0.000
		23	0.126	-0.029	335.86	0.000
		24	0.127	0.006	342.17	0.000
		25	0.181	0.043	355.06	0.000
		26	0.214	0.067	373.16	0.000
		27	0.168	0.015	384.41	0.000
		28	0.168	0.043	395.61	0.000
		29	0.104	-0.065	399.89	0.000
		30	0.106	-0.024	404.39	0.000
		31	0.084	-0.052	407.21	0.000
		32	0.043	-0.015	407.97	0.000
		33	0.008	-0.080	407.99	0.000
		34	0.025	-0.026	408.24	0.000
		35	0.070	-0.024	410.21	0.000
		36	0.048	-0.007	411.15	0.000

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.408	-0.408	60.942	0.000
		2	-0.081	-0.296	63.322	0.000
		3	0.069	-0.125	65.063	0.000
		4	-0.028	-0.092	65.352	0.000
		5	0.036	-0.006	65.834	0.000
		6	-0.099	-0.118	69.470	0.000
		7	-0.038	-0.178	70.002	0.000
		8	-0.007	-0.209	70.020	0.000
		9	0.112	-0.032	74.703	0.000
		10	-0.024	0.002	74.917	0.000
		11	-0.028	-0.005	75.214	0.000
		12	-0.025	-0.089	75.449	0.000
		13	0.141	0.079	82.951	0.000
		14	-0.123	-0.059	88.651	0.000
		15	0.021	-0.006	88.817	0.000
		16	-0.006	-0.021	88.828	0.000
		17	-0.013	-0.011	88.898	0.000
		18	0.056	0.024	90.100	0.000
		19	-0.058	-0.005	91.402	0.000
		20	0.014	-0.002	91.482	0.000
		21	-0.059	-0.099	92.810	0.000
		22	0.158	0.080	102.46	0.000
		23	-0.081	0.035	105.00	0.000
		24	-0.050	-0.015	105.96	0.000
		25	0.005	-0.071	105.97	0.000
		26	0.058	-0.006	107.27	0.000
		27	-0.035	-0.030	107.75	0.000
		28	0.050	0.083	108.73	0.000
		29	-0.054	0.023	109.88	0.000
		30	0.020	0.039	110.03	0.000
		31	0.040	0.001	110.68	0.000
		32	-0.021	0.041	110.85	0.000
		33	-0.043	-0.021	111.61	0.000
		34	-0.016	-0.011	111.70	0.000
		35	0.062	-0.006	113.24	0.000
		36	-0.068	-0.035	115.09	0.000

Figure 4.3 Example of (a) non-stationary correlogram (b) stationary correlogram.

4.3.2.4 Q-Statistic

An alternative to LB statistics is the Q statistics developed by Box and Pierce is a joint hypothesis test of all the correlation coefficients instead of individual tests (Ljung & Box, 1978).

As seen in Figure 4.3, due to the large number of samples in this study, the Q-stat values differ consistently between the two tables at lag order 36. Although each corresponding probability value is significant, only Figure 4.3(b) is acceptable due to stationary criteria.

4.3.2.5 Durbin—Watson & LM tests

Durbin Watson (DW) statistics is a way for detecting serial correlations in a regression model of for example three variables (POWER, IRRA and TEMP) as given in Eq 59.

$$\text{Power} = \alpha\ \text{IRRA} + \beta\ \text{TEMP} + C + \text{Power}\ (-1) \tag{59}$$

where Power is the dependent variable, and Power (-1) is the one-period lag dependent variable.

Eq 58 is also known as an autoregressive (AR) model. DW can be used only if there is only one lag in the AR model. For several-periods lag, Q-statistics and Lagrange Multiplier (LM) tests have to be applied to the AR and to the outcome. The corresponding probabilities obtained using Eviews software are represented in Table 4.3 for the LM tests, and Table 4.4 for the Q-statistics tests. If the null hypothesis H_0 is true, there no serial correlation. If the alternative hypothesis H_1 is true, there is a serial correlation. However, all probability values being less than 5%, H_0 can be rejected, or rather H_1 is accepted. Indeed, there is serial correlation in the AR model.

Table 4.3 Outcome of serial correlation test.

Dependent Variable: POWER				
Method Least Squares				
Date 07/11/19 Time: 09:41				
Sample (adjusted): 1/02/0365 1/01/0366				
Included observations: 365 after adjustments				
Variable	Coefficient	Std.Error	t-Statistic	Prob.
IRRA	17.98115	0.320048	56.18262	0.0000
TEMP	12.70429	10.80203	1.176102	0.2403
C	−197.9435	307.4385	−0.643847	0.5201
POWER(-1)	−0.002498	0.008110	−0.308041	0.7582
R-squared	0.979952	Mean dependent var	8588.047	
Adjusted R-Squared	0.979786	S.D dependent var	2998.959	
S.E.of regression	426.3841	Akaike info criterion	14.95946	
Sum squared resid	65631037	Schwarz criterion	15.00220	
Log likelihood	−2726.101	Hannan-Quinn criter.	14.97644	
F-statistic	5881.985	Durbin-Watson stat	1.680805	
Prob(F-statistic)	0.000000			

Table 4.4 Correlogram of serial correlation.

Q-statistic probabilities adjusted for 1 dynamic regressor						
Autocorrelation	Partial Correlation	LAGS	AC	PAC	Q-Stat	Prob*
		1	0.160	0.160	9.3646	0.002
		2	0.094	0.070	12.621	0.002
		3	0.405	0.391	73.213	0.000
		4	0.072	−0.049	75.162	0.000
		5	0.074	0.036	77.226	0.000
		6	0.093	−0.097	80.424	0.000
		7	0.104	0.114	84.494	0.000
		8	0.096	0.035	87.967	0.000
		9	0.116	0.120	93.009	0.000
		10	0.073	−0.045	95.046	0.000

Usually, the LM test is used for higher order errors with variables that are dependent on longer periods of time. This is the purpose of the Breush-Godfrey test for the estimation of least squares or second-order least squares. The result is given in Table 4.5.

Table 4.5 LM Test of serial correlation.

Breush-Godfrey Serial Correlation LM Test			
F-Statistic	5.885122	Prob.F(2,359)	0.0031
Obs*R-squared	11.58707	Prob.Chi-Square(2)	0.0030

Analyzing the observed R squared and the corresponding probability, less than 5% significant level, the null hypothesis can be rejected. Therefore, there is a serial correlation in the AR model. According to both tests, there is a serial correlation in the AR model, which cannot be used for forecasting.

4.3.2.6 Histogram—Normality test & jarque bera statistic

If residuals are normally distributed, a bell form shaped curve can be superimposed on the histogram (Hoffman, 2015). Plotting residuals of the latter is a rough method to test the normality hypothesis. The histogram is usually given with significant value of the Jarque Bera (JB) statistics which has two degrees of freedom for the null hypothesis of normality, i.e., the residuals are normally distributed. JB must be used for very large samples. Considering the great number of observations, this is very suitable for this study. The JB formula for the null hypothesis of normality is given in Eq 60:

$$JB = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right] \tag{60}$$

where S is the skewness coefficient (symmetrical form), K is the kurtosis coefficient (flattening form) and n is the sample size. For example, the JB is expected to be null if S = 0 and K = 3 for a normality test.

The horizontal axis of the histogram represents variables of interest such as an ordinary least squares residuals values. The vertical axis is the expected value of this variable if it were normally distributed. Figure 4.4 represents the histogram with a normal distribution considering the JB value and the corresponding probability value, indicating the null hypothesis of normal distribution for this large number of observations, cannot be rejected.

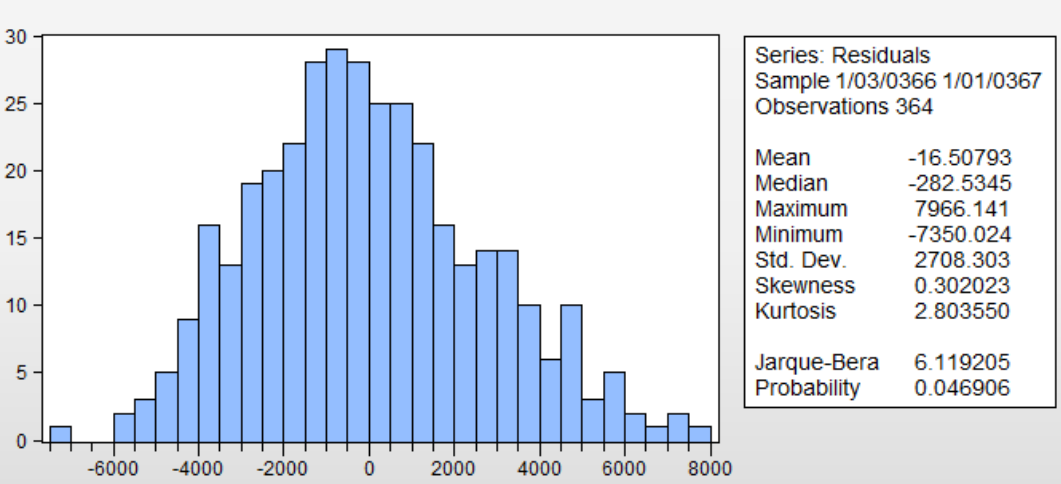


Figure 4.4 Example of a normal residual distribution.

More probability distributions shall be used in this study. Some definitions and technical terms are explained in the following section for a better understanding, and results outcomes through different Eviews tables.

4.3.2.7 Further definitions of statistical distributions

(a) Chi-Square

Considering random variables such as x_1, x_2, x_3… x_n that are normally and independently distributed. xi follows the normal distribution given as in Eq 61:

$$x_i = NID\ (\mu_i, \sigma_i^2) \tag{61}$$

where μ and σ^2 are the mean value and variance of x, respectively. Variances measure the dispersal of the data points around the mean. If data fall far from the mean value, variance shall increase.

The chi-square denoted as χ^2 is given by the Eq 62:

$$x_1^2 + x_2^2 + x_3^2 \ldots\ldots x_n^2 = \sum x_i^2 \cong \chi_n^2 \tag{62}$$

where n denotes the degree of freedom. This test is sometimes referred to as the median test. Chi-squares are specifically tabulated for different uses of the null hypothesis.

(b) Coefficient of Determination

A linear regression model is given as in Eq 63:

$$y_t = \beta_1 x_t + \beta_2 + e_t \tag{63}$$

where e_t is the residual term. The sum squared residual SSR is given in Eq 64:

$$SSR = \sum_t e_t^2 \tag{64}$$

The coefficient of determination R^2 is used to evaluate the goodness-of-fit of a regression model. Its value lies between 0 and 1. If this value is closer to 1, better is the fit. However, R^2 is a measure of the accuracy of the relationship between the model and the dependent variable. However, it is not a formal test for relationship determination as expressed by the Eq 65:

$$R^2 = 1 - \frac{SSR}{TSS} \tag{65}$$

where TSS is the total sum of squares given as in Eq 66:

$$TSS = \sum_t (y_t - \bar{y_t}) \tag{66}$$

where $\bar{y_t}$ is the estimated value, and $e_t = y_t - \bar{y_t}$ is the difference between the observed value of the variable and the adjusted value using the estimated coefficient of the model.

(c) F-Statistic

The F-statistic denoted as F test is a joint test that indicates whether a linear regression model provides a better fit to the data than a model that contains no independent variables. The F-test is related to the R-squared as given in Eq 67:

$$F = \frac{R^2/k}{(1-R^2)/(N-k-1)} \tag{67}$$

where k is the degree of freedom and N the number of observations. The F-test as already mentioned is based on the hypothesis test. The null hypothesis H_0 of F-test states that the model with no independent variables fits the data as well as the model under test. If the F-test is significant, then it can be concluded that the correlation between the model under test and the dependent variable is statistically significant. Consequently, R^2 value from Eq 67 is not equaled to zero. The F-test is given with its corresponding probability value at a significant level. If the probability value is less than the significance level, then data provide sufficient evidence to conclude that the regression model fits the data better than the model with no independent variables.

(d) Lag length criteria

To determine the best relationship between the model and the dependent variable is to choose the optimal lag length as an essential element for relationship stability. Including too many lagged terms will consume degrees of freedom and possibility of multicollinearity whereas too few lags will lead to specification errors. Several criteria have been defined such as Akaike, Schwartz, Hannan and probably at a lesser extent the Durbin Watson. Akaike and Schwartz make it possible to intercede when introducing one or more explanatory variables, between the loss of degrees of freedom and the information endowment. The lower the lag length is, the better the model is. In this study, the Akaike information criteria (AIC) and Schwartz information criteria (SIC) are considered and will be indicated in the outcome table form of Eviews software for all regression determination. The AIC is defined as in Eq 68:

$$AIC = e^{2k/N} \frac{SSR}{N} = e^{2k/N} \frac{\sum e_t^2}{N} \tag{68}$$

The SIC relationship is similar to AIC with a half value of exponential term. However, it should be noted that for both AIC and SIC the lowest lag length should give a better model. Table 4.6 is an example of the results that come up for a regression test under Eviews. All the specified parameters are indicated in this table with the corresponding values. Similar tables will be seen in this study and must be analyzed if the values are statistically significant.

Table 4.6 Statistical parameters

R-Squared	0.031745	Mean dependent var	8.84E-13
Adjusted R-Squared	0.018260	S.D dependant var	424.6234
S.E of Rgression	420.7288	Akaike info criterion	14.93816
Sum squared resid	63547555	Schwarz criterion	15.00226
Log likelihood	−2720,213	Hannan-Quinn criter.	14.96363
F-statistic	2.354049	Durbin-Watson stat	2.053022
Prob(F-statistic)	0.040231		

4.4 Properties of cointegration and error correction mechanism

4.4.1 Properties of cointegration

The method of cointegration in regression analysis is based on an assumption of stationary increments with fixed time lag called I(d). These terminology and notation have been established in upper sections. The development of the cointegration technique is based on I(d) integration to infer a short time as well as long-time equilibrium relations between non-stationary variables via regression analysis. Here, it should be pointed out that the regression of a non-stationary time series (on another non-stationary time series) may produce a spurious regression. One way to lookout against it is to find out if the time series are cointegrated. A combination of two or more individual non-stationary series may result in a stationary series. The properties of cointegration are explained as follows.

When regressing using the least squares regression including two non-stationary variables as given in Eq 63, and rewritten as integrated order I(1) in Eq 69:

$$e_t \quad = \quad P_t \; - \; \beta_1 G_t \; + \; \beta_2 \quad = \quad I(1) \tag{69}$$

e_t is non-stationary and auto correlated as the DW is very small. Basically, Granger demonstrated that if Pt and Gt are I(d) series, a linear combination of et is also I(d) that may result in a spurious regression. The last one is characterized by a high R^2 and t Student value even though there is no meaningful relationship between the two variables. To avoid such situation, regression is performed on variables in first difference which are stationary (ΔPt and ΔGt are I(0)) as represented in Eq 70:

$$\Delta P_t = \alpha \, \Delta \, x G_t + \beta + u_t \quad \Rightarrow \; u_t = \Delta P_t - \alpha \, \Delta \, G_t - \beta = I(0) \tag{70}$$

However, sometimes a regression with variables at level is preferred rather than at first difference. In that case, it is important to know how to regress non-stationary variables, and if the regression is not a spurious regression. Then, the concept of cointegration is applied.

The idea behind cointegration is as follows: in a short term G_t and P_t may have a divergent evolution but they will evolve together in the long term. There exist then a long-term relationship between P_t and G_t that is stable denoted as the cointegration relationship given as in Eq 71:

$$P_t = a\,G_t + b \tag{71}$$

A summary of cointegration concepts and conditions is given below:

Cointegration of two or more-time series suggests that there is a long-run, or equilibrium, relationship between them. The two cointegration conditions are, firstly, these series have to be of the same integrated order I(d). Secondly, a linear combination of these series allows to reduce the integrated series to a lower order.

To reconcile the short-run behavior with its long-run behavior, an error correction mechanism (ECM) has been developed by Engle & Granger (EG).

4.4.2 Error correction mechanism

In this section, to help understanding, only two variables P and G are considered to have only one cointegration relationship between them. The same principle is extended to five variables in the final study. If two variables P and G are cointegrated (P_t – â G_t –b is I(0)), then the relationship between them can be expressed as an ECM, such as Eq 72:

$$\Delta P_t = \gamma\,\Delta G_t + \delta\left(P_{(t-1)} - a\,G_{(t-1)} - b\right) + \upsilon_t \tag{72}$$

where must have a negative sign, Pt behaves as spring recall force and can go back to its long-term equilibrium value given as (aG(t-1) + b). Otherwise the specification of ECM type is not valid. The ECM allows to model jointly short-term dynamics (variables in first difference) and long-term dynamics (variables at level).

The short-term dynamic is given as in Eq 73(a):

$$P_t = \alpha_1 P_{t-1} + \alpha_2 G_t + \alpha_3 G_{t-1} + \alpha_0 + \upsilon_t \tag{73a}$$

The long-term dynamic is given as in Eq 73(b) as cointegration of two-time series suggests that there is a long-run, or equilibrium, relationship between them.

$$P_t = a\,G_t + b + \varepsilon_t \tag{73b}$$

Eq 73(b) is deduced from Eq 73(a) as for the long term by using $P_{t-1} = P_t$ and $G_{t-1} = G_t$.

This EG method is valid when the number of variables N is equal to two but as N > 2, up to N-1 cointegration relations can exist. Therefore, EG method is a limited technique, as the study of N variables simultaneously does not make it possible to distinguish several cointegration relations. To overcome this difficulty, the study of a multivariate approach of Johansen cointegration is proposed as discussed in the next section.

4.5 Johansen VECM cointegration

The Johansen test can be considered as a multivariate generalization of the augmented Dickey-Fuller test, but the former is a strategic test that makes it possible to estimate all

cointegrating vectors when more than two variables are considered. In this study, the Johansen test is applied to 5 variables where Power (P) is the dependent variable, whereas irradiation (IRRA), temperature (Temp), wind speed (Wind), humidity (Humi) are explanatory variables. Indications in brackets are notation used in Eviews software for this study.

It should be recalled that if non-stationary series are integrated of the first order I(1) and found to be cointegrated, a vector error correction mechanism (VECM) can be used so as to enable the examination of short run as well as long-run dynamics of the cointegration series. This is the subject of the next section.

4.5.1 Multiple cointegration equation

Considering a vector auto regression (VAR) Pt of order p and N variables of non-stationary I(1) as given in Eq 74:

$$P_t = A_1 P_{t-1} + A_2 P_{t-2} + ... A_p P_{t-p} + \varepsilon_t \quad (74)$$

where the matrix rank are respectively, Pt (N,1), A1 (N, N), Pt -1 (N,1)……. Ap (N, N), Pt (N,1) and εt (N,1). For example, considering 5 variables lagged 2, Eq 74 is transformed as in Eq 75:

$$P_t = A_1 P_{t-1} + A_2 P_{t-2} + \varepsilon_t \quad (75)$$

And in the matrix form as represented in Eq 76:

$$\begin{bmatrix} P_{1t} \\ P_{2t} \\ P_{3t} \\ P_{4t} \\ P_{5t} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix} \begin{bmatrix} P_{1t-1} \\ P_{2t-1} \\ P_{3t-1} \\ P_{4t-1} \\ P_{5t-1} \end{bmatrix} + \begin{bmatrix} a_{16} & a_{17} & a_{18} & a_{19} & a_{110} \\ a_{26} & a_{27} & a_{28} & a_{29} & a_{210} \\ a_{36} & a_{37} & a_{38} & a_{39} & a_{310} \\ a_{46} & a_{47} & a_{48} & a_{49} & a_{410} \\ a_{56} & a_{57} & a_{58} & a_{59} & a_{510} \end{bmatrix} \begin{bmatrix} P_{1t-2} \\ P_{2t-2} \\ P_{3t-2} \\ P_{4t-2} \\ P_{5t-2} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \\ \epsilon_{4t} \\ \epsilon_{5t} \end{bmatrix} \quad (76)$$

The system equations are given in Eq 77:

$$\begin{aligned} P_{1t} &= a_{11} P_{1t-1} + a_{12} P_{2t-1} \ldots a_{15} P_{5t-1} + a_{16} P_{1t-2} + \cdots a_{110} P_{5t-2} + \epsilon_{1t} \\ P_{2t} &= a_{21} P_{1t-1} + a_{22} P_{2t-1} \ldots a_{25} P_{5t-1} + a_{26} P_{1t-2} + \cdots a_{210} P_{5t-2} + \epsilon_{2t} \\ P_{3t} &= a_{31} P_{1t-1} + a_{32} P_{2t-1} \ldots a_{35} P_{5t-1} + a_{36} P_{1t-2} + \cdots a_{310} P_{5t-2} + \epsilon_{3t} \\ P_{4t} &= a_{41} P_{1t-1} + a_{42} P_{2t-1} \ldots a_{45} P_{5t-1} + a_{46} P_{1t-2} + \cdots a_{410} P_{5t-2} + \epsilon_{4t} \\ P_{5t} &= a_{51} P_{1t-1} + a_{52} P_{2t-1} \ldots a_{55} P_{5t-1} + a_{56} P_{1t-2} + \cdots a_{510} P_{5t-2} + \epsilon_{5t} \end{aligned} \quad (77)$$

This first difference VAR (2) model can be written in a vector error correction model (VECM) as a function of only P_{t-1} as in Eq 78:

$$\Delta P_t = - A_2 \Delta P_{t-1} + \Pi P_{t-1} + \varepsilon_t \quad (78)$$

where $\Pi = A_1 + A_2 - I$ and I is the unit matrix. Eq 77 can also be written as function of Pt-1 and Pt-2 as given in Eq 79:

$$\Delta P_t = (A_1 - I) \Delta P_{t-1} + \Pi P_{t-2} + \varepsilon_t \quad (79)$$

If the coefficient matrix Π has reduced rank r < k, where k is the vector variables of I(1), r is the number of cointegration equations. The matrix Π can be written in terms of vector of adjustment parameters α and matrix of cointegration vectors β' given by Eq 80:

$$\Pi = \alpha\,\beta'\text{, where } \beta' P_t \text{ is } I(0) \tag{80}$$

where α is a (N,r) matrix with r < N, and β' has r cointegration vectors such that 0 < r < N as to highlight the VECM model. If this is applied for N = 5 as for this study. It results in Eq 81:

$$\begin{bmatrix}\Delta P_{1t}\\ \Delta P_{2t}\\ \Delta P_{3t}\\ \Delta P_{4t}\\ \Delta P_{5t}\end{bmatrix} = -\begin{bmatrix}a_{16} & a_{17} & a_{18} & a_{19} & a_{110}\\ a_{26} & a_{27} & a_{28} & a_{29} & a_{210}\\ a_{36} & a_{37} & a_{38} & a_{39} & a_{310}\\ a_{46} & a_{47} & a_{48} & a_{49} & a_{410}\\ a_{56} & a_{57} & a_{58} & a_{59} & a_{510}\end{bmatrix}\begin{bmatrix}\Delta P_{1t-1}\\ \Delta P_{2t-1}\\ \Delta P_{3t-1}\\ \Delta P_{4t-1}\\ \Delta P_{5t-1}\end{bmatrix} + \begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\\ a_{31} & a_{32}\\ a_{41} & a_{42}\\ a_{51} & a_{52}\end{bmatrix}\begin{bmatrix}\beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} & \beta_{15}\\ \beta_{21} & \beta_{22} & \beta_{23} & \beta_{24} & \beta_{25}\end{bmatrix}\begin{bmatrix}P_{1t-1}\\ P_{2t-1}\\ P_{3t-1}\\ P_{4t-1}\\ P_{5t-1}\end{bmatrix} + \begin{bmatrix}\epsilon_{1t}\\ \epsilon_{2t}\\ \epsilon_{3t}\\ \epsilon_{4t}\\ \epsilon_{5t}\end{bmatrix} \tag{81}$$

To estimate a VECM model, the matrix rank must be equal to r, meaning that Π has r non zero Eigen values and thus β'.

The Johansen test and estimation strategy which is a maximum likelihood test makes it possible to estimate all cointegrating vectors for N variables, which all have unit roots and there are at most N-1cointegrating vectors. The Johansen test provides estimates of all cointegrating vectors if cointegration relationship does exist, and a rank test is useful. Thereby, if:

- Rank (Π) = 0, then r = 0 meaning that none cointegration relationship and VECM cannot be applied,
- Rank (Π) = r, and meaning that variables are cointegrated and the number of cointegration relationship is equal to r. VECM model can be estimated.
- Rank (Π) = N, meaning that none cointegration relationship.

Johansen procedure is based on the maximum Eigenvalue and Trace tests that are conducted on the error correction model foundation. For both test statistics, the initial Johansen test is a null hypothesis test of no cointegration against the alternative of cointegration.

The first test of maximum Eigenvalues is to determine whether the rank of the matrix is zero, and the null hypothesis is rank (Π) = 0 whereas the alternative hypothesis is rank (Π) = 1.

The second test of Trace is to determine whether the rank of the matrix is r0, the null hypothesis is rank (Π) = r0 and the alternative hypothesis is that r0 < rank (Π) ≤ r, where r is the maximum number of possible cointegration vectors.

The Johansen technique described in this section is basic and further discussions or more technical details are beyond the scope of this paper. Interested readers can consult literatures (Johansen, 1988, 1991; Kitamura, 1998).

The Johansen test that will be conducted in the following sections is summarized below in five steps.

- Step 1: Performing series stationarity (correlogram & ADF) tests to determine whether there is cointegration relationship or not.
- Step 2: If step1 is true, meaning that series are of the same order of integration and cointegration is likely, therefore VECM model can be estimated. Determining the lag length using Akaike and Schwarz criteria.
- Step 3: Implementing the Johansen test to determine the amount of cointegration relationships.
- Step 4: Identifying the cointegration relationships or long-term relationships between variables.

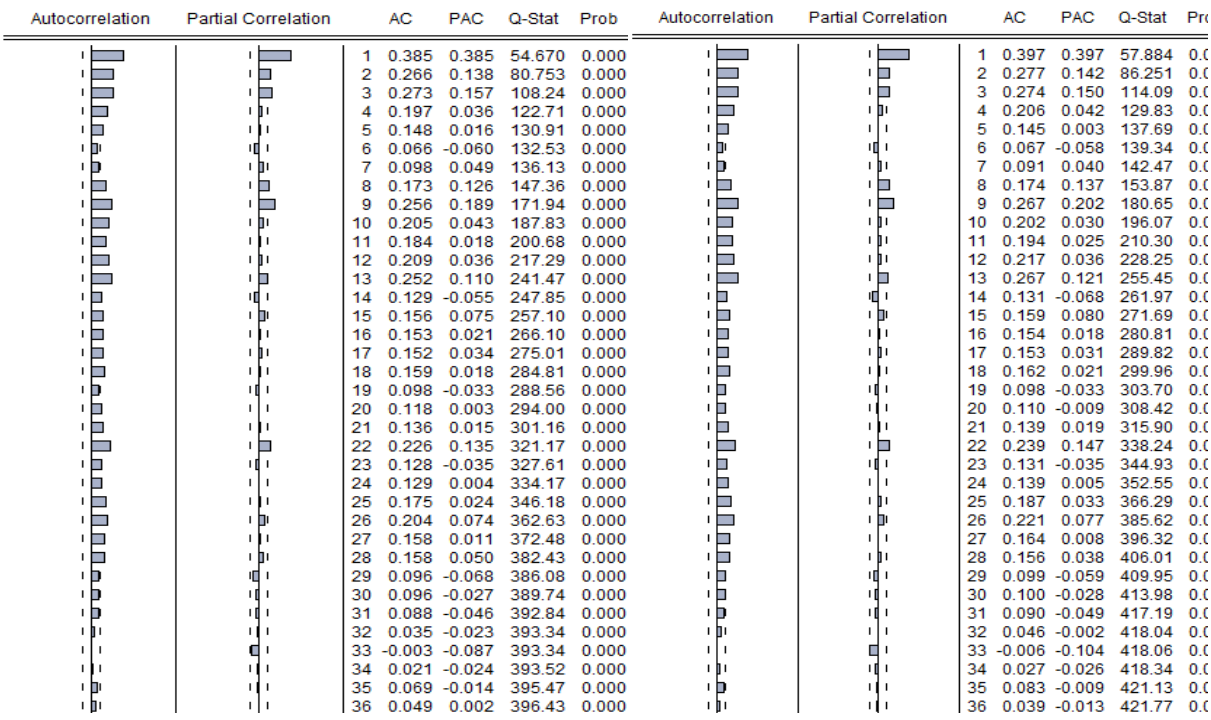
- Step 5: Estimating the VECM model by maximum likelihood method, test validations by visual diagnostic or correlogram, and checking that residuals from the model are white noise.

4.6 Applying Johansen tests to experimental data

Data that is used for this study comes from a building-mounted PV plant designed for a grid connected system. Inclined at 21°, which is Reunion Island latitude, for optimal energy extraction, the modules that make up the PV plant are polycrystalline type of 180 W each, equipped with solar, temperature and wind sensors. For this study, the sample of one-year daily mean data of year 2012, with 365 observations, is retrieved among 7 years of measurements from the COREX building, located at La Possession in the west coast of the island. The determined cointegrating relationship is then applied and compared to other data in real conditions for the years 2013 to 2018. With a 10-minute sampling step, this represents more than 17,000 data per year. This cointegration relationship is also applied to the second half of year 2019 to make the PV output prediction. The following notations are used for each variable: POWER, IRRA, TEMP, WIND, and HUMI.

4.6.1 Visual diagnostic of stationary series of the 5 variables

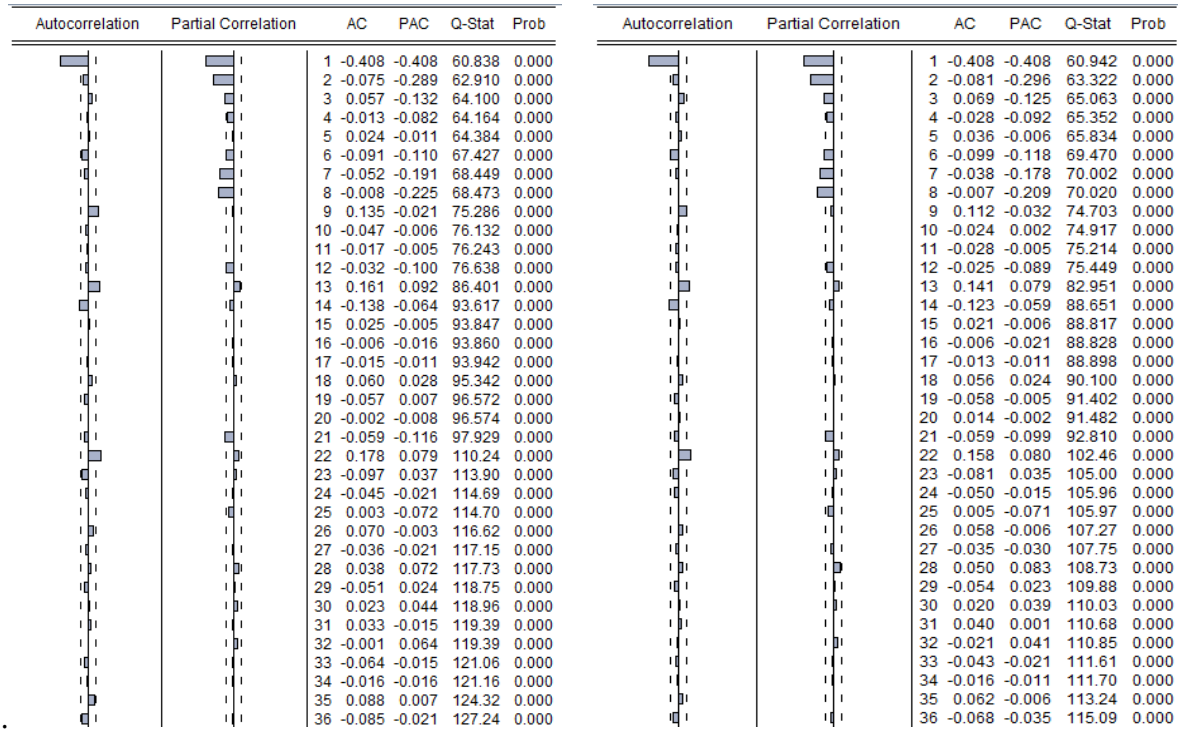
As mentioned in step 1, the following Figures 4.5 (a) and (b) show the correlograms of non-stationary series of Power and Irra (irradiation) at level as explained in section 4.3.2.3. Similar figures for other variables at level such as Temp, Wind and Humi (humidity) are given in appendix 1. The autocorrelation coefficient starts at a high value and declines very slowly toward zero as the lag increases the autocorrelation coefficients at various lags are high even up to lag 26 for correlograms. The last values in the Q-stat columns are significant indicating serial correlation in the residuals. More precisely, if we consider that at level each series is a non-stationary series as a visual diagnostic. Figures 4.6 (a) and (b) show stationary series of first difference of variable Power and Irra, as the spikes are beyond the vertical line of the autocorrelation column, except for the first value, which means that we have to consider the first lag as it will be indicated by the ADF test. These correlograms seem to indicate white noise time series. Similar diagrams of first difference of variables Temp, Wind & Humi are shown in appendix 2.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.385	0.385	54.670	0.000
		2	0.266	0.138	80.753	0.000
		3	0.273	0.157	108.24	0.000
		4	0.197	0.036	122.71	0.000
		5	0.148	0.016	130.91	0.000
		6	0.066	-0.060	132.53	0.000
		7	0.098	0.049	136.13	0.000
		8	0.173	0.126	147.36	0.000
		9	0.256	0.189	171.94	0.000
		10	0.205	0.043	187.83	0.000
		11	0.184	0.018	200.68	0.000
		12	0.209	0.036	217.29	0.000
		13	0.252	0.110	241.47	0.000
		14	0.129	-0.055	247.85	0.000
		15	0.156	0.075	257.10	0.000
		16	0.153	0.021	266.10	0.000
		17	0.152	0.034	275.01	0.000
		18	0.159	0.018	284.81	0.000
		19	0.098	-0.033	288.56	0.000
		20	0.118	0.003	294.00	0.000
		21	0.136	0.015	301.16	0.000
		22	0.226	0.135	321.17	0.000
		23	0.128	-0.035	327.61	0.000
		24	0.129	0.004	334.17	0.000
		25	0.175	0.024	346.18	0.000
		26	0.204	0.074	362.63	0.000
		27	0.158	0.011	372.48	0.000
		28	0.158	0.050	382.43	0.000
		29	0.096	-0.068	386.08	0.000
		30	0.096	-0.027	389.74	0.000
		31	0.088	-0.046	392.84	0.000
		32	0.035	-0.023	393.34	0.000
		33	-0.003	-0.087	393.34	0.000
		34	0.021	-0.024	393.52	0.000
		35	0.069	-0.014	395.47	0.000
		36	0.049	0.002	396.43	0.000

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Pr
		1	0.397	0.397	57.884	0.0
		2	0.277	0.142	86.251	0.0
		3	0.274	0.150	114.09	0.0
		4	0.206	0.042	129.83	0.0
		5	0.145	0.003	137.69	0.0
		6	0.067	-0.058	139.34	0.0
		7	0.091	0.040	142.47	0.0
		8	0.174	0.137	153.87	0.0
		9	0.267	0.202	180.65	0.0
		10	0.202	0.030	196.07	0.0
		11	0.194	0.025	210.30	0.0
		12	0.217	0.036	228.25	0.0
		13	0.267	0.121	255.45	0.0
		14	0.131	-0.068	261.97	0.0
		15	0.159	0.080	271.69	0.0
		16	0.154	0.018	280.81	0.0
		17	0.153	0.031	289.82	0.0
		18	0.162	0.021	299.96	0.0
		19	0.098	-0.033	303.70	0.0
		20	0.110	-0.009	308.42	0.0
		21	0.139	0.019	315.90	0.0
		22	0.239	0.147	338.24	0.0
		23	0.131	-0.035	344.93	0.0
		24	0.139	0.005	352.55	0.0
		25	0.187	0.033	366.29	0.0
		26	0.221	0.077	385.62	0.0
		27	0.164	0.008	396.32	0.0
		28	0.156	0.038	406.01	0.0
		29	0.099	-0.059	409.95	0.0
		30	0.100	-0.028	413.98	0.0
		31	0.090	-0.049	417.19	0.0
		32	0.046	-0.002	418.04	0.0
		33	-0.006	-0.104	418.06	0.0
		34	0.027	-0.026	418.34	0.0
		35	0.083	-0.009	421.13	0.0
		36	0.039	-0.013	421.77	0.0

Figure 4.5 (a) POWER correlogram at level; (b) IRRA correlogram at level



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.408	-0.408	60.838	0.000
		2	-0.075	-0.289	62.910	0.000
		3	0.057	-0.132	64.100	0.000
		4	-0.013	-0.082	64.164	0.000
		5	0.024	-0.011	64.384	0.000
		6	-0.091	-0.110	67.427	0.000
		7	-0.052	-0.191	68.449	0.000
		8	-0.008	-0.225	68.473	0.000
		9	0.135	-0.021	75.286	0.000
		10	-0.047	-0.006	76.132	0.000
		11	-0.017	-0.005	76.243	0.000
		12	-0.032	-0.100	76.638	0.000
		13	0.161	0.092	86.401	0.000
		14	-0.138	-0.064	93.617	0.000
		15	0.025	-0.005	93.847	0.000
		16	-0.006	-0.016	93.860	0.000
		17	-0.015	-0.011	93.942	0.000
		18	0.060	0.028	95.342	0.000
		19	-0.057	0.007	96.572	0.000
		20	-0.002	-0.008	96.574	0.000
		21	-0.059	-0.116	97.929	0.000
		22	0.178	0.079	110.24	0.000
		23	-0.097	0.037	113.90	0.000
		24	-0.045	-0.021	114.69	0.000
		25	0.003	-0.072	114.70	0.000
		26	0.070	-0.003	116.62	0.000
		27	-0.036	-0.021	117.15	0.000
		28	0.038	0.072	117.73	0.000
		29	-0.051	0.024	118.75	0.000
		30	0.023	0.044	118.96	0.000
		31	0.033	-0.015	119.39	0.000
		32	-0.001	0.064	119.39	0.000
		33	-0.064	-0.015	121.06	0.000
		34	-0.016	-0.016	121.16	0.000
		35	0.088	0.007	124.32	0.000
		36	-0.085	-0.021	127.24	0.000

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.408	-0.408	60.942	0.000
		2	-0.081	-0.296	63.322	0.000
		3	0.069	-0.125	65.063	0.000
		4	-0.028	-0.092	65.352	0.000
		5	0.036	-0.006	65.834	0.000
		6	-0.099	-0.118	69.470	0.000
		7	-0.038	-0.178	70.002	0.000
		8	-0.007	-0.209	70.020	0.000
		9	0.112	-0.032	74.703	0.000
		10	-0.024	0.002	74.917	0.000
		11	-0.028	-0.005	75.214	0.000
		12	-0.025	-0.089	75.449	0.000
		13	0.141	0.079	82.951	0.000
		14	-0.123	-0.059	88.651	0.000
		15	0.021	-0.006	88.817	0.000
		16	-0.006	-0.021	88.828	0.000
		17	-0.013	-0.011	88.898	0.000
		18	0.056	0.024	90.100	0.000
		19	-0.058	-0.005	91.402	0.000
		20	0.014	-0.002	91.482	0.000
		21	-0.059	-0.099	92.810	0.000
		22	0.158	0.080	102.46	0.000
		23	-0.081	0.035	105.00	0.000
		24	-0.050	-0.015	105.96	0.000
		25	0.005	-0.071	105.97	0.000
		26	0.058	-0.006	107.27	0.000
		27	-0.035	-0.030	107.75	0.000
		28	0.050	0.083	108.73	0.000
		29	-0.054	0.023	109.88	0.000
		30	0.020	0.039	110.03	0.000
		31	0.040	0.001	110.68	0.000
		32	-0.021	0.041	110.85	0.000
		33	-0.043	-0.021	111.61	0.000
		34	-0.016	-0.011	111.70	0.000
		35	0.062	-0.006	113.24	0.000
		36	-0.068	-0.035	115.09	0.000

Figure 4.6 (a) POWER correlogram at first difference; (b) IRRA correlogram at first difference

4.6.2 Augmented dickey fuller test of the 5 variables

The ADF test is applied to verify the null hypothesis of whether there is a unit root in a time series as explained in section 4.3.2.2. In this section, only the stationary outcome of each series is displayed and the header portion of each Table 4.7 and 4.8 indicates the null hypothesis test for Power and IRRA and is rejected if the ADF t-statistic value is less at 5% significant level. Similar figures of the ADF test at first difference for other variables of Temp, Wind and Humi are indicated in appendix 3.

Table 4.7 First difference of power

Null Hypothesis : D(POWER) has a unit root Exogenous : None Lag length : 7 (Automatic-based on SIC, maxlag = 16)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		−12,51960	0.0000
Test critical values	1% level	−2,571643	
	5% level	−1,941740	
	10% level	−1,616087	

Table 4.8 First difference of irradiance.

Null Hypothesis: D(IRRA) has a unit root Exogenous: None Lag length: 7 (Automatic-based on SIC, maxlag = 16)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		−12,70921	0.0000
Test critical values	1% level	−2,571643	
	5% level	−1,941740	
	10% level	−1,616087	

As all series are of the same order of integration that is I(1), cointegration is probable (OR expected). Therefore, VECM model can be estimated. The next section goes to step 2.

4.6.3 Lag length determination

As indicated in section 4.3.2.7(d), the AIC and SC are used to determine the lag length. As a reminder, lower is AIC value, better is the model. To determine the lag length, it is assumed that variables are not cointegrated, and the unrestricted vector auto-regression is processed under the Eviews software with a corresponding number of lag. The outcome for lag 1, that is p =1, is indicated in Table 4.9 and 4.10.

Table 4.9 First part of the VAR result.

	Power	**IRRA**	**TEMP**	**HUMI**	**WIND**
POWER(-1)	0.138750	−0.000269	0.000471	4.33E-06	−1.18E-05
	(0.34297)	(0.01845)	(0.00053)	(6.9E-06)	(8.3E-05)
	[0.40456]	[−0.01457]	[0.89419]	[0.62533]	[−0.14099]
IRRA(-1)	3.792247	0.391502	−0.011122	−9.36E-05	0.001265
	(6.65388)	(0.35798)	(0.011122)	(0.00013)	(0.00162)
	[0.56993]	[1.09363]	[-1.08825]	[-0.69673]	[0.78154]
TEMP(-1)	24.43531	−0.004351	0.507372	0.001161	0.032680
	(73.5611)	(3.95764)	(0.11299)	(0.00148)	(0.01790)
	[0.33218]	[−0.00110]	[4.49021]	[0.78181]	[−1.82579]
HUMI(-1)	2617.939	120.9769	6.474258	0.627388	0.822792
	(2120.54)	(114.086)	(3.25701)	(0.04280)	(0.51597)
	[1.23456]	[1.06040]	[1.98779]	[14.6579]	[1.59466]
WIND(-1)	359.7670	23.23764	0.592929	0.003879	0.422704
	(198.714)	(10.6910)	(0.30521)	(0.00401)	(0.04835)
	[1.81048]	[2.17358]	[1.94268]	[−0.96722]	[8.74240]
C	1901.391	138.9732	14.75872	0.229401	1.655465
	(2202.29)	(118.485)	(3.38257)	(0.04445)	(0.53586)
	[0.86337]	[1.17292]	[4.36317]	[5.16061]	[3.08936]
R-squared	0.1666062	0.175396	0.230779	0.428917	0.199223
Adj.R-squared	0.154317	0.163782	0.219945	0.420874	0.187944
Sum sq.resids	2.63E+09	7616334	6207.477	1.072036	155.7846
S.E equation	2722.517	146.4734	4.181609	0.054953	0.662442
F-statistic	14.13824	15.10190	21.30117	53.32521	17.66386
Log likelihood	−3364.473	−2309.462	−1025.693	538.1501	−360.5440
Akaike AIC	18.67298	12.82805	5.715750	−2.948200	2.030714
Schwarz SC	18.73762	12.89268	5.780385	−2.883565	2.095350
Mean dependent	8530.460	458.3989	39.97784	0.69945	2.43440
D dependent	2960.511	160.1766	4.734572	0.072211	0.735115

Table 4.10 Second part of the VAR result.

Determinant resid covariance (dof adj.)	13270892
Determinant resid covariance	12204103
Log likelihood	-5506.454
Akaike information criterion	30.67287
Schwarz criterion	30.99605

The AIC value indicated in Table 9 is for each system. For Power as the dependent variable, the corresponding AIC is 18.67, whereas in Table 4.10, the value of the AIC is for the whole VAR system and this is the chosen one. For the test with lag = 2, AIC value of the system is

greater than for p =1. The result obtained is not represented here. In Table 4.9, it can be noted that values in brackets '()' are standard errors while values in square brackets '[ ]' are the corresponding t-statistic value. Table 4.11 shows the result obtained for another test that confirms the order determination by comparing various criteria.

Table 4.11 Different lag criteria

Lag	LogL	LR	FPE	AIC	SC	HQ
0	−5748.302	NA	47911324	31.87425	31.92811	31.89566
1	−5506.454	475.6570*	14411011*	30.67287*	30.99605*	30.80136*

The star "*" indicates the lag order selected by the following criterion and their corresponding definition: LR is the likelihood ratio, FPE is final prediction error and HQ is the Hannan-Quinn criterion. Therefore, the lag length for p = 1 will be used from now on in this study. The next section goes to step 3.

4.6.4 Determining of the number of cointegration relationships

As explained in section 4.4.2, the number of cointegration relationships is based on both the Trace and the Eigen value tests. This is reported in two blocks through Eviews software, denoted the Trace statistics and the maximum Eigenvalue statistics, respectively. For this study, this test is performed with the deterministic trend assumption, which means that there is no intercept or trend in the cointegration equation or VAR test. The two outcome blocks of this test through Eviews are indicated in Table 4.12 and 4.13. In Table 4.12, the header portion indicates the concerning series with the lag length equal to one and no deterministic trend. The columns of each block are as follows. The first column is the number of cointegration relations under the null hypothesis. The second column is the ordered Eigenvalues of the Π matrix as explained in section 4.4.2. The third column is the test statistic and the fourth column is the 5% critical value. The Trace test indicates 4 cointegration equations at 5% significant level as the probability value is nearly 47% and greater than 5%. The star "*" denotes rejection of the hypothesis at 5% level. Therefore, all variables such POWER, IRRA, TEMP, level of HUMI and SWIND are linked by a long run relationship.

Table 4.12 Cointegration Trace test.

Trend assumption: No deterministic trend Series: POWER IRRA TEMP level of HUMI WIND Lags interval (in first differences): 1 to 1 Unrestricted Cointegration Rank Test (Trace)				
Hypothesized No. of CE(s)	EigenValue	Trace Statistic	0.05 Crititical Value	Prob.**
None*	0.304862	371.9636	60.06141	0.0001
At most 1*	0.245336	241.7790	40.17493	0.0001
At most 2*	0.207530	141.0082	24.27596	0.0001
At most 3*	0.147316	57.73709	12.32090	0.0000
At most 4	0.001909	0.684013	4.129906	0.4677

Table 4.13 Cointegration Eigenvalue test.

Unrestricted Cointegration Rank Test (Maximum Eigen Values)				
Hypothesized No. of CE(s)	EigenValue	Max-Eigen Statistic	0.05 Crititical Value	Prob.**
None*	0.304862	130.1847	30.43961	0.0001
At most 1*	0.245336	100.7708	24.15921	0.0001
At most 2*	0.207530	83.27109	17.79730	0.0001
At most 3*	0.147316	57.05307	11.22480	0.0000
At most 4	0.001909	0.684013	4.129906	0.4677

In Table 4.13, the maximum eigenvalue test indicates four equations at 5% level. The outcome of this test is in line with what is indicated in section 4.5.1.The Johansen VECM test is then performed through Eviews with one lagged. The final outcome is given in Table 4.14 - 4.16. The entire Eq 76 in section 4.5.1 can be deduced from Table 10. The target model D (POWER) which is the dependent variable given in Eq 82 (a) (b), (c) between Table 10a & c. D (POWER) is identified as ΔP.

$$\Delta P = -0.738\ CointEq1 + 2.76\ CointEq2 + 29.00\ CointEq3 + 3493.349\ CointEq4 + \varepsilon_{it} \quad (82a)$$

$$\Delta P - 0.738\ (P_{t-1} - 3521{,}54W_{t-1}) + 2.764(I_{t-1} - 189{,}05) + 29.004(T_{t-1} - 16.52W_{t-1}) + 3493.349\ (H_{t-1} - 0.289W_{t-1}) + \varepsilon_{1t} \quad (82b)$$

$$P_t = 833.33\ W_t + 3.74\ I_t + 36.38\ T_t + 4758.96H_t + \varepsilon_{1t} \quad (82c)$$

Eq 82(c) is the long-term relationship as each variable at (t-1) is equal to each variable at t

It should be noted that Eq 82(c) is determined with one outlier in square brackets '[ ]' removed from the number of observations. An outlier may be defined as an observation with a large residual that represents the difference (positive or negative) between the actual value and the estimated value from the regression model. When the residual is large, it is in comparison with the other residuals. Usually, a large residual catches attention because of its rather large vertical distance from the estimated regression line. The relationship of ΔP is deduced from Table 20(c) using the error correction model where the values in brackets '()' is the standard error, and the values in square brackets '[]' is the t statistic value. There is no probability value to determine whether each coefficient is significant.

Table 4.14 The four cointegration equations.

Cointegration Eq:	CointEq1	CointEq2	CointEq3	CointEq4
POWER(−1)	1.000000	0.000000	0.000000	0.000000
IRRA(−1)	0.000000	1.000000	0.000000	0.000000
TEMP(−1)	0.000000	0.000000	1.000000	0.000000
HUMI(−1)	0.000000	0.000000	0.000000	1.000000
WIND(−1)	−352.536	−189.0507	−16.52614	−0.289727
	(115.133)	(6.10357)	(0.43571)	(0.00856)
	[−30.5867]	[−30.9738]	[−37.9295]	[−33.8617]

Table 4.15 The error correction coefficients.

Error Correction	D(POWER)	D(IRRA)	D(TEMP)	D(HUMI)	D(WIND)
CointEq1	−0.738061	0.003671	0.000668	−6.30E-06	−0.000101
	(0.44968)	(0.02418)	(0.00069)	(9.4E-06)	(0.00011)
	[−1.64130]	[0.15180]	[0.96372]	[−0.66729]	[−0.90932]
CointEq2	2.764963	−0.633466	-0.020299	−6.30E-06	0.001919
	(8.49770)	(0.45697)	(0.01309)	(0.00018)	(0.00211)
	[0.32538]	[−1.38624]	[-1.55056]	[−0.35314]	[0.91023]
CointEq3	29.00405	1.202475	−0.168783	0.007478	0.014007
	(57.8867)	(3.11288)	(0.08918)	(0.00122)	(0.56441)
	[0.50105]	[0.38629]	[−1.89266]	[6.15305]	[0.97515]
CointEq4	3493.349	170.2571	10.64644	−0.318148	1.006675
	(2274.49)	(122.312)	(3.50397)	(0.04776)	(0.56441)
	[1.53588]	[1.39200]	[3.03840]	[−6.66201]	[1.78360]
D(POWER(−1))	−0.061329	0.000548	−0.000373	8.67E−06	0.000102
	(0.34563)	(0.01859	(0.00053)	(7.3E−06)	(8.6E−05)
	[−0.17744]	[0.02947]	[−0.70012]	[1.19434]	[1.18555]
D(IRRA(-1))	−0.385590	−0.119264	0.009972	−4.70E-05	−0.001840
	(6.98093)	(0.37540)	(0.01075)	(0;00015)	(0.00173)
	[−0.05523]	[−0.31770]	[0.92728]	[−0.32055]	[−1.06198]
D(TEMP(-1))	−41.98734	−1.471655	−0.312499	−0.004109	−0.012250
	(79.2692)	(4.26273)	(0.12212)	(0.00166)	(0.01967)
	[−0.52968]	[−0.34524]	[−2.55899]	[−2.46904]	[−0.62278]
D(HUMI(-1))	−2443.284	−91.71841	−4.715188	−0.009332	0.646033
	(2670.53)	(143.609)	(4.11409)	(0.05607)	(0.66268)
	[-0.52968]	[−0.63867]	[−1.14611]	[−0.16643]	[0.97487]
D(WIND(-1))	−388.2601	−22.43381	−0.629913	−0.003435	−0.025043
	(219.111)	(11.7828)	(0.33755)	(0.00460)	(0.05437)
	[−1.77198]	[−1.90395]	[−1.86612]	[−0.74676]	[−0.46059]

Table 4.16 Statistical data of the outcome with the AIC.

R-squared	0.339711	0.335370	0.313075	0.152640	0.280574
Adj.R-Squared	0.324575	0.320135	0.297329	0.133216	0.264083
Sum sq.resids	2.54E+09	7335040	6019.884	1.118186	156.1900
S.E equation	2695.906	144.9735	4.153186	0.0566604	0.668981
F-statistic	22.44456	22.01305	19.88265	7.858403	17.01326
Log likelihood	−3331.440	-2285.028	−1013.170	524.6397	−359.5067
Akaike AIC	18.66167	12.81580	5.710447	−2.880669	2.058697
Schwarz SC	18.75923	12.91335	5;808002	−2.783113	2.156252
Mean dependent	−24.29050	−1.402235	−0.064246	−0.000726	−0.002601
S.D dependent	3280.321	175.8236	4.954562	0.060789	0.779829
Determinant resid covariance (dof adj)			14268562		
Determinant resid covariance			12562967		
Log likelihood			−5465.881		
Akaike information criterion			30.89878		
Schwarz criterion			31.60334		

This is done by using system equations through Eviews, where the residual of the cointegration equation can be derived when D (POWER) is the dependent variable. This allowed to determine the residual of the cointegration equation as given in Eq 83:

$$\begin{aligned}\Delta P = {} & C(1) \times (P(-1) - 3526.23\, W(-1) + C(2) \times I(-1) - 189.36 \times W(-1) + C(3) \times (T(-1) - 16.44 \\ & \times W(-1) + C(4) \times (H(-1) - 0.2869 \times W(-1)) + C(5) \times D(P(-1)) + C(6) \\ & \times D(I(-1) + C(7) \times D(T(-1)) + C(8) \times D(H(-1)) + C(9) \times D(W(-1))\end{aligned} \tag{83}$$

The probability of each coefficient C (1) to C (9) is given in Table 4.17.

Table 4.17 Cointegration coefficient with the corresponding probability.

	Coefficient	Std.Error	t-Statitic	Prob.
C(1)	−0.732626	0.450017	-1.627997	0.1044
C(2)	2.741201	8.504443	0.322326	0.7474
C(3)	26.67434	57.90265	0.460676	0.6453
C(4)	3488.323	2276.294	1.532457	0.1263
C(5)	−0.049931	0.345787	−0.144399	0.8853
C(6)	−0.639548	6.983515	−0.091580	0.9271
C(7)	−43.51793	79.32281	−0.548618	0.5836
C(8)	−2442.330	2672.656	−0.913822	0.3614
C(9)	−349.8633	217.1123	−1.611440	0.1080

From sections 4.4.4 and 4.4.5, the coefficient of C(1) which is the speed of adjustment towards the long run relationship, must be of negative sign and statistically significant whereas coefficients from C(2) to C(9) are short run coefficients. Negative implies a departure in one direction. The correction would have to pull back to the other direction. In this case, this is satisfying for the model as it implies that the model is converging in the long-run equilibrium. To test the short run causality, the Wald test was performed as given in the next section.

4.6.5 Wald test

The Wald statistic test is a joint test for short run coefficients and the null hypothesis is that all short run coefficients are jointly zero. In this case C (2) = … C (9) = 0. This is given in Table 4.18, where probability of the chi-square value as explained in section 4.3.2.7, is greater than 5% significant level, meaning that there is no short-run relationship as all coefficients C(2) to C(9) are zero. The null hypothesis cannot be rejected.

Table 4.18 Wald statistic test for short-run equilibrium.

Wald Test Equation: Untitled			
Test Statistic	Value	df	Probability
F-statistic	1.932705	(5,355)	0.0882
Chi-square	9.663527	5	0.0854

4.6.6 Lagrange multiplier test and jarque bera statistic

The long run relationship of Eq.82(c) is significant. However, the residual property of white noise needs to be tested. This is verified using the LM test as described in section 4.3.2.5, and the outcome is given in Table 4.19.

Table 4.19 LM test of serial correlation.

Breush-Godfrey Serial Correlation LM Test			
F-Statistic	0.008380	Prob.F(1,341)	0.9271
Obs*R-squared	0.008847	Prob.Chi-Square(1)	0.9251

The observed R squared and the corresponding probability which is greater than 5% significant level mean that the null hypothesis can be rejected, and the AR model has serial correlation. To see if the residual is normally distributed, the Jarque Bera statistic is applied as displayed in Figure 4.7.

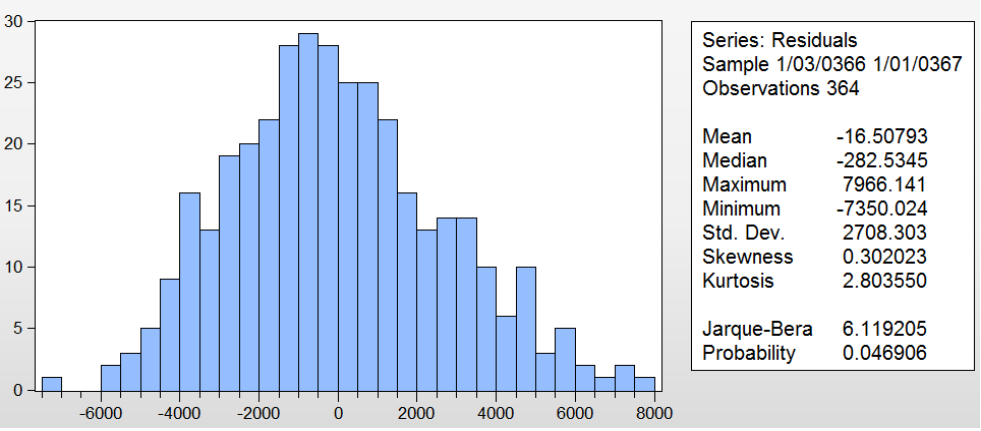


Figure 4.7 Jarque Bera residual normal distribution.

As explained in section 4.3.2.6, the Jarque Bera coefficient is very significant due to the large number of observations, and the bell shape indicates that the residual follows a normal distribution. The obtained model outcome is good. The stability diagnostic needs to be tested to make sure that the model is dynamically stable. For this purpose, the CUSUM test is performed.

4.6.7 The CUSUM test

The CUSUM test (Durbin Test) is based on the cumulative sum of the recursive residuals. It plots the cumulative sum together with the 5% critical lines. The test attain parameter stability if the cumulative sum goes inside the area between the two critical lines. In Figure 4.8, the blue curve, also known as the trade line, lies between the red boundaries. Therefore the model is set to be dynamically stable.

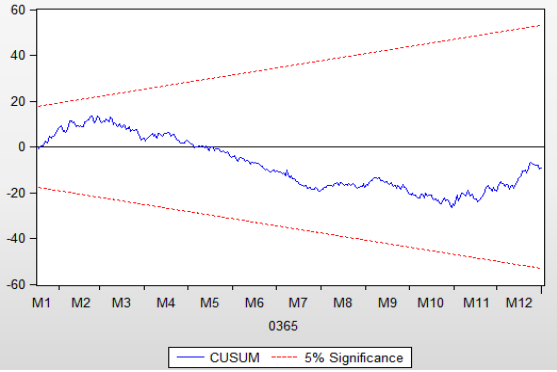


Figure 4.8 The CUSUM test.

The residual is a random or white noise process. The final long run relationship of this study linking the environmental parameters to the power output in a tropical zone is given in Eq 84:

$$P_t = 833.33 \times W_t + 3.74 \times I_t + 36.38 \times T_t + 4758.96 \times H_t \quad (84)$$

This equation is then applied for several years of data, and the outcome is compared to real data. This is given is the next section.

4.7 Experimental Results

We used the model cointegration regression of Eq 84 determined from year 2012 to calculate the power output for each year of 2013, 2014 and 2016. The goal was to design a model from data of year 2012 and trying to forecast the power output from the model for the following years. Then, we compared each year Johansen cointegration power output to the measured power output in real conditions of the corresponding year. Figures 4.9, 4.10, 4.11 represent the plot of each year with the corresponding R^2 value. The year sample time is 10 minutes giving more than 17,000 values per year.

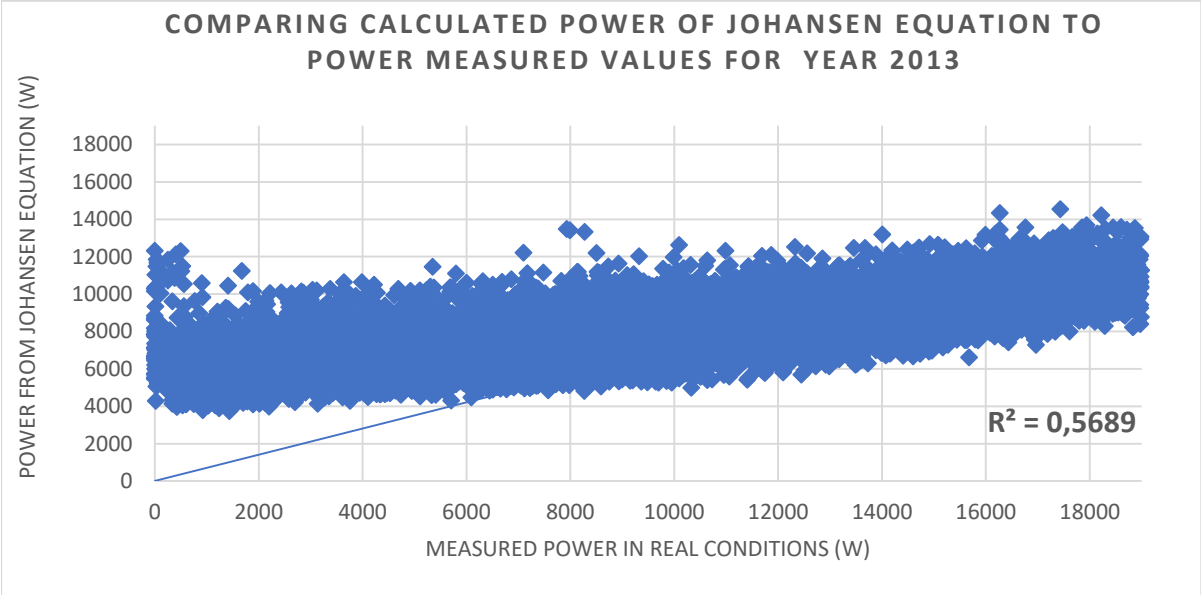


Figure 4.9 Comparing Model Power output to Measured Power for year 2013.

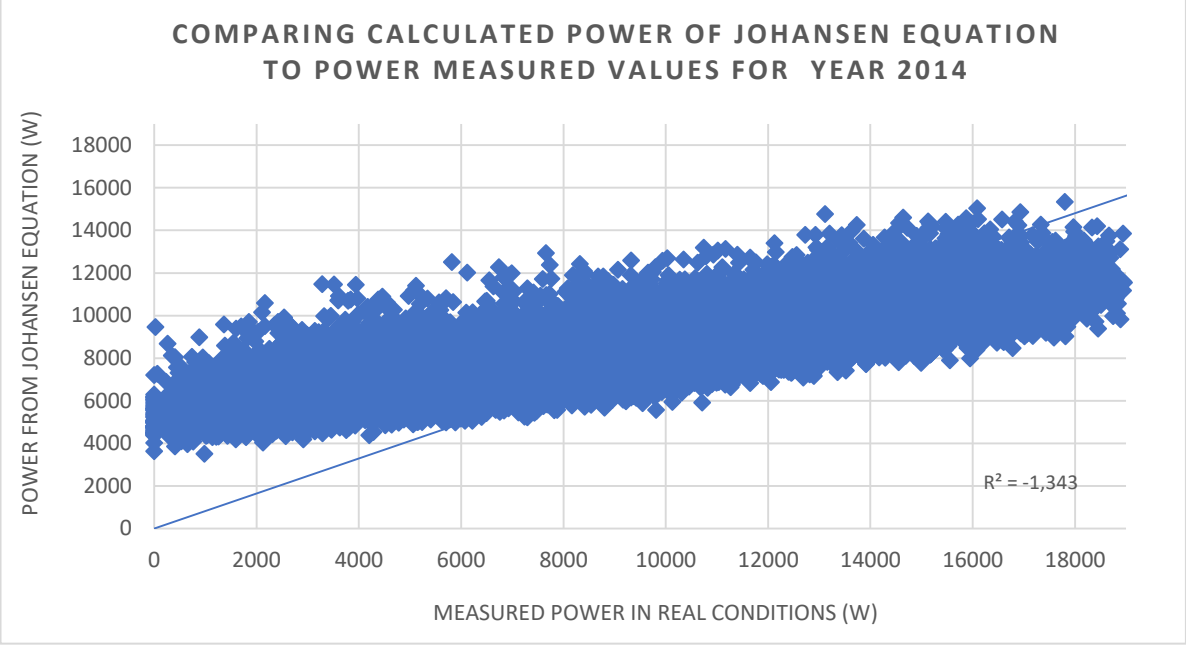


Figure 4.10 Comparing Model Power output to measured Power for year 2014

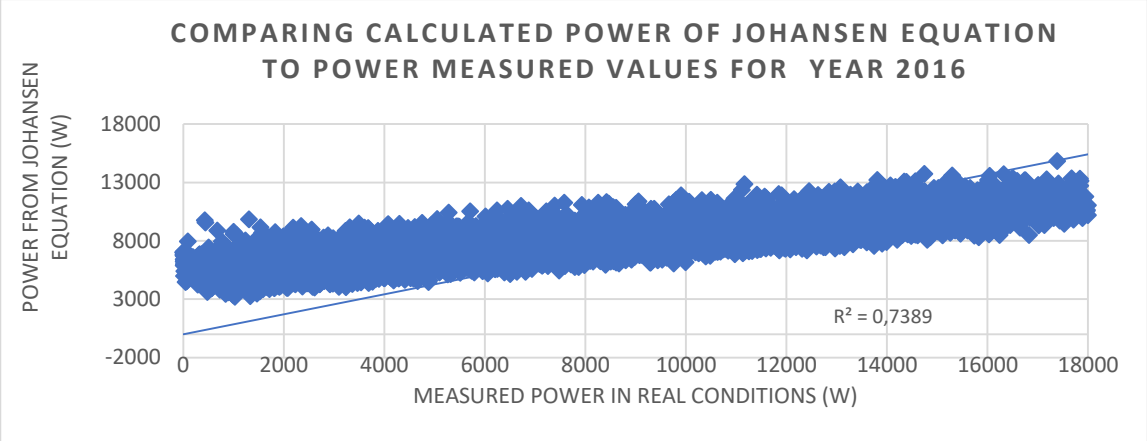
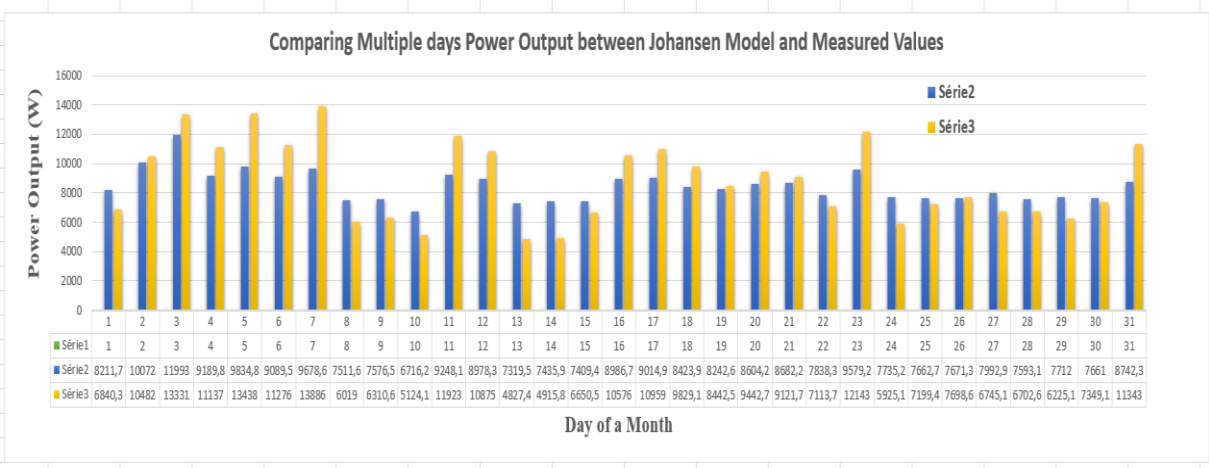


Figure 4.11 Comparing Model Power output to measured Power for year 2016

The accuracy of the fit for a regression model is characterized by the coefficient of determination R^2 or Pearson's correlation coefficient (Fu et al., 2020). It shows how correlated the forecasted and real values are. Applying the Johansen cointegration equation to the different years, it can be observed that R^2 is between nearly 65% and 74%. The R^2 value could have been better but the data has been randomly chosen. Some data are far from the regression lines as in Figure 4.10 and 4.11, but can be explained by the fact that these data were related to a few days before and after a storm period.

We then applied the Johansen cointegration model for a long-term forecasting that is multiple days ahead. This is represented in Figure 4.12, where the yellow and blue colors of the bar chart are respectively for the measured power output and Johansen model power output. The x axis is the day number of a particular month. For this test, we used the month of January 2016.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Série1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Série2	8211,7	10072	11933	9189,8	9834,8	9089,5	9678,6	7511,6	7576,5	6716,2	9248,1	8978,3	7319,5	7435,9	7409,4	8986,7	9014,9	8423,9	8242,6	8604,2	8682,2	7838,3	9579,2	7735,2	7662,7	7671,3	7992,9	7593,1	7712	7661	8742,3
Série3	6840,3	10482	13331	11137	13438	11276	13886	6019	6310,6	5124,1	11923	10875	4827,4	4915,8	6650,5	10576	10959	9829,1	8442,5	9442,7	9121,7	7113,7	12143	5925,1	7199,4	7698,6	6745,1	6702,6	6225,1	7349,1	11343

Figure 4.12 Comparing multiple days for long term forecasting.

The values below the bar chart diagram are real power output (yellow series) and model power output (blue series).

We also applied the test for an immediate-short-term forecasting that is hours ahead. This is represented in the bar chart diagram of Figure 4.13. The blue bar chart (serie 1) is the measured power output and the orange one (serie 2) is the Johansen cointegration power output model. The horizontal axis is an hourly interval of a particular day. For this test we have used data of 11th April 2014 from 9 a.m. to 5 p.m.

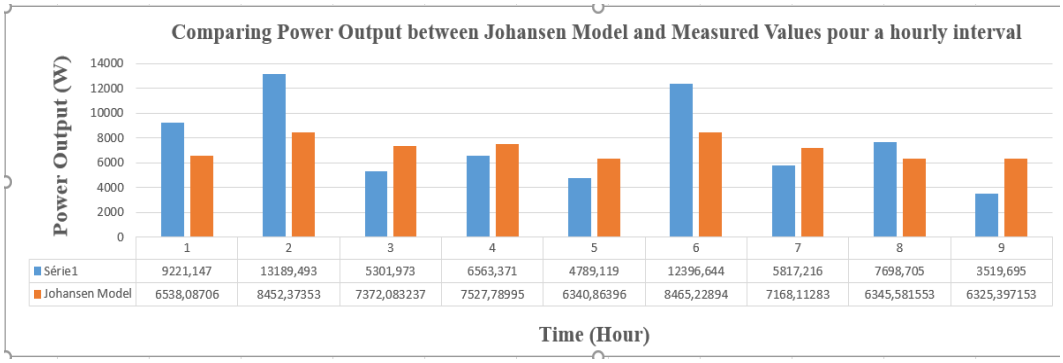


Figure 4.13 Comparing an hourly interval for immediate short term forecasting.

This is a promising model which obviously need to be improved nothing when comparing the bar chart diagrams at various time interval forecasting.

4.8 Discussion

The Johansen cointegration principle is an appropriate applied method to determine the cointegration relationship between PV output and environmental parameters such as solar irradiation, module temperature, wind speed and relative humidity.

The present work focuses on the methodology used for developing the forecasting method and therefore does not address more general questions such as the precise sensitivity of the environmental factors, the effect of a very long (multiple years) training data or the quantitative comparison with other similar researches. Nevertheless, the following observations can be drawn from this study:

- Indeed, it provides more efficient estimators and can also be carried out when distributions of residuals are not normal and heteroscedastic.
- The weakness of the Johansen approach is that it is sensitive to the lag length. This last one has been determined in a systematic and accurate manner to make the technique perfectly reliable.
- The Johansen cointegration relationship is determined from data of a specific year, and the outcome has been applied and compared to other years of data under real conditions. Like all the current statistical methods of solar photovoltaic generation forecast, it uses past meteorological parameters, hourly irradiance and hourly PV power output to reconstruct the relationship, which means that it is severely dependent of in-site data and requires a sufficient of past measures to be more precise.
- This multiple linear regression between PV power output and the four chosen environmental parameters has a prediction accuracy for the following years between 65%

and 74% (Figures 4.9–4.11). The precision is not as high as we expected but can be explained by the fact that this model is a regression model which as we know is better suited for short-term or medium-term forecast and not long term forecast. However, the performance accuracy is higher since we deal with short term forecast (Figure 4.13). Its performance is hardly comparable as it is to Machine Learning (ML) or deep learning models such as Artificial Neural Networks (ANN) or Support Vector Machine (SVM) which are widely use nowadays but by building a hybrid model combining the Johansen cointegration principle and a Machine Learning technique the model's performance can be widely improved.

- In this research, the four environment factors namely solar irradiation, module temperature, wind speed and relative humidity were chosen because their data were available from various sensors on site. Solar irradiation and cell temperature are the two most sensitive factors in the PV power generation.

The main goal is this paper was to propose an original statistical approach that can estimate and forecast PV generation based on meteorological parameters in a tropical island such as Reunion Island.

4.9 Conclusion and perspectives

Johansen cointegration principle has been applied to non-stationary economic variables for cointegration analysis of equilibrium relationships, but has never been applied to renewable energy domain. The determined model is free from serial correlation or heteroscedasticity and it can then be used for forecasting. The outcomes show that the Johansen test is an appropriate applied model able to build a cointegration relationship between PV output generation and meteorological parameters such as solar radiation, module temperature, wind speed and humidity. This promising model is only at the beginning of a new facet of research with multidisciplinary competence in this field.

In future research works, the cointegration equation determined in this paper requires improvement by additional robust statistical methods and more robust residual tests, including additional environmental parameters such as ambient temperature, dust, as well as physical effects of air convention, heat transfer by conduction and radiation to PV technology. The resulting cointegration equation should then be applied to a residential area where the consumption profile of residents is known, in order to integrate other back up energy systems such as wind turbines, fuel cells, biomass to move towards smart buildings.

A thorough benchmarking of statistical multiple linear and non-linear regression in order to forecast PV power generation will be proposed in a future work for the sake of comparing this proposed method with several statistical regression forecast models, according to some objective criterion. In order to characterize the quality of the forecasts of each of these models, commonly used error metrics such as: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Bias Error (MBE) or their relative counterparts (rRMSE, RMAE, rMBE) will be applied.

The evaluation of the performance of the final model did not consider possible interactions among independent variables or even powered variables. In the next paper, these interactions will be discussed.

This regression model as it is does not possess evolutionary techniques such as heuristics and artificial intelligence (neural networks, etc.). A future work will tackle this problem by building a hybrid model combining the Johansen cointegration principle and a Machine Learning technique and therefore will be able to consider more accurately the climate variability and climate change effect.

From the perspective side in a near future, the mathematical aspects behind the statistical theories will be computed in line code using Python 3.7 and integrated on an FPGA chip in order to be applied at minute sampling time to make accurate daily prediction. The whole process should be identified as the RTF (Ramenah-Tanougast-Fanchette) respectively for physical aspects, statistical technique, FPGA implementation and artificial intelligence for predictive principle of energy systems to smart building and smart city. The international University of Mascareignes of Republic of Mauritius in the Indian Ocean has been approached in order to perform the final test.

Conclusion

Ce chapitre est extrait de l'article publié dans la revue internationale AIMS Energy (Fanchette et al., 2020). Il a pu mettre en évidence l'approche de la cointégration par le modèle à correction d'erreurs vectoriel (MCEV) de Johansen pour proposer un modèle de conversion de la production d'énergie photovoltaïque à l'île de la Réunion à partir d'un échantillon de données de moyennes journalières d'un an est extrait de sept années de mesures du bâtiment COREX et à ces données sont appliquées la méthode MCEV de Johansen. Les résultats se sont montrés satisfaisants avec un R^2 situé entre 65% et 74%.

Cependant, plusieurs limites ont été mises en lumière. Le modèle est un modèle de conversion et non pas un modèle de prévision à proprement parler de production PV. Par ailleurs, ce modèle est peu adapté pour un jeu de données climatiques conséquent comme celui obtenu en sortie du modèle WRF, d'où la nécessité d'un algorithme prédictif plus robuste. Le choix s'est porté sur un type de réseau de neurones récurrent particulier : le LSTM, présenté dans le chapitre 3.

La prochaine étape est la prévision de la production PV sur l'ensemble du territoire réunionnais à partir des données de sortie climatiques WRF. Dans un premier temps, il faut convertir les données climatiques à notre disposition en données PV en utilisant le modèle à correction d'erreurs vectoriel (MCEV) de Johansen. Puis dans un deuxième temps, à partir du modèle prédictif photovoltaïque LSTM nous allons obtenir les données de prévision PV sur toute la carte de l'île de la Réunion sur les horizons temporels mensuels, journaliers et horaires. Enfin la dernière partie traitera d'une discussion sur les résultats obtenus.

Chapitre 5

PREVISION PHOTOVOLTAÏQUE SUR L'ILE DE LA REUNION

Introduction

Dans ce chapitre, la prévision de la production photovoltaïque est effectuée à différents horizons temporels. Cela a pour objectif de confirmer la pertinence de l'approche neuronale récurrente face à des jeux de données de taille conséquente et une topographie complexe spécifique à l'île de La Réunion.

Pour élaborer ce modèle de prévision de production PV basé sur un modèle neuronal LSTM discuté dans le chapitre 3, plusieurs étapes sont nécessaires.

La première étape consiste à créer une base de données de production PV à partir des données climatiques simulées par le modèle régional de climat WRF qui ont été validées dans le chapitre 2. Nous rappelons l'objectif de notre générateur de données WRF : générer de manière cohérente les données climatiques pour l'estimation et la prévision de la production PV sur l'ensemble du territoire réunionnais sur une grille de l'ordre du km² grâce à la méthode de cointégration de Johansen puis un modèle neuronal LSTM et Bi-LSTM. La deuxième partie est d'utiliser le prédicteur de production PV modélisé sur l'île à trois horizons temporels : M+1, j+1 et h+1.

Ce chapitre est organisé comme suit : premièrement sont présentées les données de production PV produites sur l'île de La Réunion à l'aide des données climatiques WRF et du modèle statistique de Johansen. La deuxième partie est consacrée au modèle de prévision et aux données de prévision de production PV à moyen, court et très court terme. Enfin, la dernière partie concerne l'analyse et une discussion sur notre modèle de prévision de production solaire PV aux différents horizons temporels face aux modèles « benchmark » que sont les modèles de persistance et SARIMA et des résultats.

5.1 Conversion de la production PV

5.1.1 Modèle de cointégration de Johansen et choix des données climatiques

Le but des chapitres précédents était de mettre en avant la capacité pour ce nouveau modèle statistique, le modèle de cointégration de Johansen, d'estimer la production photovoltaïque (variable expliquée) à partir les données climatiques pertinentes (variables explicatives) générées par un modèle régional climatique qu'est WRF sur tout le territoire réunionnais. À partir des données récoltées auprès de l'entreprise COREX Réunion et de sa centrale photovoltaïque située à la Possession, il a été possible dans le chapitre 4 de modéliser une méthode innovante de conversion de production PV à partir des variables météorologiques suivantes :

- Rayonnement solaire (GHI)
- Température à 2 mètres (T)
- Vitesse de vent (V)
- Humidité relative (RH)

Ces données ont été extraites des sorties WRF et utilisées dans un modèle de conversion linéaire en utilisant la méthode de cointégration de Johansen.

Ainsi, comme il a été vu dans le chapitre précédent, il en résulte une équation linéaire :

$$PV_t = 833.33 \times V_t + 3.74 \times \mathrm{GHI}_t + 36.38 \times T_t + 4758.96 \times RH_t \quad (85)$$

La relation de causalité de Granger a mis en avant le lien entre les variables explicatives climatiques et la variable à prévoir, à savoir la production PV. On a ainsi obtenu la série temporelle de la production PV sur l'île de la Réunion à haute résolution (~1km) pour les années 2017 et 2018 en utilisant les données de température, de la vitesse du vent, d'irradiance solaire et d'humidité relative. La section suivante met en avant les cartes de production solaire PV.

5.1.2 Cartes réunionnaises de production PV

De cette relation, des cartes mensuelles, journalières et horaires climatiques et de production PV sur les années 2017 et 2018 ont pu être faites. Ces cartes sont présentées pour janvier et juillet 2017 afin de représenter la variabilité saisonnière à La Réunion (janvier représente l'été austral et juillet, l'hiver austral). Ainsi, dans la figure 5.1 sont présentées les cartes du territoire réunionnais concernant le rayonnement solaire, la température, la vitesse du vent et la production PV en janvier correspondant l'été austral et en juillet correspondant à l'hiver austral sur l'île de La Réunion.

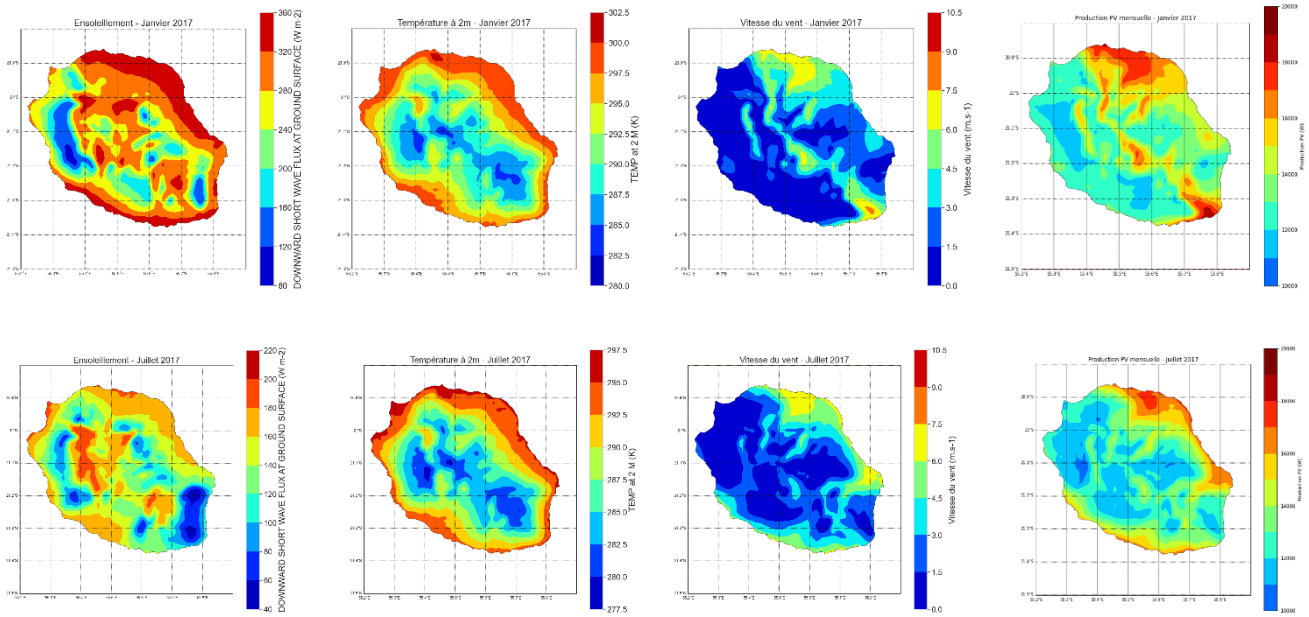


Figure 5.1 Cartes mensuelles du rayonnement solaire, de la température à 2 mètres, de la vitesse du vent et de la production PV en janvier 2017 (en haut) et en juillet 2017 (en bas)

En analysant la figure 5.1 et sans considérer la relation statistique, on constate une corrélation entre les cartes mensuelles climatiques WRF et la carte mensuelle de production PV. En effet, des zones parfaitement visibles de la carte notamment sur le littoral de l'île montrent des productions PV plus élevées, régions où la température et le rayonnement solaire sont justement les plus élevées aussi. De plus, une corrélation peut être faite entre vitesses de vent plus élevées et plus grande production PV, notamment sur le littoral nord, nord-est et sud.

Ce qui est à observer sur ces cartes notamment en comparant le mois de janvier 2017 à juillet 2017 (d'ailleurs les mêmes observations peuvent être faites avec les données de 2018) est la différence de rayonnement solaire global et de température sur le territoire réunionnais avec des valeurs plus élevées en janvier correspondant à l'été austral et une période globalement dans l'année plus ensoleillée et chaude par rapport à juillet (hiver austral). Cela se répercute de manière légitime sur la production PV sur l'île de La Réunion avec des valeurs de production PV dépassant les 19kW sur le nord et sud-est de l'île, là où les plus grandes valeurs de température, rayonnement solaire global et vitesse de vent sont notés.

Les cartes journalières et horaires sont présentées dans les sections suivantes. Dans le paragraphe suivant, vont être développés les modèles de prévision de production PV basés sur des réseaux de neurones, le LSTM et le Bi-LSTM.

5.2 Modèle de prévision de production

Le but de ce paragraphe est de créer un modèle LSTM (dit « vanilla » ou classique) et Bi- LSTM (cf. Chapitre 3.2.5) de prévision de production photovoltaïque qui permettra de prévoir les données PV sur l'ensemble du territoire selon trois horizons temporels choisis et dont la performance et la robustesse seront évaluées par des métriques de performance. Pour cela, ces modèles seront confrontés à deux autres modèles : le modèle de persistance et le modèle Seasonal Autoregressive Integrated Moving Average (SARIMA), une extension du modèle ARIMA (Autoregressive Integrated Moving Average) qui prend en compte la saisonnalité dans les données temporelles. Ces méthodes ont été décrites en détail dans les sections 1.3.1 et 1.3.3 du chapitre 1.

On peut classer en plusieurs catégories les horizons temporels de prévision de production PV. Trois horizons temporels de prévision ont toutefois été choisis pour être étudiés :

- Moyen - long terme (M+12) avec une résolution temporelle mensuelle
- Court terme (j+1, j+2,…) avec une résolution temporelle journalière
- Très court terme (h+1) avec une résolution temporelle horaire

Les données utilisées étant des donnés climatiques issues du modèle climatique régional WRF (cf. chapitre 2.4). Le chapitre 3 a permis de cerner les modèles que nous allions utiliser : le modèle de Johansen a explicité la méthodologie générale de conversion PV à partir de ces données (chapitre 4) et les modèles à apprentissage automatique LSTM et Bi-LSTM qui vont permettre leurs prévisions à trois horizons de prévision temporels : M+12, j+1 et h+1.

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

Ainsi, le premier paragraphe est dédié à l'étude du cas mensuel (horizon M+12). Cet horizon est certainement celui pour lequel la prédiction a plus d'impact pour de la planification et maintenance de système PV que pour un gestionnaire de réseau. Cependant, ce cas relativement simple (une seule périodicité, cumul annihilant le bruit des données, etc.) a permis de tester différentes approches et surtout différents outils de comparaison et d'évaluation de l'erreur.

Nous traiterons ensuite le cas de la prédiction journalière (horizon j+1), horizon temporel indispensable pour la planification opérationnelle des générateurs.

Puis sera décrit le cas horaire, mais sous un horizon 24 heures, 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. La méthodologie, bien que relativement proche du cas précédent, est plus complexe. Dans tous les cas de figures, une comparaison a été faite entre quatre modèles : le modèle de persévérance, le modèle SARIMA, le modèle LSTM et le Bi-LSTM.

Enfin, une discussion sera ouverte sur les résultats obtenus, leurs pertinences et les perspectives qui en découlent.

5.2.1 Les méthodes de prévision

Les deux modèles retenus vis-à-vis de nos jeux de données sont les modèles neuronaux LSTM et Bi-LSTM (cf. chapitre 2.2.4). Toutefois, il est primordial de vérifier la pertinence de ces modèles complexes par des modèles de référence. De ce fait, deux autres modèles ont été retenus :

- un modèle de type « naïf » : le modèle de persistance (Y. Zhang et al., 2020) (cf. chapitre 1.3.1).
- un modèle de référence de par le nombre d'études l'ayant utilisé, issus de la famille des modèles autorégressifs à moyenne mobile saisonniers : le modèle SARIMA (Kushwaha & Pindoriya, 2017) (cf. 1.3.3).

Les modèles de prévision ont été exécutés sur un ordinateur doté de 96 Gb de RAM, un processeur Intel Core i7-6700K et un CPU de 8 × 4GHz, avec un logiciel de rendu graphique open source llvmpipe LLVM 15.0.6 et une capacité de disque de 5,5 Tb et sont codés sous le langage de programmation Python.

5.2.2 Méthodologie

Pour établir la prévision de production PV issue des modèles de persistance et SARIMA, ces derniers ont été entrainés sur une base d'entrainement ou d'apprentissage qui couvre les

données de production PV sur l'année 2017 et les prévisions sont faites sur l'année 2018 et sont évaluées par différentes mesures de performances (MAE, RMSE, nRMSE, r, par rapport aux données de la référence. De même pour la prévision sur l'année 2019, les modèles sont entrainés sur les données de production 2018 avant d'établir la prévision en 2019. Toutefois, nous ne possédons pas suffisamment de données de production PV en 2019 pour évaluer et valider les méthodes sur cette année.

Étant donné qu'il est impossible d'utiliser des données nouvelles et non aperçues pour effectuer une comparaison, les données disponibles sont divisées en des sous-ensembles d'entrainement et de test comme le montre la figure 5.2. Le modèle est construit et entrainé à partir des données du sous-ensemble d'entrainement (2017). L'intérêt du sous-ensemble de test est de voir les performances du modèle sur un autre jeu de données connues et de même nature (2018). La séparation des données en ces sous-ensembles dépend de différents facteurs comme le type de données, le but de l'analyse et la taille des données. Donc, il a été décidé de conserver toute une année pour l'apprentissage et une autre pour les tests. La plupart du temps, il y a le besoin d'améliorer le modèle en réglant les hyperparamètres (comme le nombre de couches cachées, le nombre de neurones actifs par couche, le taux d'apprentissage ou encore la fonction d'activation). Ce réglage est explicité et se fait manuellement dans chaque section traitant des prévisions à chaque horizon temporel.

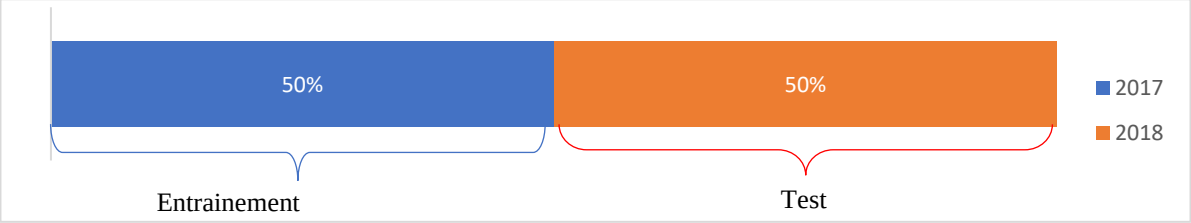


Figure 5.2 Répartition des données entre sous-ensemble d'entrainement et de test pour les modèles neuronaux LSTM et Bi-LSTM

5.2.3 Le cas mensuel : M+1

La planification de la capacité et des opérations comprend les décisions d'investissement et de planification à long terme. La planification de la capacité, ou du système, consiste à évaluer les besoins et à investir dans de nouvelles infrastructures de production, de transport et de distribution sur une période de plusieurs années. La planification des opérations implique la programmation des ressources disponibles pour répondre à la demande saisonnière prévue et s'étend sur une période de plusieurs mois, d'où l'importance d'une prévision sur les prochains mois.

La série temporelle utilisée pour la prévision PV mensuelle n'est autre que les données de production PV à haute résolution sur les années 2017 et 2018 obtenues par le modèle de conversion de cointégration de Johansen (cf. Chapitre 5.1). Le but est de faire de la prévision de production PV sur l'année 2019 à partir de la série spatio-temporelle de production PV de 2017 et 2018. Toutefois, ces données sont moyennées aux mois. Nous avons ainsi un jeu de données contenant l'équivalent d'une carte mensuelle avec 100 points en longitude et en

latitude (représentant un carré recouvrant l'intégralité de l'île), d'où une matrice de données de production PV de 1×100×100 pour un mois et donc de 12×100×100 pour une année.

Le but des manipulations initiées dans cette section est de sélectionner les modèles les plus performants pour prédire la production PV pour le ou les mois d'après. Les modèles qui seront évaluées sont la persistance, le SARIMA, le LSTM et le Bi-LSTM. Chaque modèle sera optimisé de manière manuelle (l'optimisation automatique des hyperparamètres des modèles neuronaux sera évoquée dans les Perspectives) afin de retenir les meilleurs paramètres possibles pour cette étude. Après optimisation manuelle, les modèles retenus sont :

- la persistance qui ne nécessite pas d'optimisation
- SARIMA avec p = 5 et q = 0 (cf. chapitre 1.3.3)
- LSTM (cf. chapitre 2.2.3) avec 10 neurones sur la couche d'entrée, 10 nœuds cachés, une fonction d'activation de tangente hyperbolique « *tanh* », 100 époques avec une optimisation de type « *Adam* », une fonction de perte d'erreur moyenne quadratique « *rmse* » et une évaluation du modèle durant l'apprentissage et le test par une métrique de performance de type « *mape* » qui calcule le pourcentage d'erreur absolue moyen. Tous les autres paramètres sont pris par défaut (Toolbox Tensorflow keras sur Python) (Team Keras)
- Bi-LSTM (cf. chapitre 2.2.4) utilise les mêmes paramètres que le modèle LSTM précédent.

Il faut préciser concernant le modèle LSTM et Bi-LSTM les hyperparamètres suivants :

- Le nombre de neurones sur la couche d'entrée définit le nombre de variables d'entrée que le modèle prend en compte pour la prédiction. Dans ce cas, il y a 10 neurones sur la couche d'entrée, ce qui signifie qu'il y a 10 variables d'entrée.
- Le nombre de nœuds cachés définit le nombre de neurones dans la couche cachée du réseau de neurones. Dans ce cas, il y a 10 nœuds cachés.
- La fonction d'activation est utilisée pour introduire de la non-linéarité dans le modèle et pour transformer la sortie des neurones. Dans ce cas, la fonction d'activation utilisée est la tangente hyperbolique, qui comprime les valeurs en sortie entre -1 et 1.
- La fonction de perte est utilisée pour mesurer l'erreur entre la sortie réelle et la sortie prédite du modèle. Dans ce cas, la fonction de perte utilisée est l'erreur moyenne quadratique, qui mesure la racine carrée de la moyenne des carrés des différences entre les valeurs prédites et les valeurs réelles.
- Le nombre d'époques (*epoch*) définit le nombre de fois que le modèle va être entraîné sur l'ensemble de données. Dans ce cas, le modèle est entraîné sur 100 époques.
- La métrique de performance est utilisée pour mesurer la qualité de la prédiction du modèle. Dans ce cas, la métrique de performance utilisée est le pourcentage d'erreur absolue moyen, qui mesure le pourcentage moyen de l'erreur absolue entre les valeurs prédites et les valeurs réelles.
- Enfin, l'optimisation est utilisée pour ajuster les poids du réseau de neurones en fonction de l'erreur entre la sortie réelle et la sortie prédite. Dans ce cas, l'optimisation utilisée est Adam. Ce choix sera détaillé dans le paragraphe suivant.

L'optimisation Adam (Adaptive Moment Estimation) est une méthode de descente de gradient stochastique qui repose sur l'estimation adaptative des moments de premier et de second ordre. (Kingma & Ba, 2017) considèrent que la méthode est "efficace sur le plan informatique, peu exigeante en termes de mémoire, invariante à la remise à l'échelle diagonale des gradients, et bien adaptée aux problèmes de grande taille en termes de données/paramètres". Cette optimisation a été choisie, après plusieurs tests de performance sur le modèle LSTM, parmi plusieurs optimisations. Ainsi, il a été comparé à l'optimisation par l'algorithme RMSprop (Root Mean Squared prop). (Hinton et al., 2012) calculent la moyenne exponentiellement pondérée des carrés des gradients pour chaque paramètre du modèle. Cette moyenne est utilisée pour normaliser le gradient lors de la mise à jour des poids. RMSprop est donc une variante de la descente de gradient stochastique (SGD) et l'optimisation par algorithme NAdam (Nesterov-accelerated Adaptive Moment Estimation) qui est une extension de l'algorithme Adam en utilisant une version modifiée de la dynamique de Nesterov (Dozat, 2015).

En comparant, les trois optimiseurs, le choix s'est porté sur l'optimiseur Adam, car les performances des prévisions étaient meilleures avec un nRMSE plus faible et une plus forte corrélation sur 10 tests effectués, le nombre de tests ayant été choisi arbitrairement. Ces résultats sont résumés dans le tableau 5.1. De ce fait, à partir de là toutes les simulations LSTM utilisent l'optimiseur Adam.

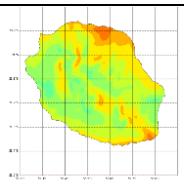
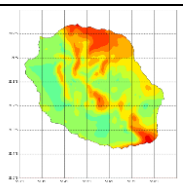
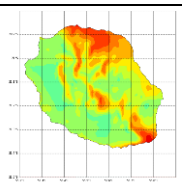
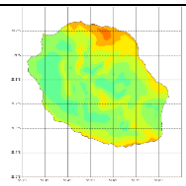
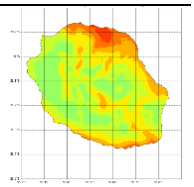
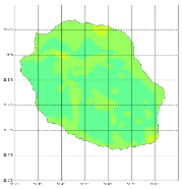
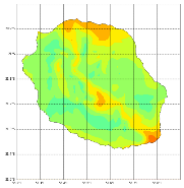
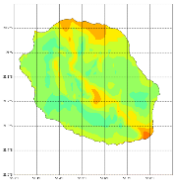
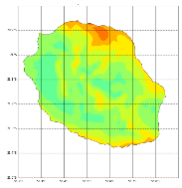
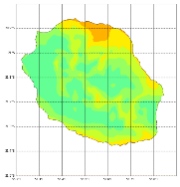
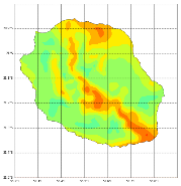
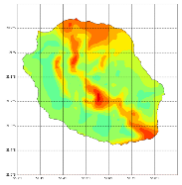
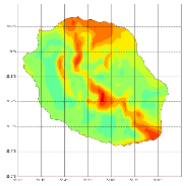
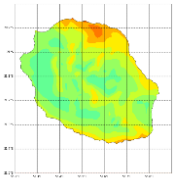
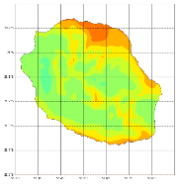
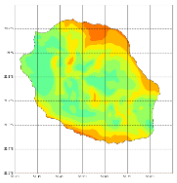
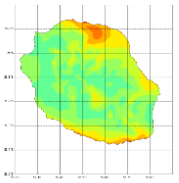
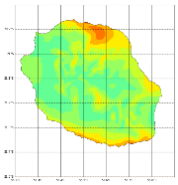
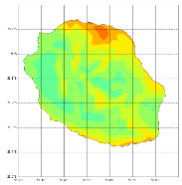
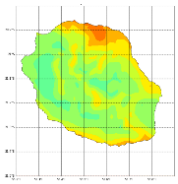
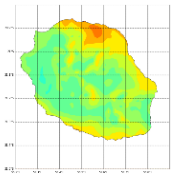
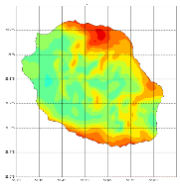
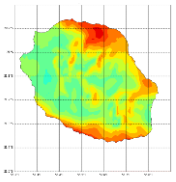
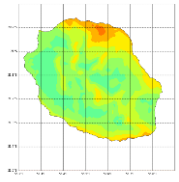
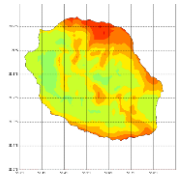
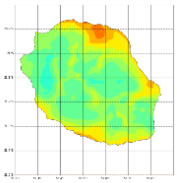
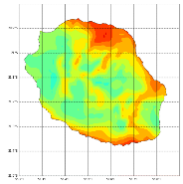
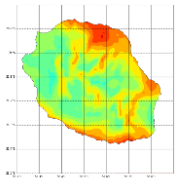
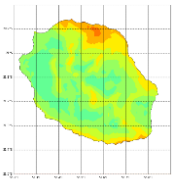
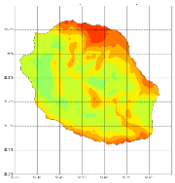
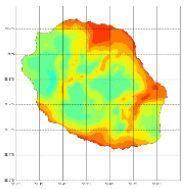
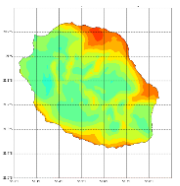
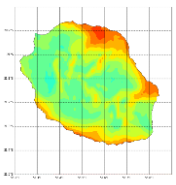
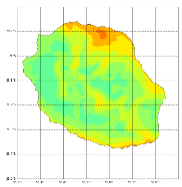
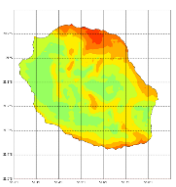
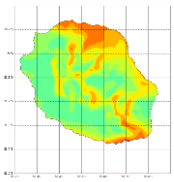
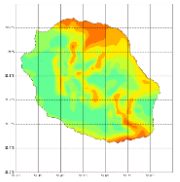
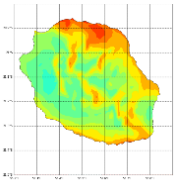
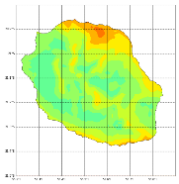
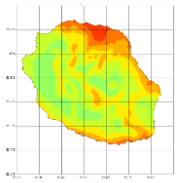
Tableau 5.1 Résultats de la comparaison entre les différents optimiseurs sur des 10 simulations de prévisions de production PV utilisant le modèle LSTM. Sont inclus l'erreur quadratique moyenne normalisée en pourcentage et le coefficient de Pearson

	Adam	**RMSprop**	**Nadam**
nRMSE (%)	8.90 ± 0.1	15.40 ± 0.15	14.84 ± 0.18
r	0.79 ± 0.01	0.78 ± 0.02	0.50 ± 0.02

Pour tous les modèles étudiés, vu que nous n'avions que deux années de données, les 12 mois de l'année 2017 ont servi de sous-ensemble d'apprentissage et les 12 mois de l'année 2018 concerne le test de prédiction. L'année 2018 de production PV sera par la suite considérée comme la référence pour toutes les sorties de prédictions des différentes méthodes. Une prédiction sur l'année 2019 est aussi faite, mais il nous est impossible de vérifier la performance de cette prédiction car nous n'avons pas assez de données énergétiques au sol pour pouvoir valider ces prévisions. Cette méthodologie sera appliquée durant tout le chapitre 5 et pour tous les horizons temporels de prévision.

5.2.3.1 Cartes de prévisions de production PV mensuelles en 2018

La figure 5.3 représente les cartes de prévision de production PV pour les quatre modèles étudiés pour les 12 mois de 2018 ainsi que la référence en termes de production PV sur l'ensemble du territoire réunionnais.

Mois (2018)	Référence	Persistance	SARIMA	LSTM	Bi-LSTM
Janvier					
Février					
Mars					
Avril					
Mai					
Juin					
Juillet					
Août					
Septembre					

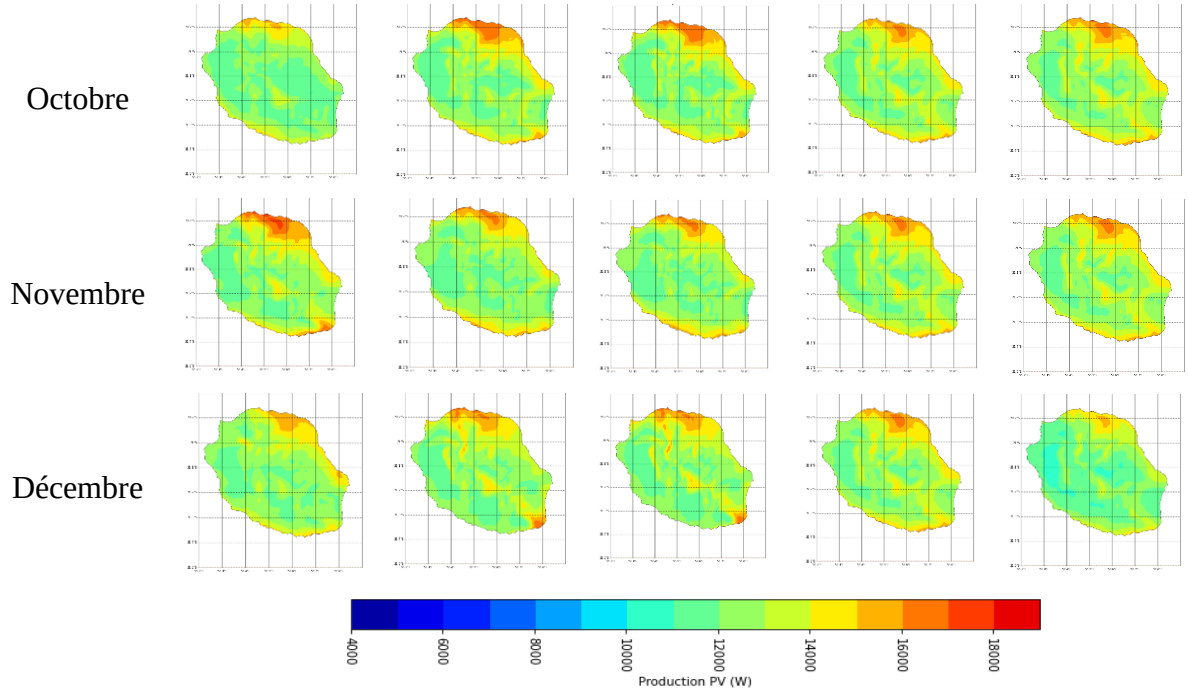


Figure 5.3 Carte de production PV de référence (à gauche) face aux cartes de prévision de production PV mensuelle sur les 12 mois de 2018 pour les modèles de persistance, SARIMA, LSTM et Bi-LSTM (de gauche à droite). L'échelle des couleurs de la production PV, comprise entre 4000 W et 19000 W est située en bas de la figure

Tableau 5.2: Résultats de la comparaison entre les différentes méthodes de prévision de production PV mensuelles basées sur des données mensuelles. Sont inclus l'erreur moyenne absolue, l'erreur quadratique moyenne, l'erreur quadratique moyenne normalisée, l'écart-type, le coefficient de Pearson et le temps de calcul pour chacune des simulations.

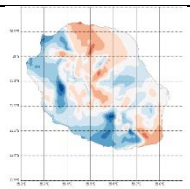
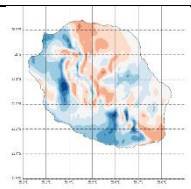
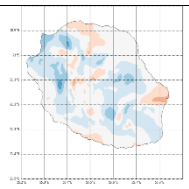
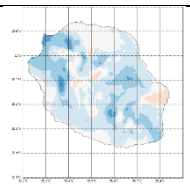
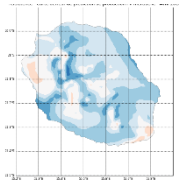
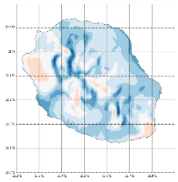
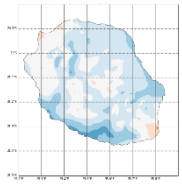
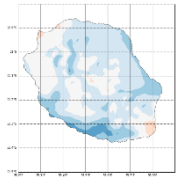
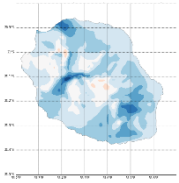
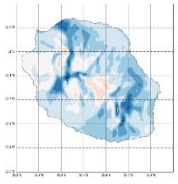
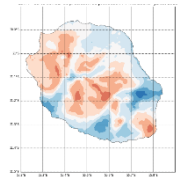
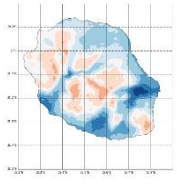
Métriques	Persistance	SARIMA	LSTM	Bi-LSTM
MAE (W)	1160	1174	**1065**	1103
RMSE (W)	1410	1423	**1313**	1319
nRMSE (%)	9.01	9.09	**8.35**	8.46
StD (W)	2223	**2205**	1824	1779
r	0.81	0.80	**0.82**	0.81
Temps de calcul (s)	**99**	260	444	1901

Le tableau 5.2 résume les résultats de la comparaison des quatre méthodes de prévision de production PV mensuelle. Le faible jeu de données en entrée avait formulé l'hypothèse que les modèles simples comme le modèle de persistance donneraient probablement les meilleurs résultats. Toutefois, nous pouvons constater en gras que le modèle LSTM classique affiche de manière générale des performances meilleures que les autres modèles, que ce soit pour un *MAE* = 1065 *W* soit une erreur de 7% par rapport à la carte de référence, un *RMSE* = 1313 *W*, le *nRMSE* =8,35% et *r* = 0.82. Il performe mieux que le modèle Bi-LSTM qui affiche des résultats satisfaisants mais pour un grand temps et puissance de calcul. Néanmoins, comme nous pouvions nous y attendre, le modèle de persistance affiche un temps de calcul plus faible (99s de temps de calcul soit 19 fois plus rapide que le modèle Bi-LSTM (1901s)). Notons que le

modèle SARIMA a un écart-type StD = 2205 W proche de celui de la référence (StD_{ref} = 2204 W). Le fait que ces deux valeurs soient proches suggère que le modèle SARIMA est capable de reproduire la variabilité de la référence de manière précise, ce qui peut indiquer une bonne performance du modèle. Il est toutefois important de souligner que l'écart-type ne représente qu'une mesure de la dispersion des données autour de la moyenne, et c'est la raison pour laquelle les autres mesures d'évaluation sont également prises en compte pour évaluer la performance globale d'un modèle.

Il faut noter que les métriques telles que celles présentées dans le tableau 5.2 sont importantes car elles permettent d'évaluer globalement statistiquement chaque modèle, mais ne mettent pas en avant les erreurs d'un point de vue spatial. Il parait alors essentiel de présenter dans un deuxième temps les cartes d'erreurs de prévision PV mensuelle. La figure 5.4, quant à elle, représente les cartes d'erreurs de prévision sur certains mois en particulier (janvier, avril, juillet et octobre) c'est-à-dire l'erreur à chaque point de grille entre la carte de référence et la carte issue du modèle choisi. Ce type de carte est souvent utilisé pour évaluer la précision d'un modèle de prévision géospatial en visualisant les différences entre les valeurs « réelles » (ici celles de la carte de référence) et les prévisions du modèle de prévision en question à travers une zone géographique. Ces cartes permettent d'identifier les zones où le modèle est performant c'est-à-dire où les erreurs sont faibles voire nulles. Elles permettent aussi d'évaluer la précision du modèle globalement en fournissant une vue d'ensemble de la performance des différentes méthodes.

5.2.3.2 Cartes d'erreurs de prévision de production PV mensuelles en 2018

Mois (2018)	Persistance	SARIMA	LSTM	Bi-LSTM
Janvier				
Avril				
Juillet				

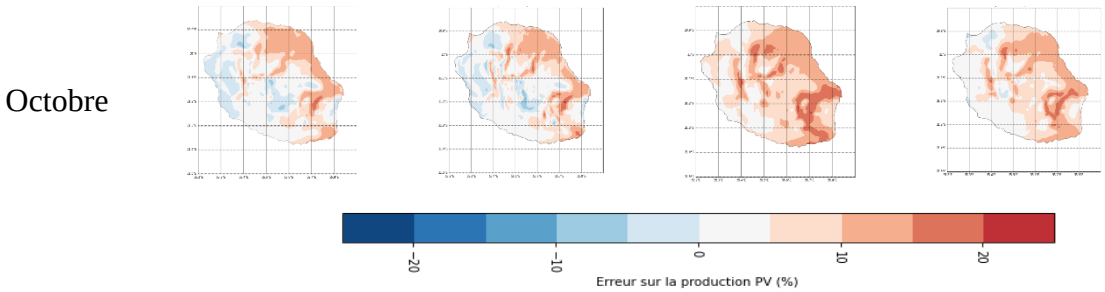
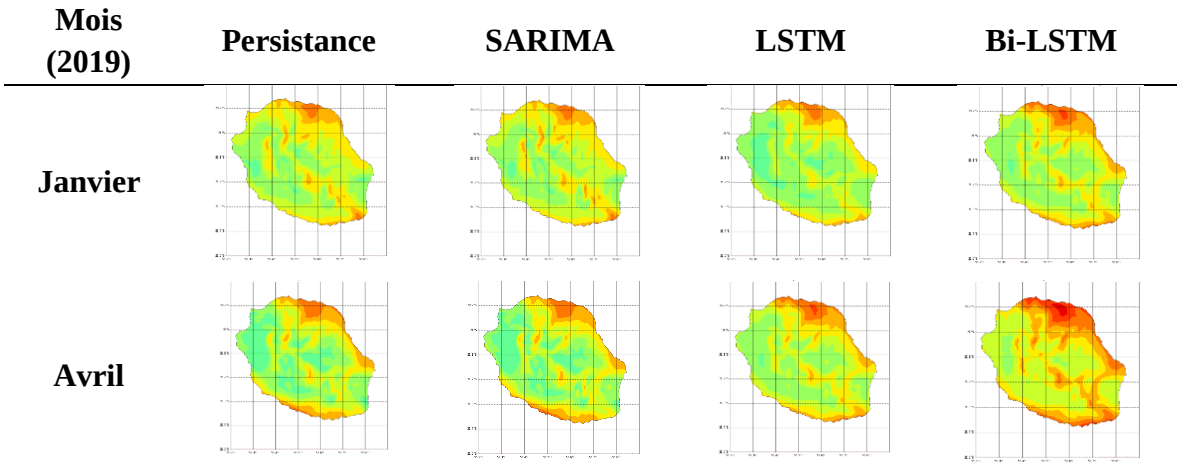


Figure 5.4 Cartes mensuelles d'erreur de production PV sur les mois de janvier, avril, juillet et octobre de 2018 pour les modèles de persistance, SARIMA, LSTM et Bi-LSTM (de gauche à droite). L'échelle des couleurs de l'erreur de production PV, compris entre -20% et 20%, est située en bas de la figure.

En analysant plus en détail la performance des modèles, le tableau de performance, les cartes de prévision mensuelles en 2018 et les cartes d'erreurs, nous voyons l'intérêt de présenter les cartes produites selon les 4 méthodes présentées face aux cartes de référence. Les cartes représentent visuellement les erreurs spatialement mais aussi temporellement, selon les mois de l'année. Les erreurs pour les 4 modèles se révèlent être comprises, peu importe le point de grille sur la carte, comprise entre -20% et 20%. Le modèle LSTM respecte le plus la spatialité de la production PV, sauf comme nous le constatons sur les mois de juillet et octobre où la production est sous-estimée notamment dans les hauts de l'île au centre et dans les cirques. Nous pouvons en conclure que le faible volume de données mensuelles d'entrainement ne permet pas au modèle pour cet horizon temporel de reproduire la production à tous les mois de l'année et à toute altitude. Cette remarque peut être d'ailleurs faite pour le modèle Bi-LSTM qui connait les mêmes limites aussi dues au faible jeu de données et à la faible optimisation des algorithmes.

5.2.3.3 Cartes de prévision de production PV mensuelles 2019

La figure 5.5 représente les cartes de prévision de production PV en 2019 selon les 4 différents modèles. Ces cartes sont obtenues en entrainant les modèles sur les données de 2018.



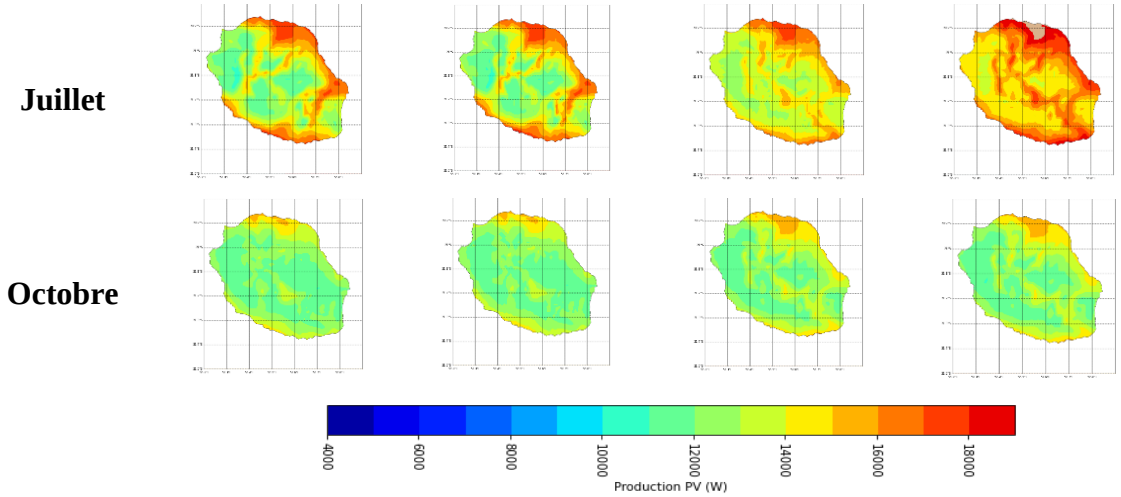


Figure 5.5 Cartes de prévision de production PV mensuelle sur les mois de janvier, avril, juillet et octobre de 2019 pour les modèles de persistance, SARIMA, LSTM et Bi-LSTM (de gauche à droite). L'échelle des couleurs de la production PV, comprise entre 4 kW et 19 kW est située en bas de la figure

5.2.3.4 Variables exogènes

Nous avons vu dans le paragraphe précédent que les prévisions de production PV mensuelles réalisées en utilisant uniquement, en entrée des modèles, des données de nature identique à la nature de la sortie, ne permettent pas de faire mieux que la méthodologie Persistance et SARIMA. Nous allons essayer d'établir si l'utilisation d'entrées de natures exogènes modifie cette conclusion, et améliore ainsi la prédiction de la production 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 le rayonnement solaire, l'humidité, la vitesse du vent, la température, les précipitations ou la nébulosité ont une incidence directe sur la production PV. Ces variables dénommées « exogènes » peuvent être intégrées au vecteur des données d'entrée du modèle.

La méthode consiste à utiliser des variables météorologiques en plus des entrées endogènes. Les quatre variables « exogènes » définies dans la section 5.1.1 découlent de l'équation utilisée pour déterminer la production PV, notre variable endogène. Ainsi, ces variables sont le rayonnement solaire (GHI), la température au sol (T), la vitesse de vent (V) et l'humidité relative (RH). L'idée est ainsi d'avoir en entrée des deux modèles neuronaux, LSTM et Bi-LSTM, une variable endogène : la production PV (PV) et quatre variables exogènes : GHI, T, V et RH afin de constater s'il y a une amélioration de la performance de ces modèles.

Deuxième paramètre que nous voulons vérifier l'influence de la qualité et de la quantité de données en entrée du modèle. Pour résumer, ici, nous voulons voir si le choix du nombre de mois d'apprentissage a une forte influence sur la performance du modèle. Donc, nous avons testé trois configurations pour les 4 modèles. Pour chaque méthode, l'apprentissage a été fait sur un seul mois et prévision d'un mois et en n'utilisant que les données de production PV pour faire la prévision. Par exemple : prévoir la production PV du mois de janvier 2018 avec

seulement la production PV du mois de janvier 2017 comme base d'apprentissage. C'est le modèle « *one month* ». La deuxième configuration est plus classique, nous utilisons les 12 mois de production PV de 2017 pour prévoir la production PV en 2018. C'est le modèle « *all month* ». Enfin, pour la dernière configuration, nous utilisons les 12 mois de production PV et des variables climatiques en 2017 en plus des données de production PV historiques pour prévoir la production PV en 2018. C'est le modèle « *all variables* » qui est valable seulement pour les modèles neuronaux. Les modèles plus simples comme le modèle de persistance et SARIMA n'utilisent qu'une variable en entrée pour faire de la prévision.

Les résultats des performances de ces 4 modèles sont résumés dans le diagramme de Taylor (figure 5.6) (Taylor, 2001) en n'utilisant qu'un mois de données pour entrainer les modèles (one_month), les douze mois (all_month) et enfin pour les deux modèles neuronaux l'utilisation en entrée de données exogènes en plus des données de production PV (all_variables).

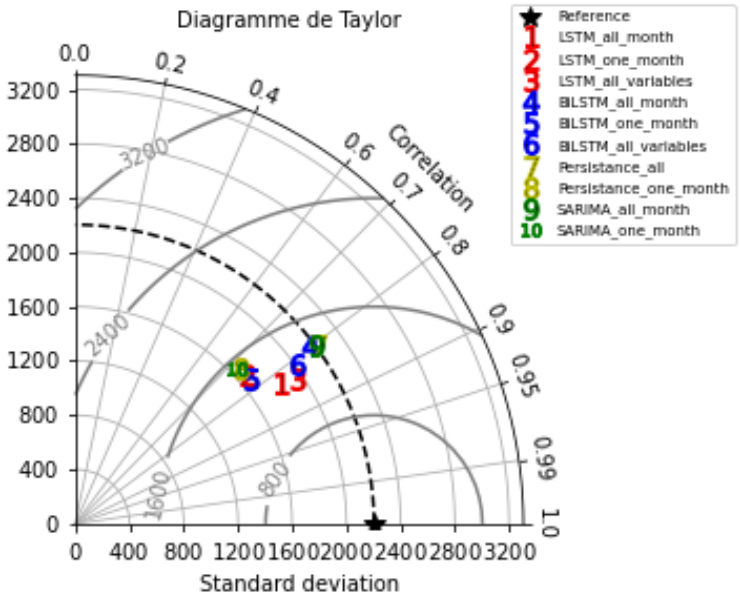


Figure 5.6 Diagramme de Taylor – Performance du modèle de persistance, SARIMA, LSTM et Bi-LSTM pour les prévisions de production PV mensuelles avec la configuration « one_month », « all_month » et « all_variables »

Ce que nous pouvons conclure de ce diagramme est que le modèle LSTM parait d'ores et déjà d'un point de vue performance meilleure que les trois autres. De manière générale pour les 4 modèles, utiliser les douze mois à notre disposition comme données d'entrée améliorent considérablement les performances et l'utilisation des données exogènes met en avant une meilleure efficacité tout en gardant à l'esprit que cela rallonge les temps de calcul et que la puissance de calcul nécessaire aux simulations est plus importante. L'apprentissage sur une plus faible base de données réduit la performance des modèles, mais permet d'obtenir des simulations plus rapides. Toutefois, pour le cas journalier et horaire des prochaines sections, les modèles neuronaux utiliseront les données des variables exogènes pour les prévisions.

5.2.4 Le cas journalier : j+1

La programmation des opérations désigne le processus consistant à déterminer quels générateurs fonctionnent pour répondre à la demande prévue à court terme. Il s'agit généralement de prendre des engagements à l'avance sur la base des prévisions de la demande du jour suivant, avec des ajustements effectués dans une période allant de quelques heures à 15 minutes pour tenir compte des écarts entre les prévisions de la demande à l'avance et celles du jour suivant, ainsi que pour tenir compte des pannes imprévues des centrales ou des problèmes des lignes de transport.

Comme dans la section 5.2.2, la série temporelle utilisée pour la prévision PV journalière n'est autre que les données de production PV à haute résolution sur les années 2017 obtenues par le modèle de conversion de cointégration de Johansen (cf. Chapitre 5.1). Le but est de faire de la prévision de production PV sur l'année 2018 à partir de la série spatio-temporelle de production PV de 2017. Toutefois, ces données sont moyennées sur une journée. Nous avons ainsi un jeu de données contenant l'équivalent d'une carte journalière avec 100 points en longitude et en latitude comme précédemment avec une matrice de données de production PV de $1 \times 100 \times 100$ pour une journée soit $365 \times 100 \times 100$ pour une année.

Le but des manipulations initiées dans cette section est de sélectionner les modèles les plus performants pour prédire la production PV pour le ou les jours d'après. Les modèles qui seront évaluées sont la persistance, le SARIMA, le LSTM et le Bi-LSTM. Comme la section 5.2.2, chaque modèle a été optimisé manuellement afin de retenir les meilleurs pour cette étude. Après optimisation, les modèles retenus sont les mêmes que précédemment :

- la persistance qui ne nécessite pas d'optimisation
- SARIMA avec p = 5 et q = 0
- LSTM avec 10 neurones sur la couche d'entrée, 10 nœuds cachés, une fonction d'activation de tangente hyperbolique « *tanh* », 100 époques avec une optimisation de type « *adam* », une fonction de perte d'erreur moyenne quadratique « *rmse* » et une évaluation du modèle durant l'apprentissage et le test par une métrique de performance de type « *mape* »
- Bi-LSTM utilise les mêmes paramètres que le modèle LSTM.

Nous conserverons l'optimisation « adam » durant tout ce chapitre 5. Pour tous les modèles étudiés, les 365 jours de l'année 2017 ont servi de base d'apprentissage et les données de l'année 2018 concernent le test de prédiction. Les modèles ont été entrainés sur les différents jours d'une même classe avant d'être testé sur les jours appartenant à la même classe journalière en 2018. L'année 2018 de production PV est considérée comme la référence pour toutes les sorties de prédictions des différentes méthodes.

5.2.4.1 Classification ascendante hiérarchique

Toutefois, il ne serait pas judicieux et pertinent de simuler des prévisions journalières sur des jours pris au hasard. Au lieu de considérer des séries temporelles complètes constituées de 365 jours pour chaque année d'intérêt, nous adoptons une approche plus compacte dérivée

d'une technique de classification objective qui regroupe les champs de productions PV quotidiennes les plus similaires en un nombre limité de classes statistiquement cohérentes.

Ainsi pour documenter la variabilité annuelle sans produire une série de 365 cartes quotidiennes, les jours les plus ressemblants sont objectivement regroupés à l'aide d'une technique de regroupement hiérarchique (Cretat et al., 2011), formant ainsi des classes statistiquement homogènes. La technique ascendante signifie qu'au début de l'algorithme, chaque jour constitue une classe distincte. Puis, de manière itérative, l'algorithme fusionne deux par deux toutes les classes (c'est-à-dire les jours ou groupes de jours), en identifiant les deux plus similaires. L'algorithme peut être arrêté à n'importe quelle étape, avant que les classes ne soient finalement fusionnées pour former une seule classe. Cette technique de clustering non paramétrique, pour laquelle il n'existe pas de critère pour le choix du nombre de clusters, présente également l'avantage par rapport aux autres techniques basées sur les k-means d'être applicable à des séries non gaussiennes comme le rayonnement solaire ou la production PV quotidienne. Les k-means est un algorithme de clustering qui fonctionne en divisant un ensemble de données en un nombre prédéfinis de groupes (ou clusters) selon une distance mesurée entre les observations. (Cheng & Wallace, 1993), (Casola & Wallace, 2007), ou plus récemment (Govender, 2017) ont démontré l'utilité de cette technique pour l'identification de régimes météorologiques. La classification hiérarchique est basée sur la matrice de distance ψ, calculée comme la distance euclidienne entre chaque observation (jour) et toutes les autres. Chaque jour est défini par N variables (c'est-à-dire la production PV). La méthode de « clustering » de Ward (Ward, 1963) est ensuite utilisée pour regrouper les jours en minimisant l'inertie intra-classe (c'est-à-dire l'hétérogénéité entre les modèles d'une classe donnée). Les événements individuels les plus similaires, notés A et B, sont d'abord identifiés en utilisant la distance minimale dans ψ, et regroupés pour former un premier cluster R. La distance d(R, Q) entre R et tout autre cluster Q (événement individuel ou groupe d'événements), constitué par m_R et m_Q événements, respectivement, est une fonction linéaire des distances de Q par rapport aux clusters originaux A et B fusionnés en R. Elle est définie comme suit :

$$d(R,Q) = \frac{m_A + m_Q}{m_A + m_B + m_Q} d(A,Q) + \frac{m_B + m_Q}{m_A + m_B + m_Q} d(B,Q) - \frac{m_Q}{m_A + m_B + m_Q} d(A,B) \quad (85)$$

où m_A et m_B sont les nombres d'événements de A et B. Les distances d(A, Q), d (B, Q) et d(A, B) sont obtenues directement à partir de la matrice de distance ψ (ou, à d'autres étapes de la procédure de classification, à partir des distances calculées aux étapes précédentes).

La méthodologie a été appliquée séparément pour les deux années, mais nous présentons ici les résultats pour 2018 uniquement. L'arbre de classification ou dendrogramme (figure 5.7a) montre les regroupements successifs effectués par l'algorithme pour l'année 2018. Dans la figure 5.7b, le dendrogramme est tronqué en six classes, car cela apparaît comme un bon compromis entre compacité et hétérogénéité intra-clusters : comme il n'existe pas de test statistique pour déterminer le nombre optimal de classes dans cette méthodologie, ce choix dépend surtout du niveau de détail attendu du regroupement (Morel et al., 2014).

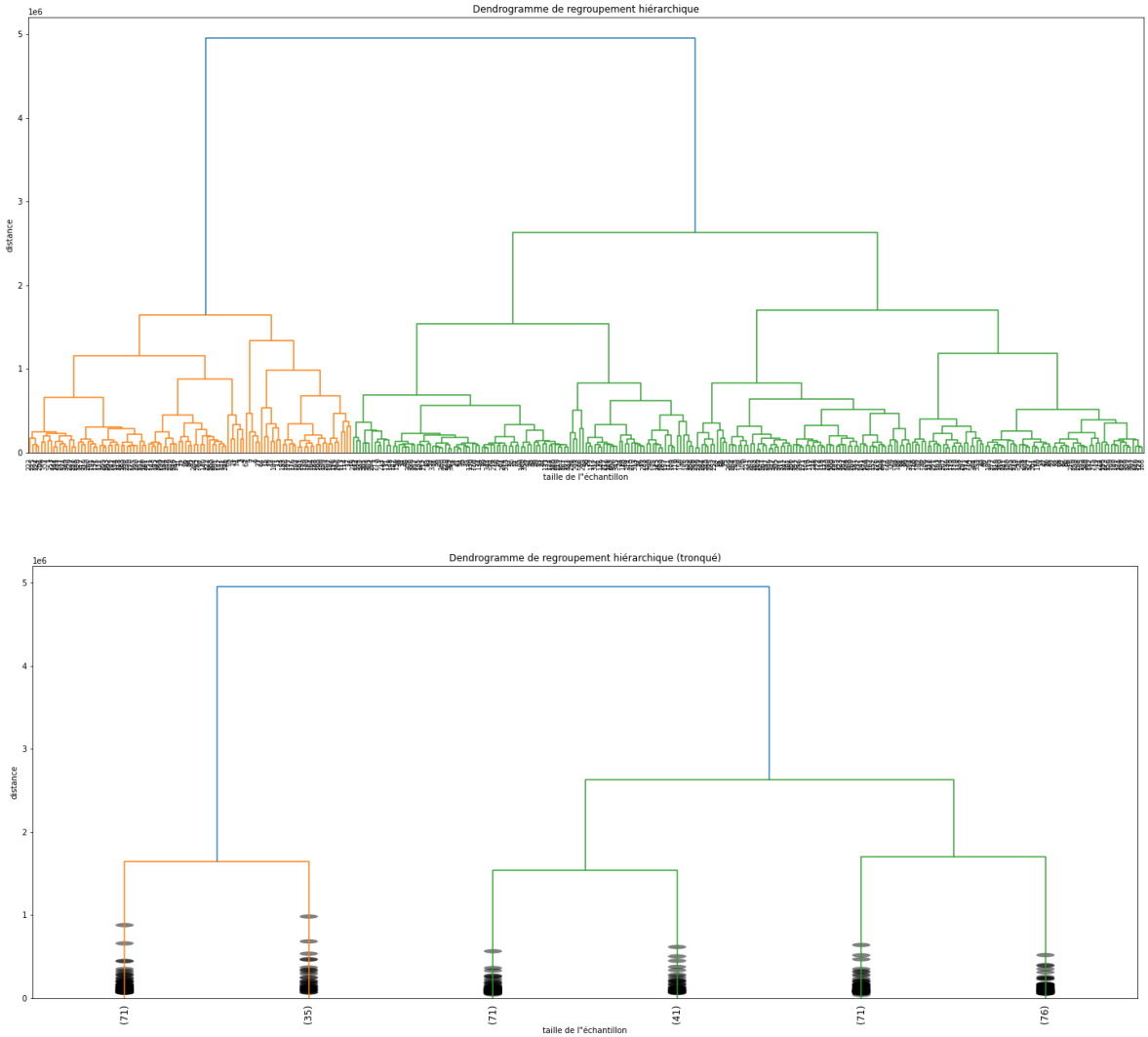


Figure 5.7 a. Dendrogramme (arbre de classification) de la classification ascendante hiérarchique effectuée sur les champs de production PV quotidiennes simulés pour 2018 avec en abscisse le numéro du jour
b. Dendrogramme de la figure 5.7a tronqué en 6 classes avec en abscisse le nombre de jours dans chaque classe

5.2.4.2 Profils journaliers

La figure 5.7 montre qu'à l'échelle de l'île entière, que nous pouvons distinguer six classes de journées différentes (« clusters ») qui séparent fortement les jours à fort et faible ensoleillement et donc à fort et faible taux de production PV (attribués aux profils journaliers 1-2 et 3-6, respectivement). Parmi les 6 classes, le profil journalier 1 contient 71 jours, le profil journalier 2 inclut 35 jours, le profil journalier 3, lui, 71 jours, le profil 4 englobe 41 jours, le profil 5 71 jours et enfin le profil journalier 6 contient 76 jours. Les configurations spatiales associées aux six profils sont présentées à la figure 5.8. Il est notable de distinguer les deux premiers profils non pas par la différence de production PV sur la globalité de l'île, mais sa répartition géographique.

En effet, sur le profil journalier 1, la production PV est supérieure sur un axe du Nord-Ouest au Sud-Est de l'île, tandis que sur le profil 2, nous observons un fort taux de production PV sur le Nord-Est - Est et dans une moindre mesure sur la côte Sud-Ouest de l'île de La Réunion. Les profils 3 à 6 affichent des taux de production PV globalement plus bas que les deux premières classes avec des répartitions différentes sur l'île. En effet, le profil 3 se démarque par des journées avec un faible rayonnement solaire et de plus basses températures sur tout le sud de l'île avec comme résultat une plus faible production PV localisée dans le sud. Le profil 4 se distingue par une production plus importante que le reste de l'île sur la zone géographique nord-ouest et sud-est. Quant au profil journalier 5, il ressemble au profil 1 mais avec une production PV globalement plus basse. Nous retrouvons une production PV plus élevée que sur le reste de l'île sur cet axe nord/ouest-sud-est. Finalement, le profil journalier 6 présente une production PV proche spatialement du profil journalier 2 mais globalement plus faible. La production PV plus élevée dans la partie nord-nord-est/sud que sur le reste de La Réunion.

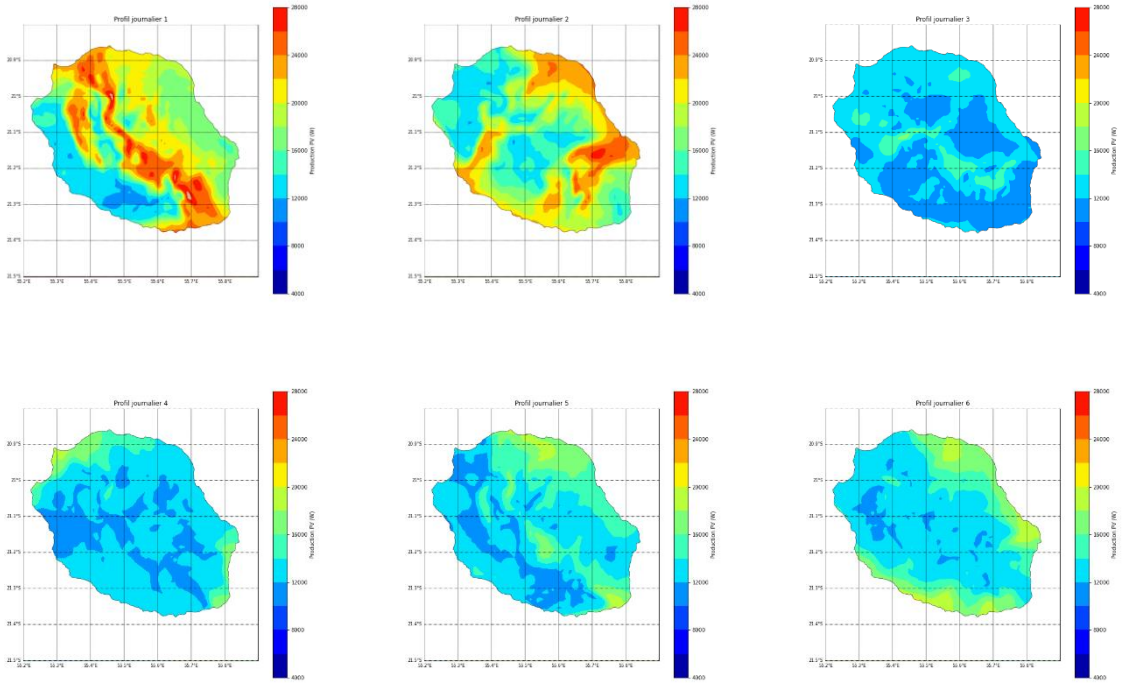


Figure 5.8 Les 6 profils journaliers issus de la classification ascendante hiérarchique

5.2.4.3 Cartes de prévisions de production PV journalière en 2018

La figure 5.9 représente les cartes de prévision de production PV journalière en 2018 pour les quatre modèles étudiés pour les 6 profils journaliers précédemment expliqués ; la référence en termes de production PV étant les profils de la figure 5.8 sur l'ensemble du territoire réunionnais.

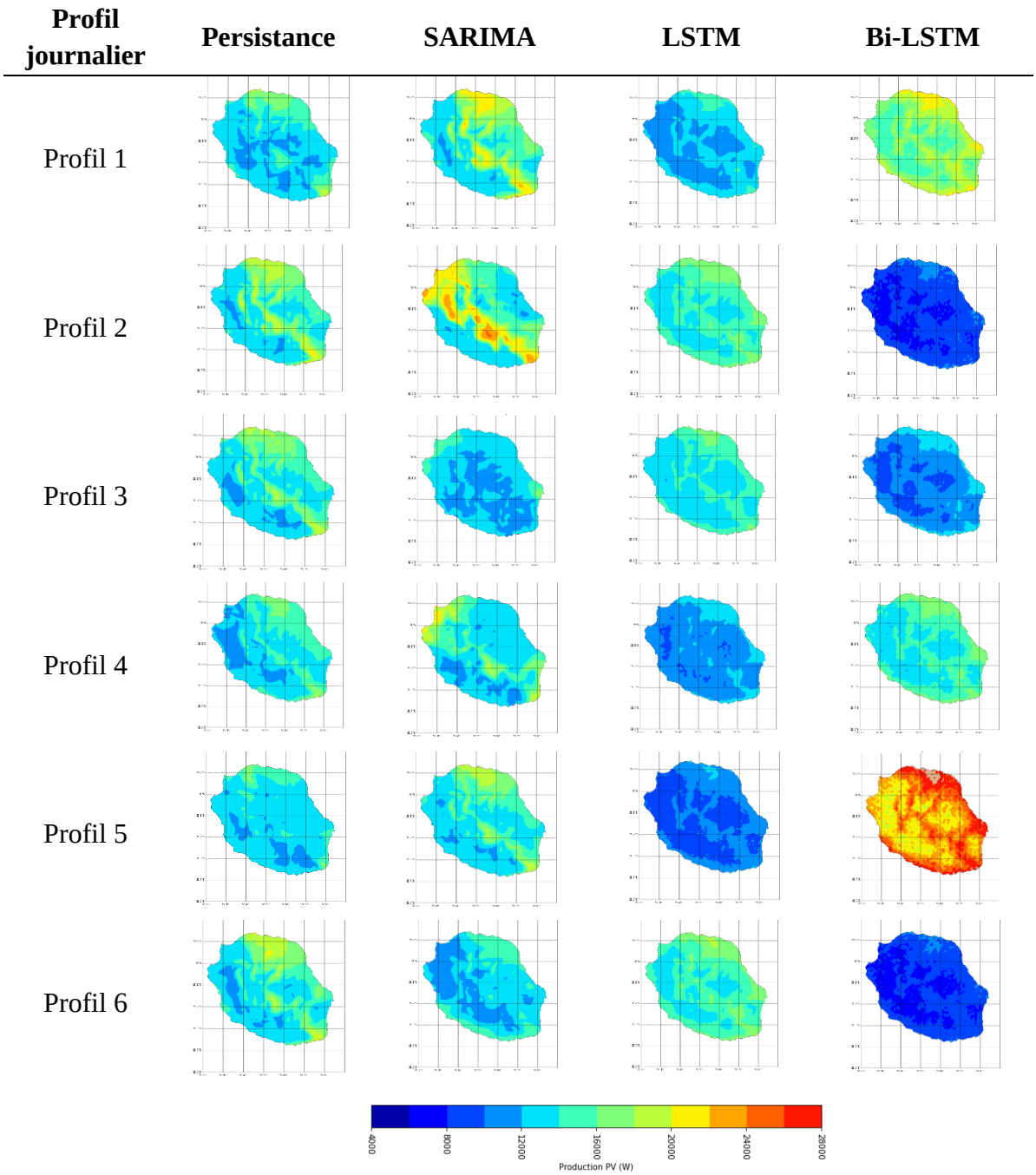


Figure 5.9 Cartes de prévision de production PV journalière selon les 6 profils journaliers pour les modèles de persistance, SARIMA, LSTM et Bi-LSTM (de gauche à droite). L'échelle des couleurs de la production PV, comprise entre 4 kW et 28 kW est située en bas de la figure

Tableau 5.3 Résultats de la comparaison entre les différentes méthodes de prévision de production PV journalière selon les 6 classes de jours déterminées. Sont inclus l'erreur quadratique moyenne, l'erreur quadratique moyenne normalisée, l'écart-type, le coefficient de Pearson et le temps de calcul pour chacune des simulations.

Profil	Métriques	Persistance	SARIMA	LSTM	Bi-LSTM
1	RMSE (W)	3906	3931	**2612**	4959
1	nRMSE (%)	24.8	25	**13.26**	34.67
1	StD (W)	**2466**	2436	2114	1682
1	r	0.65	0.64	**0.78**	0.76
2	RMSE (W)	4089	**4065**	4951	5692
2	nRMSE (%)	23.75	**23.63**	33.15	25.74
2	StD (W)	**2913**	2878	1575	2568
2	r	0.31	**0.32**	0.23	0.19
3	RMSE (W)	3707	3671	**3264**	6731
3	nRMSE (%)	22.44	22.23	**19.49**	33.1
3	StD (W)	2640	2607	**1769**	2331
3	r	0.24	0.25	**0.45**	0.41
4	RMSE (W)	3214	**3183**	6263	4987
4	nRMSE (%)	21.16	**20.94**	28.89	24.87
4	StD (W)	1822	1801	2364	**2377**
4	r	0.28	0.29	**0.36**	0.31
5	RMSE (W)	**2635**	2643	7654	6144
5	nRMSE (%)	**18.52**	18.54	32.35	58.62
5	StD (W)	1371	1355	**2620**	1654
5	r	0.71	0.69	**0.72**	0.49
6	RMSE (W)	3450	3418	**991**	7951
6	nRMSE (%)	20.22	20.06	**6.89**	35.98
6	StD (W)	2778	2744	**1519**	2572
6	r	0.65	0.66	**0.79**	0.77
1 à 6	Temps de calcul moyen (s)	**43**	186	3789	31936

Le tableau 5.3 résume les résultats de la comparaison lors de simulations les 4 méthodes de prévision de production PV journalière selon la classe du jour.

Deux constats peuvent être faits. Dans un premier temps, de manière générale, les performances de la prévision à j+1 de tous les modèles chutent en comparaison de celles des prévisions mensuelles à M+12. En effet, le *nRMSE* du modèle passe de plus ou moins 10% à plus de 30% dans certains cas et un coefficient de corrélation *r* plus bas. Dans un deuxième

temps, les temps de calculs sont plus longs pour les deux modèles neuronaux, le modèle le plus rapide étant toujours le modèle de persistance avec un temps de calcul de 43s. Le modèle Bi-LSTM a un temps de calcul de 31936 soit 700 fois plus long que le modèle simple.

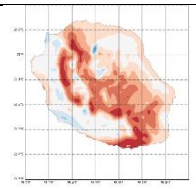
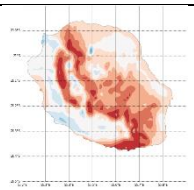
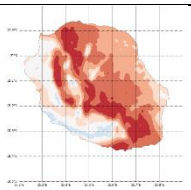
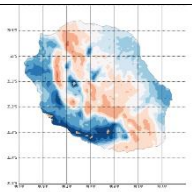
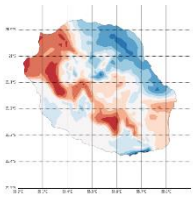
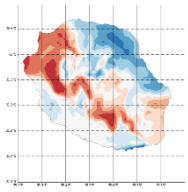
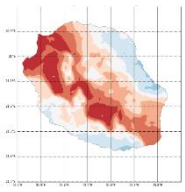
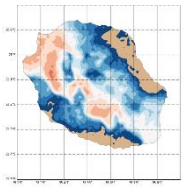
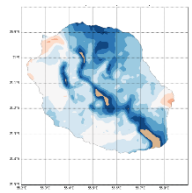
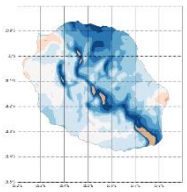
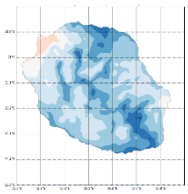
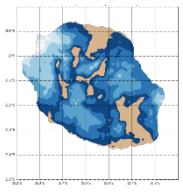
Il est plus difficile de mettre en avant un modèle plutôt qu'un autre si l'on considère les 6 profils dans leur intégralité. D'un point de vue statistique, en prenant en considération les erreurs dans leur globalité sur les 10000 points de grille, nous obtenons pour :

- le profil journalier 1 : le modèle LSTM est le plus performant avec respectivement un *nRMSE* de 13.26 % et r = 0.78.
- le profil journalier 2 : le modèle SARIMA est performant même nous notons une chute de précision de tous les modèles avec un r = 0.32 et un *nRMSE* = 23.63%.
- le profil journalier 3 : le modèle LSTM est le plus performant aussi avec un r = 0.45 et un *nRMSE* = 19.5%.
- le profil journalier 4 et 5 : les modèles de persistance et SARIMA se distinguent pour ces profils de jours avec pour le profil 3 des *nRMSE* ≃ 21% et pour le profil 4 des *nRMSE* ≃ 18.5%.
- le profil journalier 6 : le modèle LSTM est de nouveau le plus performant avec un r=0.79 et un *nRMSE* = 6.9 %.

Bien que nécessitant plus de puissance et de temps de calcul, le modèle Bi-LSTM est moins performant dans l'ensemble sur les différents profils journaliers pour des prévisions à j+1.

5.2.4.4 Cartes d'erreurs de prévision de production PV journalière en 2018

Comme pour le paragraphe précédent, notons que les métriques telles que celles présentées dans le tableau 5.3 permettent d'évaluer globalement statistiquement chaque modèle, mais ne mettent pas en avant les erreurs d'un point de vue spatial, d'où l'intérêt de présenter dans un deuxième temps les cartes d'erreurs de prévision PV mensuelle. La figure 5.10 représente les cartes d'erreurs de prévision pour les 6 profils journaliers en 2018.

Profil journalier	Persistance	SARIMA	LSTM	Bi-LSTM
Profil 1				
Profil 2				
Profil 3				

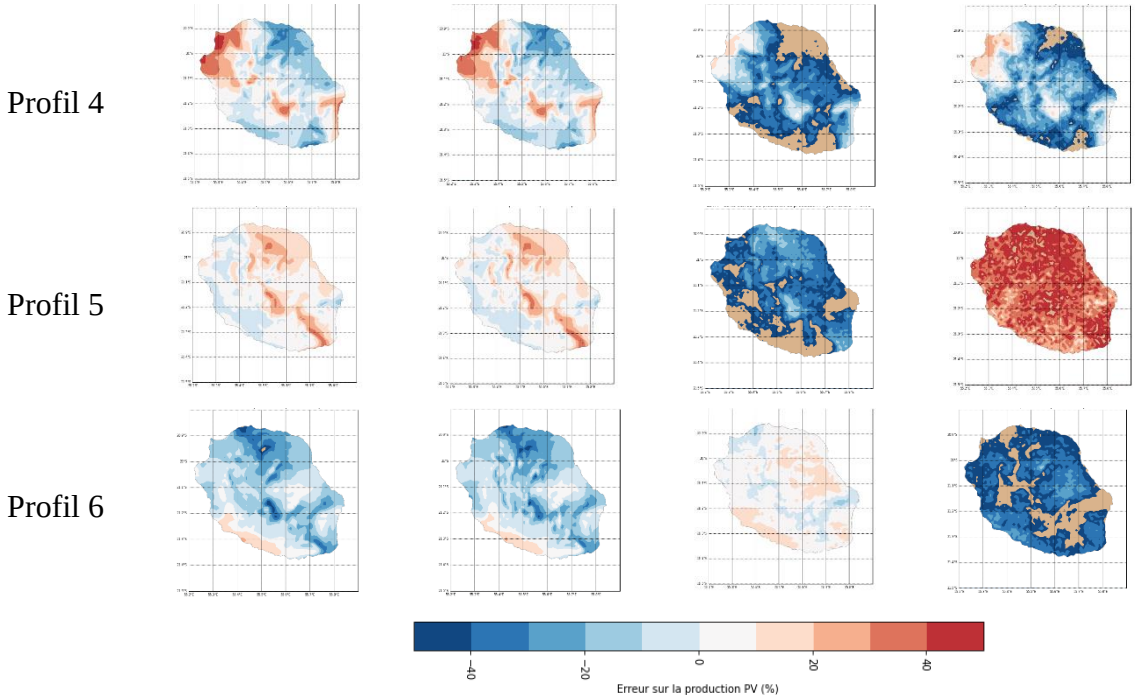


Figure 5.10 Cartes journalières d'erreur de production PV pour les 6 profils journaliers pour les modèles de persistance, SARIMA, LSTM et Bi-LSTM (de gauche à droite). L'échelle des couleurs de l'erreur de production PV, compris entre -50% et 50%, est située en bas de la figure.

En observant en détail ces cartes d'erreur sur le territoire réunionnais dans sa globalité, nous pouvons mettre en exergue que le modèle de persistance et SARIMA présentent les mêmes caractéristiques c'est-à-dire une surestimation de la production PV sur l'axe nord-ouest/sud-est de l'île pour les profils journaliers 1, 2 et 5 dans une moindre mesure mais une sous-estimation de la production PV dans la même région sur le profil journalier 3 et 6. Le modèle LSTM et Bi-LSTM présente parfois une forte erreur spatiale avec des erreurs entre la carte de prévision et la carte référence dépassant dans certaines zones (représentées en marron) les 50% d'erreur (profil 3 et 4 pour le LSTM et profil 2, 3, 4 et 6 pour le Bi-LSTM), principalement dans les zones littorales sud et nord. Cela peut s'expliquer par des modèles neuronaux pas assez optimisés et aussi par le faible jeu de données pour certains profils journaliers. Aussi, nous pouvons observer un coefficient de corrélation de Pearson faible peu importe le modèle utilisé. Ainsi pour la classe journalière 2 et 4 comportant respectivement 35 et 41 jours, les coefficients sont les plus bas avec un $r \leq 0.36$, quelle que soit la méthode de prévision utilisée.

Les cartes de prévision de production PV journalière pour les 6 classes de journée sur l'année 2019 sont en Annexe A. L'étape suivante est prévoir la production PV au pas de temps horaire au sein d'une journée pour chaque profil journalier.

5.2.5 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. Cet horizon correspond notamment au temps nécessaire pour alimenter le réseau en énergie issue des moteurs thermiques (~60 % de la puissance totale

installée à l'île de La Réunion). Dans ce paragraphe, nous allons appliquer les méthodes précédemment décrites pour le cas horaire. 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.

Nous utilisons les données brutes simulées par le modèle régional climatique WRF, en ne conservant que les données entre 8h et 17h, plage horaire constante de production solaire PV sur l'année. Nous avons un jeu de données contenant l'équivalent de 10 cartes horaires par jour avec 100 points en longitude et en latitude comme précédemment avec une matrice de données de production PV de $10 \times 100 \times 100$ par jour soit $10 \times 365 \times 100 \times 100$ par année.

Le but des manipulations initiées est de sélectionner les modèles les plus performants pour prédire la production PV pour l'(les) heure(s) suivante(s). Les modèles évalués sont la persistance, le SARIMA, le LSTM et le Bi-LSTM et sont optimisés manuellement :

- la persistance ne nécessite pas d'optimisation
- SARIMA avec $p = 5$ et $q = 0$
- LSTM avec 10 neurones sur la couche d'entrée, 20 nœuds cachés, une fonction d'activation de tangente hyperbolique « *tanh* », 200 *époques* avec une optimisation de type « *adam* », une fonction de perte d'erreur moyenne quadratique « *rmse* » et une évaluation du modèle durant l'apprentissage et le test par une métrique de performance de type « *mape* »
- Bi-LSTM utilise les mêmes paramètres que le modèle LSTM.

Pour les modèles de persistance et SARIMA, les heures comprises entre 8h et 17 pour les 365 jours de l'année 2017 ont servi de base d'apprentissage et les données de l'année 2018 ont été utilisées comme base de test. Pour les modèles neuronaux, le temps et la puissance de calcul nécessaire étant beaucoup plus important, il a été nécessaire d'entrainer et de tester sur un échantillon de données horaires plus restreint, en l'occurrence sur un seul profil journalier. Le profil journalier 1 a été choisi car la production PV est globalement plus grande et les modèles neuronaux ont été plus performants. Comme pour les autres horizons temporels de prévision, l'année 2018 de production PV sera par la suite considérée comme la référence pour toutes les sorties de prédictions des différentes méthodes. Ces six profils ont été évoqués et expliqués dans la section 5.2.4.2.

5.2.5.1 Cartes de production PV horaires (classe journalière1) en 2018

Dans un souci de lisibilité du manuscrit, seule l'étude sur le profil journalier1 (choisi arbitrairement) sera présentée dans cette partie. Les cartes horaires de production PV des profils journaliers 2 à 6 seront en Annexe B. La figure 5.11 présente, pour la classe journalière 1, les cartes de références. Ces cartes sont des cartes moyennées à chaque heure de la journée entre 8h et 17h pour l'année 2018. Rappelons que la classe 1 englobe 71 jours sur les 365 de l'année 2018.

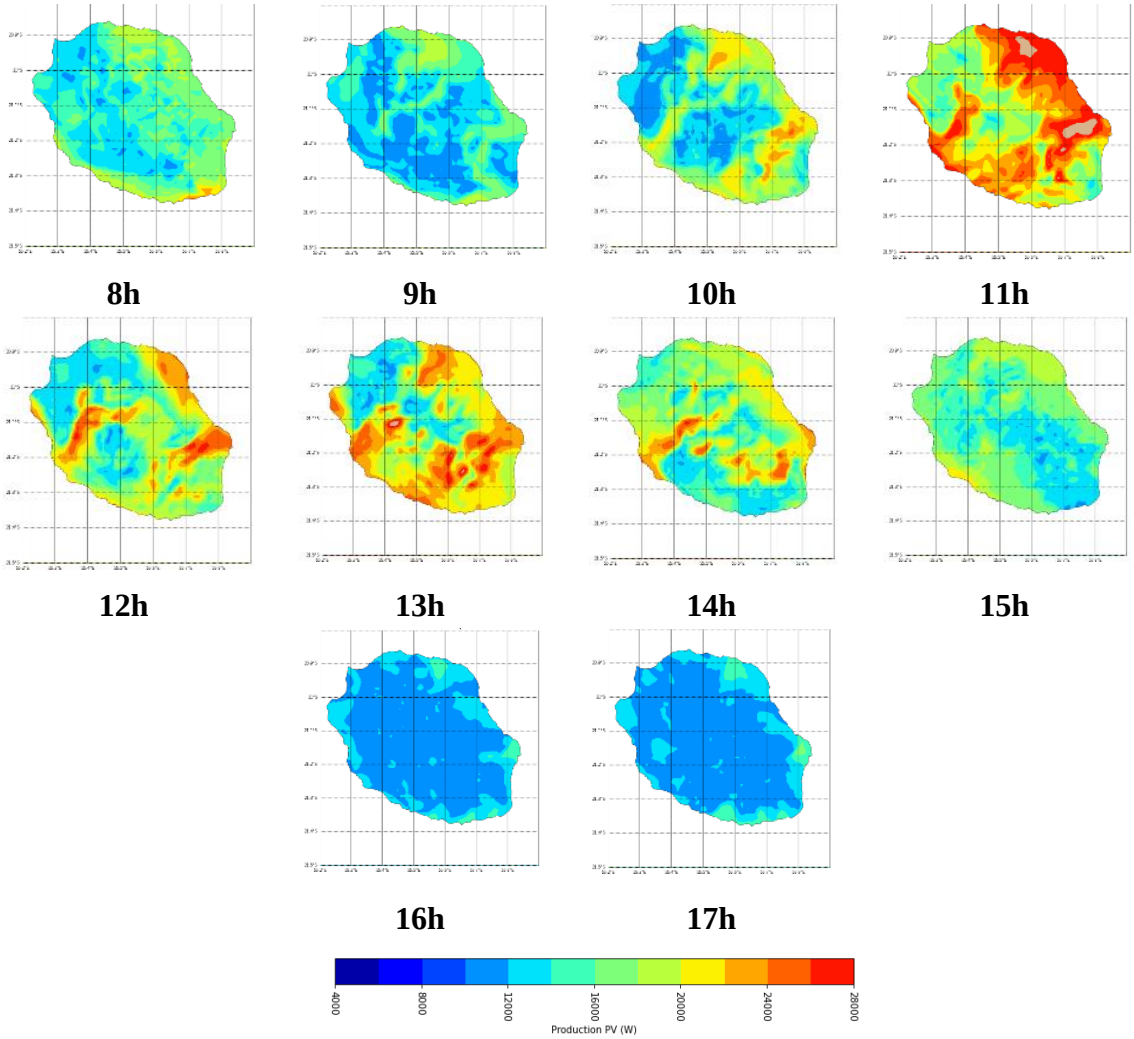
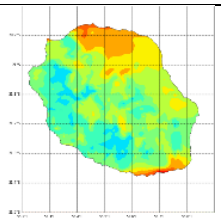
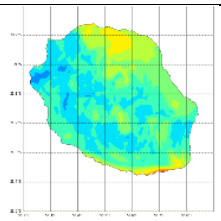
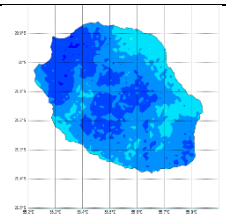
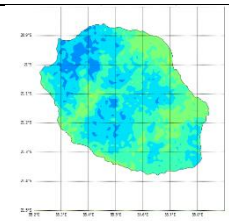
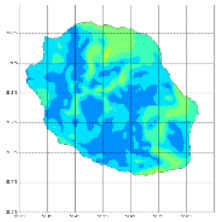
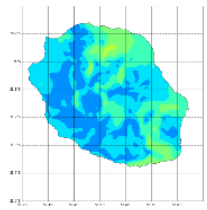
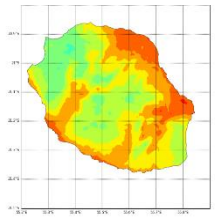
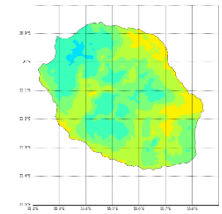
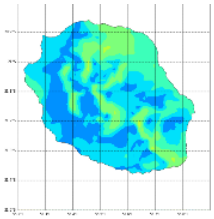
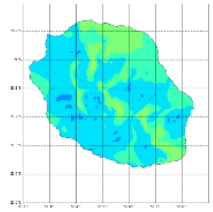
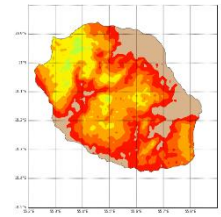
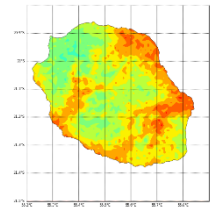
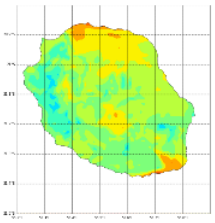
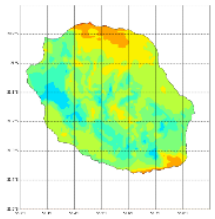
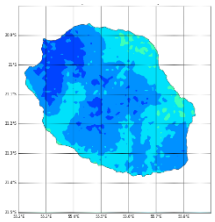
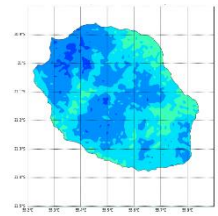
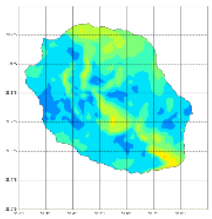
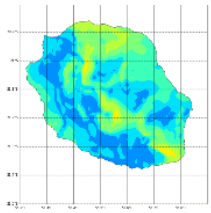
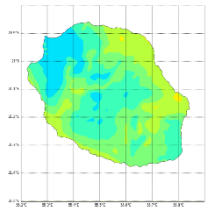
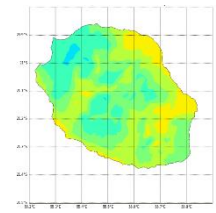
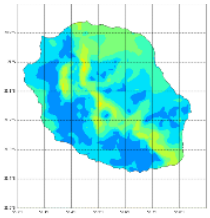
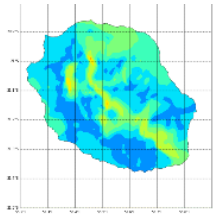
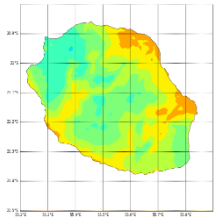
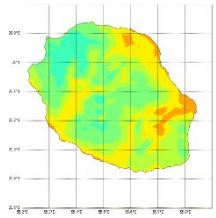
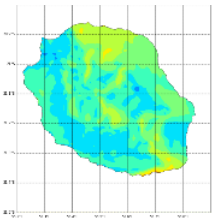
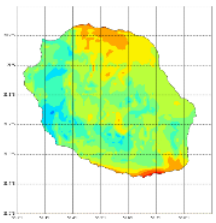
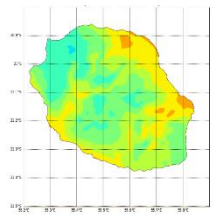
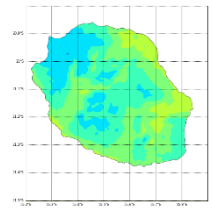


Figure 5.11 Cartes réunionnaises de référence des profils horaires de production PV pour la classe journalière 1 en 2018 entre 8h et 17h. L'échelle des couleurs de la production PV, comprise entre 4 kW et 28 kW est située en bas de la figure

Comme nous pouvons le constater la production PV, faible en début de journée, augmente progressivement au cours de la journée pour atteindre son pic de production entre 11h et 13h, avant de diminuer graduellement en fin de journée (d'où l'intérêt d'arrêter la récolte de données à partir de 17h).

5.2.5.2 Cartes de prévisions de production PV horaires (classe journalière1) en 2018

La figure 5.12 représente les cartes de prévision de production PV horaires pour la classe journalière 1 en 2018 pour les quatre modèles étudiés ; la référence en termes de production PV étant les profils de la figure 5.11 sur l'ensemble du territoire réunionnais. Les cartes prévisionnelles horaires correspondant aux autres profils journaliers sont en Annexe B.

Heures	Persistance	SARIMA	LSTM	Bi-LSTM
8h				
9h				
10h				
11h				
12h				
13h				
14h				

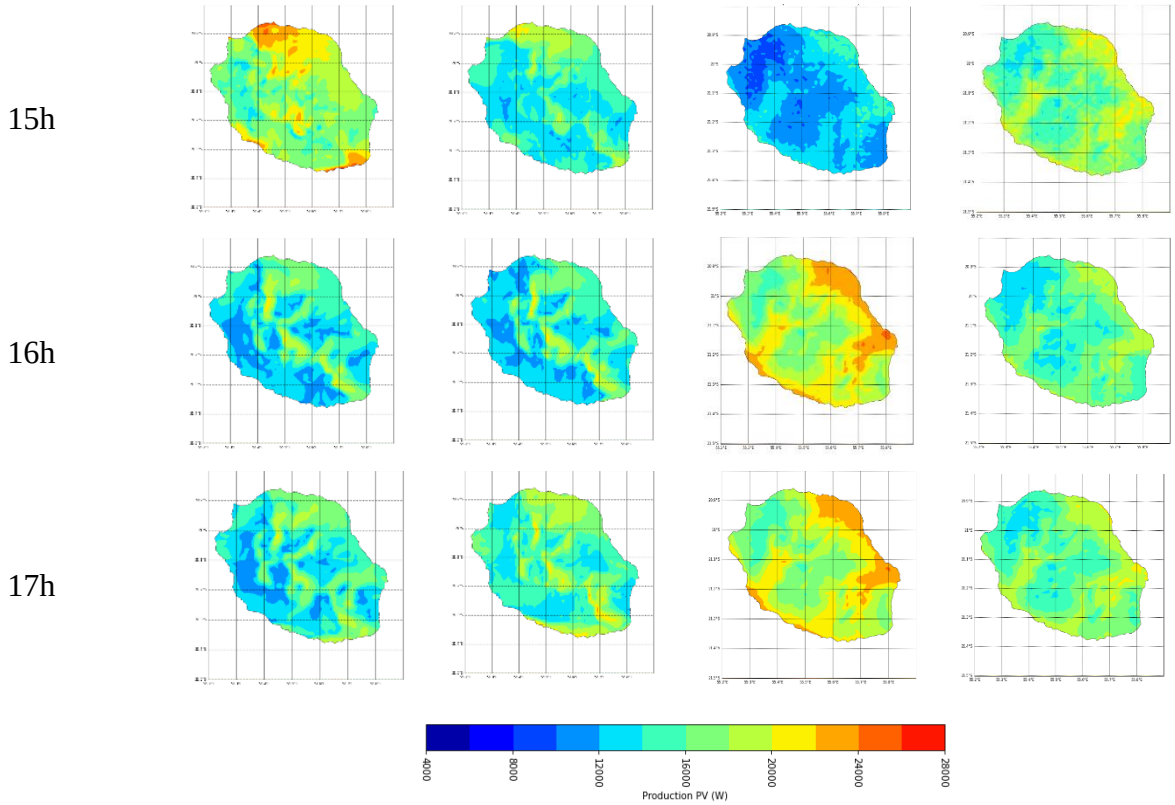


Figure 5.12 Cartes de prévision de production PV horaires sur une journée type de classe journalière 1 en 2018 pour les modèles de persistance, SARIMA, LSTM et Bi-LSTM (de gauche à droite). L'échelle des couleurs de la production PV, comprise entre 4 kW et 28 kW est située en bas de la figure

Tableau 5.4 Résultats de la comparaison entre les différentes méthodes de prévision de production PV horaires (classe journalière 1). Sont inclus l'erreur quadratique moyenne, l'erreur quadratique moyenne normalisée, l'écart-type, le coefficient de Pearson et le temps de calcul pour chacune des simulations.

Heure de la journée	Métriques	Persistance	SARIMA	LSTM	Bi-LSTM
8h	RMSE (W)	4391	4296	5187	**2072**
8h	nRMSE (%)	21.45	21.06	42.07	**12.51**
8h	StD (W)	3232	3163	1586	**2159**
8h	r	0.48	0.49	0.58	**0.59**
9h	RMSE (W)	**1748**	1774	7807	4281
9h	nRMSE (%)	**11.21**	11.35	32.53	21.12
9h	StD (W)	2387	2333	2974	**2477**
9h	r	**0.83**	0.81	0.68	0.66
10h	RMSE (W)	5394	5353	10009	**4764**
10h	nRMSE (%)	35.69	35.30	34.05	**19.96**
10h	StD (W)	2037	1989	**3710**	3035
10h	r	0.57	0.57	**0.87**	0.83

11h	RMSE (W)	**6044**	6029	10402	9916
11h	nRMSE (%)	**30.27**	30.29	81.04	69.62
11h	StD (W)	**2823**	2765	1673	1818
11h	r	0.05	0.013	**0.74**	0.65
12h	RMSE (W)	6487	6417	3217	**2499**
12h	nRMSE (%)	37.82	37.40	17.07	**12.23**
12h	StD (W)	**3410**	3331	2303	2476
12h	r	0.01	0.001	**0.83**	0.82
13h	RMSE (W)	7843	7778	2453	**2246**
13h	nRMSE (%)	49.13	48.62	11.44	**10.28**
13h	StD (W)	2766	2700	2637	**2851**
13h	r	0.12	0.13	0.79	**0.80**
14h	RMSE (W)	5122	5096	**2964**	3534
14h	nRMSE (%)	29.71	29.55	**14.05**	19.31
14h	StD (W)	2510	2456	**2593**	2211
14h	r	0.20	0.21	**0.60**	0.59
15h	RMSE (W)	5033	4940	3601	**3607**
15h	nRMSE (%)	24.47	24.10	26.19	**18.27**
15h	StD (W)	3247	3176	**1718**	2514
15h	r	0.12	0.12	**0.53**	0.47
16h	RMSE (W)	3620	**3611**	9785	5593
16h	nRMSE (%)	23.17	**23.06**	43.43	30.5
16h	StD (W)	2512	**2454**	2782	2232
16h	r	0.26	0.27	**0.79**	0.78
17h	RMSE (W)	3188	**3182**	8424	5669
17h	nRMSE (%)	19.41	**19.35**	38.10	29.29
17h	StD (W)	2403	2344	2727	**2340**
17h	r	0.62	0.62	0.80	**0.81**
1	Temps de calcul moyen (s)	**375**	30710	48889	30073

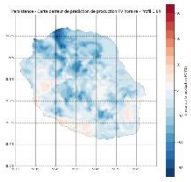
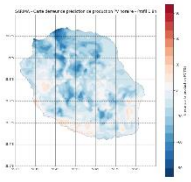
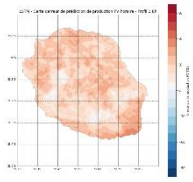
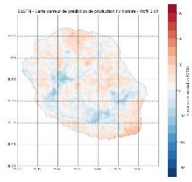
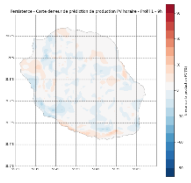
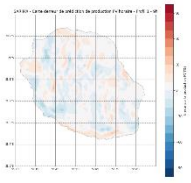
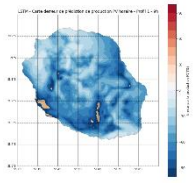
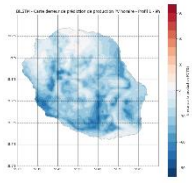
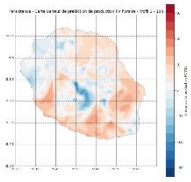
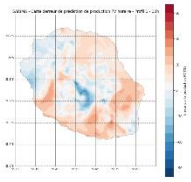
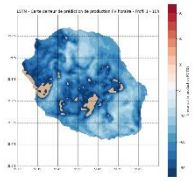
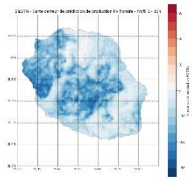
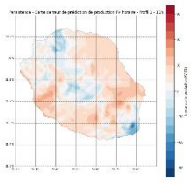
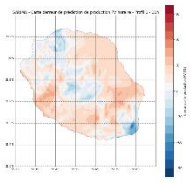
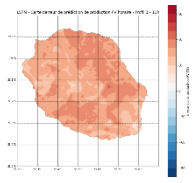
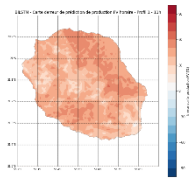
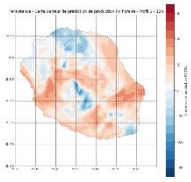
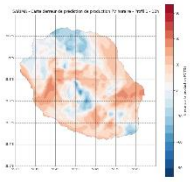
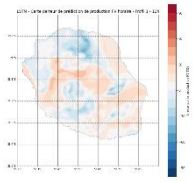
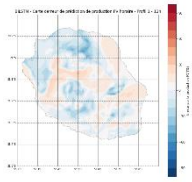
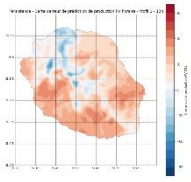
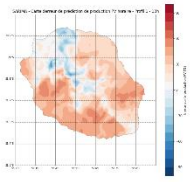
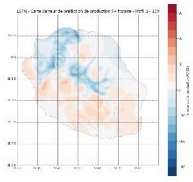
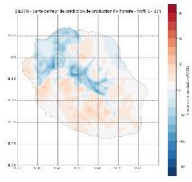
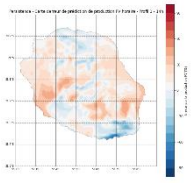
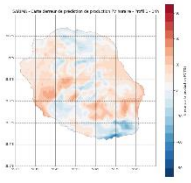
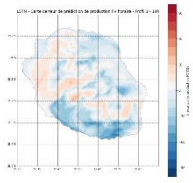
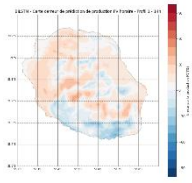
Le tableau 5.4 résume les résultats de la comparaison lors de simulations les 4 méthodes de prévision de production PV horaire uniquement pour la classe journalière 1(les autres étant en Annexe B).

Deux constats peuvent être faits. Dans un premier temps, de manière générale, les modèles sont moins performants et les temps de calculs sont rallongés. Le modèle de persistance affiche un temps de calcul moyen faible (375s) tandis que le modèle LSTM a pris près de 14 heures pour un seul profil journalier. Le second constat est que les modèles neuronaux ont été dans la plupart des cas plus performant et précis que les modèles de persistance et SARIAM. Le modèle Bi-LSTM donne des résultats très satisfaisants sur certaines plages horaires (8h, 10h, 12h et 13h), meilleur que les 3 autres avec un *nRMSE* égal à 12.5%, 20%, 12% et 10%. Notons qu'à 11h les deux modèles de réseaux de neurones LSTM et Bi-LSTM sont très mauvais avec des *nRMSE* $\simeq$ 81% et 69% respectivement. Cela peut s'expliquer par une mauvaise optimisation du modèle neuronale qui, ici, nécessite plus de précisions et de nombreuses simulations et un plus grand jeu de données comme base d'apprentissage. De plus, les données d'entrainement de ces deux modèles neuronaux étaient des données horaires sur les 71 jours regroupant la classe journalière 1 de l'année 2017 et le test sur la même classe en 2018. En effet, il s'agissait d'un choix à faire, car le temps de calcul et la mémoire nécessaire pour faire fonctionner ces simulations sont proportionnels à la taille des données utilisées en entrée de modèle. Face à ce peu de données, la performance des modèles neuronaux s'est avérée satisfaisant certes mais de meilleurs résultats sont attendus si plus de données en entrée sont apportées. Le modèle de persistance a une relative faible erreur dans son ensemble mais nous pouvons constater une augmentation du nRMSE sensiblement proportionnel à la variation de production PV au cours d'une journée type du profil journalier 1 (allant jusqu'à 49% d'erreur à 13h).

D'un point de vue statistique, si nous calculons les erreurs dans leur globalité sur les 10000 points de grille, les modèle LSTM et Bi-LSTM sont les plus performants tandis que pour les premières et dernières heures d'ensoleillement, c'est-à-dire surtout 8h-9h et 16h-17h le modèle de Persistance et/ou SARIMA sont très bien adapté. Ces modèles sont plus précis en début et fin de journée pour des plus faibles valeurs de productions PV.

Comme pour le paragraphe précédent, notons que les métriques telles que présentées dans le tableau 14 ne mettent pas en avant les erreurs de prévision d'un point de vue spatial. Nous présentons donc dans la suite les cartes d'erreurs de prévision de production PV horaire pour une journée de classe journalière 1. La figure 5.13 représente ces cartes d'erreurs de prévision obtenues.

5.2.5.3 Cartes d'erreurs de prévision de production PV horaires (classe journalière1) en 2018

Heures	Persistance	SARIMA	LSTM	Bi-LSTM
8h				
9h				
10h				
11h				
12h				
13h				
14h				

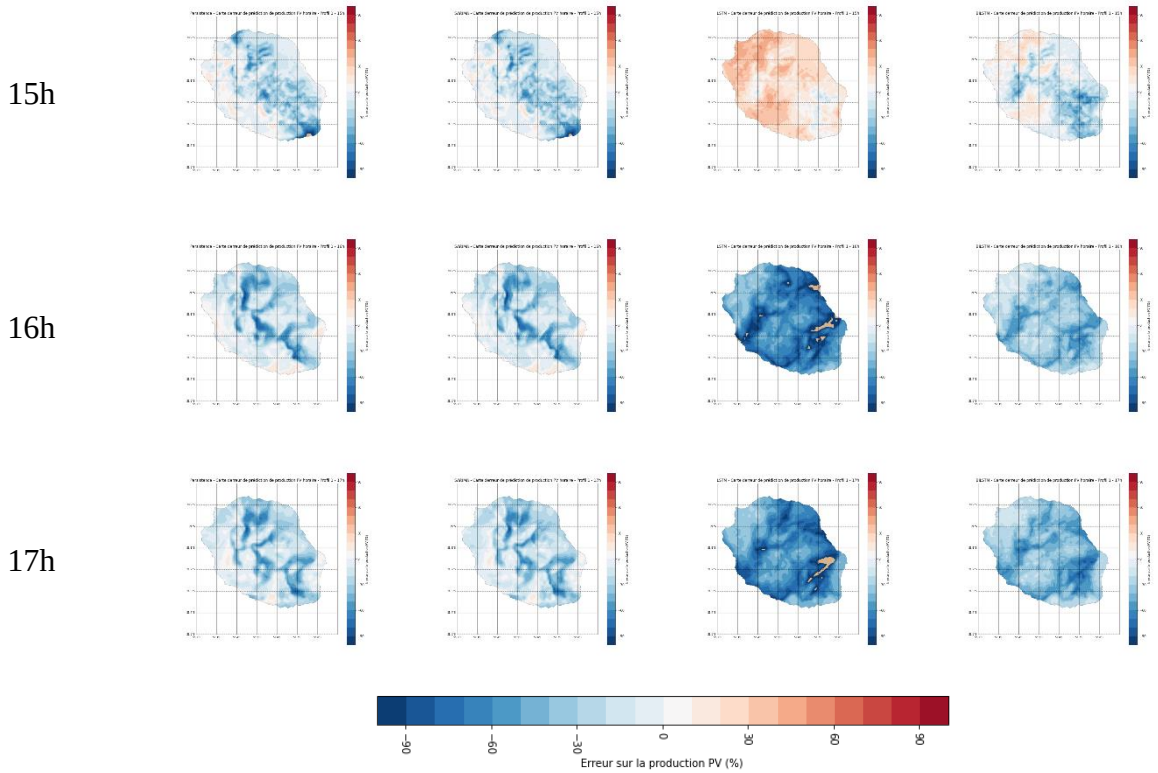


Figure 5.13 Cartes horaires d'erreur de production PV sur une journée de classe journalière 1 pour les modèles de persistance, SARIMA, LSTM et Bi-LSTM (de gauche à droite). L'échelle des couleurs de l'erreur de production PV, compris entre -100% et 100%, est située en bas de la figure

En observant en détail ces cartes d'erreur, il met en exergue que les prévisions des modèles de persistance et SARIMA ont des tendances spatiales très similaires, leurs différences se retrouvent plutôt d'un point de vue statistique comme vu dans le tableau 5.4. Il y a une incapacité de ces modèles à suivre l'évolution de la production PV sur la zone littorale nord et sud –sud-ouest, avec une sous-estimation de la production qui apparait en rouge sur les cartes d'erreurs entre 10h et 13h.

À 11h, nous voyons clairement une sous-évaluation des modèles LSTM et Bi-LSTM de la production PV sur toute l'île. Ces deux modèles neuronaux affichent de faibles erreurs d'un point de vue spatial entre 12h et 15h, heures de forte production PV, ce qui est un point intéressant à relever. Toutefois, le modèle LSTM présente de fortes erreurs spatiales avec des erreurs entre la carte de prévision et la carte référence dépassant dans certaines zones les 100% (zones en marron) notamment à 9h et 10h. Comme pour le créneau horaire de 11h, cela peut s'expliquer par des modèles neuronaux pas assez optimisés et aussi par le faible jeu de données pour ce profil journalier. Il aurait fallu une base d'apprentissage de plusieurs années pour capturer plus en détail les dépendances à long terme de la production PV à ces heures. Cela prouve qu'une plus grande base de données et une optimisation des hyperparamètres automatique seraient une perspective d'étude dans l'amélioration de la performance de ces modèles de prévision.

5.3 Discussion

Cette section va discuter des modèles de prévision développés au cours de cette thèse et des résultats obtenus pour les trois horizons temporels choisis : M+1, j+1 et h+1 dans le chapitre 5.

Concernant les prévisions mensuelles à M+1, le faible jeu de données en entrée avait formulé l'hypothèse que les modèles simples comme le modèle de persistance donneraient probablement les meilleurs résultats. Toutefois, le modèle LSTM classique a affiché de manière générale des performances meilleures que les autres modèles, avec une erreur de 7% par rapport à la carte de référence et un *nRMSE* =8,35*%*. Il performe mieux que le modèle Bi-LSTM qui affiche néanmoins des résultats satisfaisants mais pour un temps et puissance de calcul plus conséquent. Le modèle SARIMA a un écart-type *StD* proche de celui de la référence $StD_{ref.}$, qui suggère que le modèle est capable de reproduire la variabilité de la référence de manière précise.

Les cartes d'erreurs de prévision représentent les erreurs spatialement et temporellement, selon les mois de l'année. Les erreurs pour les quatre modèles ont été généralement faibles et se révèlent être comprises entre -20% et 20%. Le modèle LSTM et Bi-LSTM ne respectent pas parfaitement le plus la spatialité de la production PV sur l'île car elle est sous-estimée dans les hauts au centre de l'île et dans les cirques. Peu importe le modèle neuronal développé, nous pouvons en conclure que le faible volume de données mensuelles d'entrainement ne permet pas au modèle pour cet horizon temporel de reproduire la production à tous les mois de l'année et à toute altitude. Cette limite peut être due au faible jeu de données mais aussi à la faible optimisation des algorithmes.

Concernant la prévision journalière à j+1. Il faut noter la baisse de performance de l'ensemble des modèles. L'étude de la performance métrique s'est portée sur les différentes classes de journée à La Réunion. Il est plus difficile de mettre en avant un modèle plutôt qu'un autre si l'on considère les 6 profils dans leur intégralité. D'un point de vue statistique, pour les profils journaliers 1, 3 et 6, le modèle LSTM est le plus performant avec respectivement un *nRMSE* de 13.26 *%*, 19.5% et 6,9% et les modèles de référence persistance et SARIMA pour les autres profils journaliers (2, 4 et 5) avec respectivement un *nRMSE* de 23,6%, 21% et 18,5%.

Le modèle de persistance et SARIMA présentent les mêmes caractéristiques c'est-à-dire une surestimation de la production PV sur l'axe nord-ouest/sud-est de l'île pour les profils journaliers 1, 2 et 5 dans une moindre mesure mais une sous-estimation de la production PV dans la même région sur le profil journalier 3 et 6. Le modèle LSTM et Bi-LSTM présente une plus forte erreur spatiale que lors des prévisions mensuelles avec des erreurs dépassant parfois les 50% d'erreur (profil 3 et 4 pour le LSTM et profil 2, 3, 4 et 6 pour le Bi-LSTM), principalement dans les zones littorales sud et nord.

Concernant la prévision horaire à h+1, les temps de calculs sont rallongés avec le modèle LSTM prenant près de 14 heures pour les six profils journaliers. Les modèles neuronaux ont été dans la plupart des cas plus performant et précis que les modèles de persistance et SARIMA. Le modèle Bi-LSTM donne des résultats très satisfaisants sur certaines plages horaires (8h, 10h, 12h et 13h), meilleur que les 3 autres avec un *nRMSE* égal à 12.5%, 20%, 12% et 10%.

Les données d'entrainement de ces deux modèles neuronaux étaient des données horaires sur les 71 jours regroupant la classe journalière 1 de l'année 2017 et le test sur la même classe en 2018. Face à plus petit jeu de données, la performance des modèles neuronaux s'est avérée satisfaisant certes mais de meilleurs résultats sont attendus si plus de données en entrée sont apportées. Le modèle de persistance et SARIMA a une relative faible erreur dans son ensemble mais nous avons constaté une augmentation du nRMSE sensiblement proportionnelle à la variation de production PV, qu'ils sous-estiment, au cours d'une journée type du profil journalier 1. Ces modèles sont plus précis en début et fin de journée pour des plus faibles valeurs de productions PV.

Les modèles neuronaux LSTM et Bi-LSTM présentent de plus fortes erreurs en début de journée (à 9h, 10h et11h) cela peut s'expliquer par des modèles neuronaux pas assez optimisés et aussi par le faible jeu de données pour ce profil journalier. Nous pouvons émettre l'hypothèse qu'une base d'apprentissage de plusieurs années aurait été nécessaire pour capturer plus en détail les dépendances à long terme de la production PV à ces heures. Cela prouve qu'une plus grande base de données et une optimisation automatique des hyperparamètres seraient une perspective d'étude dans l'amélioration de la performance de ces modèles de prévision.

Conclusion

Les modèles de réseaux de neurones développés au cours de ces travaux de thèse, LSTM et Bi-LSTM, ont montré des résultats probants pour la prédiction à différents horizons temporels, notamment par rapport aux modèles de référence tels que le modèle naïf de persistance et le modèle SARIMA. Les performances des modèles neuronaux ont été évaluées sur trois horizons temporels différents, à savoir M+1, j+1 et h+1, et ont montré des résultats prometteurs.

En effet, les résultats ont montré que les modèles neuronaux ont pu capturer les tendances et les modèles complexes de la série temporelle, ce qui leur a permis de réaliser des prévisions plus précises que les modèles de référence. En particulier, le modèle LSTM a montré une grande capacité à apprendre des dépendances à long terme dans les séries temporelles, ce qui est crucial pour la prédiction à des horizons temporels plus éloignés.

Les modèles neuronaux offrent une méthode prometteuse pour la prédiction de séries temporelles dans le domaine énergétique, qui peut être améliorée par l'utilisation de techniques avancées telles que l'optimisation automatique des hyperparamètres, qui n'a pas été employé dans ces travaux.

En conclusion, les modèles neuronaux développés au cours de ces travaux de recherche ont montré de résultats concluants pour la prédiction de séries temporelles de production PV à différents horizons temporels, même si les contraintes en termes de temps et de capacité de calculs sont des freins à leur utilisation en temps réel.

Conclusions générales

Cette thèse a porté sur le développement d'un modèle de prévision spatio-temporelle de production photovoltaïque sur l'ensemble de l'île de La Réunion, afin d'apporter un outil permettant de prévoir ce productible fortement corrélée à des paramètres météorologiques et climatiques. Plusieurs d'outils de prévision existent concernant le rayonnement solaire et la production PV, mais la plupart se concentre sur l'aspect temporel de la prévision en négligeant la spatialité du problème et dans la majorité des cas, ces outils développés sont peu robustes et ne sont pas reproductibles à tout point d'un même pays. L'intérêt d'un outil de prévision spatio-temporelle prend alors tout son sens et encore plus pour un territoire comme celui de La Réunion où le photovoltaïque prend de plus en plus d'ampleur et dont la topographie est très marquée.

Après avoir exposé l'énergie solaire photovoltaïque, son fonctionnement et sa variabilité, il a été mis en évidence l'intérêt de la prévision du productible PV pour le réseau électrique réunionnais. Grâce à un état de l'art sur les modèles de prévisions photovoltaïques couramment utilisés dans la littérature, le premier chapitre a permis de sélectionner deux classes de méthodes de prévision à étudier : un modèle statistique et un modèle à réseau de neurones.

Plusieurs méthodes et outils sont alors envisageables, mais la taille des données météorologiques sous forme de séries chronologiques qui sont utilisés en entrée du modèle réduit le champ des possibilités. Ainsi, dans le chapitre 2, un besoin précis en données climatiques était indispensable. Pour obtenir une prévision spatio-temporelle, ces données devaient recouvrir la totalité de l'île de la Réunion et avec une haute résolution spatiale et temporelle. Ainsi, ces données climatiques ont été générées par le modèle régional de climat WRF à une résolution au km² et à l'heure sur une période de deux ans allant de janvier 2017 à décembre 2018. Le modèle WRF a de ce fait été utilisé comme un générateur de données climatiques cohérentes. La pertinence de ces données a été vérifiée après confrontation avec les données au sol Météo-France à certains points de la carte et avec des données satellitaires SARAH -2.1. La résolution spatiale et temporelle fine des sorties du modèle WRF permettent de respecter les variations locales climatiques en tout point de l'île de La Réunion.

Le chapitre 3 met en avant un modèle statistique et un modèle de réseaux de neurones et la combinaison pour créer un modèle hybride de prévision PV spatio-temporelle. La méthode statistique de cointégration de Johansen, empruntée à l'Économétrie, nécessite une analyse des séries temporelles météorologiques à disposition et de choisir parmi ces variables celles dites explicatives qui permettent d'améliorer la prévision d'une autre variable, celle à expliquer, notamment dans ces travaux de thèse, la production PV. Ces variables explicatives sont choisies par la méthode de la causalité de Granger qui permet de les identifier. L'autre catégorie de méthode retenue est le réseau de neurone récurrent, très performant pour la prévision de séries temporelles également. Après un nouvel état de l'art sur les modèles à mémoire court et long

terme (LSTM), deux méthodes ont été conservées : le LSTM et le Bi-LSTM, une variante du modèle LSTM car jugées très performantes sur la prévision de production PV dans la littérature.

Les modèles et les données étant confirmés, le chapitre 4 permet de tester la méthode de cointégration de Johansen sur des données à la Possession afin de valider le modèle sur un jeu de données météorologiques et énergétiques connu en un point de l'île. De ce chapitre découle une relation entre la production PV (la variable à expliquer) et le rayonnement solaire, la vitesse du vent, la température et l'humidité (les quatre variables météorologiques explicatives choisies).

Le dernier chapitre combine la méthode de cointégration de Johansen à travers cette relation entre la production PV et les variables météorologiques et le modèle neuronal, LSTM ou Bi-LSTM. Ces méthodes hybrides utilisent en entrée les données générées par le modèle régional climatique WRF sur la période de 2017 et 2018. Ces outils sont entrainés sur 2017 et testés sur 2018 et prévoient enfin 2019. Pour attester de leur performance, ils sont confrontés à deux autres méthodes de référence retrouvées fréquemment dans la littérature, le modèle naïf de persistance et le modèle SARIMA. Des prévisions de production PV mensuelles (M+1), journalières (j+1) selon différentes classes journalières préétablies par classification ascendante hiérarchique et horaires (h+1) ont pu être établies et comparées à des cartes de référence sur l'année 2018. Les modèles LSTM et Bi-LSTM se sont montrés performants aux trois horizons temporels mais nécessite toutefois une optimisation précise (automatique) des hyperparamètres soit effectuée en amont des simulations. De plus, le faible volume de données climatiques avec qu'une seule année calendaire pour entrainer le modèle le rend moins robuste et performant qu'initialement souhaité.

Pour conclure, durant ces travaux de thèse, un modèle combinant une méthode statistique novatrice dans le domaine de l'Énergétique et une méthode récurrente neuronale LSTM et Bi-LSTM a été proposé et validé par simulation. Les résultats ont montré l'efficacité de la méthode et les versatilités spatiale et temporelle en fait d'elle un atout important surtout pour des régions où des variations météorologiques peuvent apparaitre localement du fait d'un relief marqué comme cela peut être le cas pour l'île de La Réunion.

Perspectives

De nombreuse perspectives et continuités à ces travaux peuvent être menées. Ces différents points sont classés et détaillés dans les paragraphes suivants. Ils doivent être améliorés ou travaillés afin d'approfondir les travaux de recherche.

Application des travaux de recherche à l'île Maurice

Un choix a été fait de ne travailler que sur les données de sortie WRF 4.1 concernant le sous domaine 3 concernant l'île de la Réunion comme cela est représenté sur la figure 5.14. Toutefois, dans des futurs travaux de recherche, les données concernant le sous domaine 4 c'est-à-dire l'île Maurice pour les années 2017 et 2018 ont aussi été simulées. Par soucis de temps, seules les données concernant l'île de La Réunion.

De plus, dans la comparaison, la validation et l'analyse des données climatiques WRF, des données au sol précises auraient été nécessaires à différents points de la carte mauricienne qui malheureusement n'étaient pas à disposition.

Il est donc envisageable, à condition d'avoir des données météorologiques au sol et des données de production PV précises et avérées, d'appliquer tout le processus de ces travaux de recherche sur l'île Maurice et d'en constater la pertinence. De plus, étant donné le paysage relativement plat avec seulement quelques collines et montagnes isolées de cette île, les résultats pourraient s'avérer plus probants que sur l'île de La Réunion. Des travaux de recherche, dans ce sens, sont menés actuellement à l'ile Maurice (Ramenah et al., 2023).

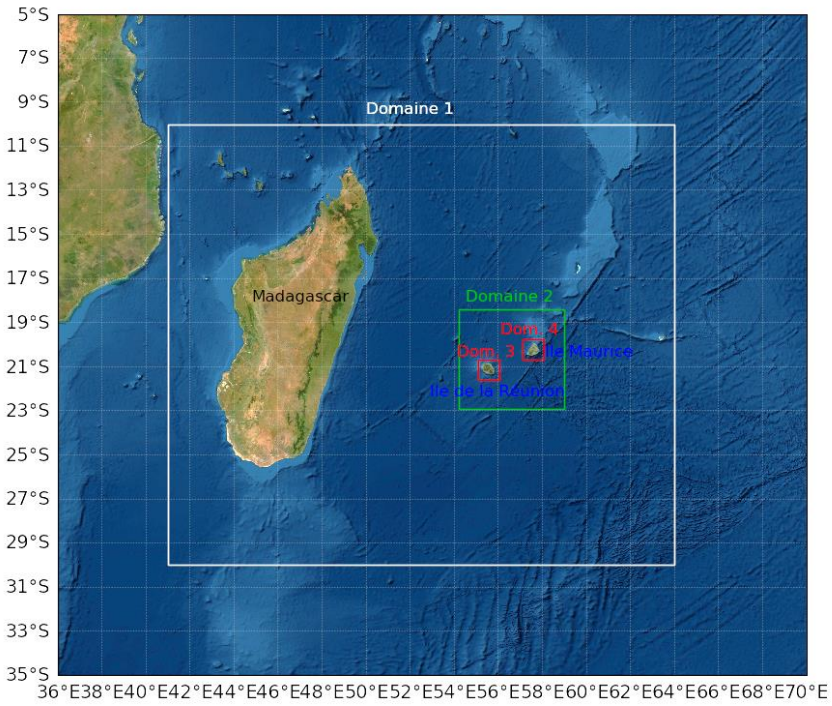


Figure P.14 Position des domaines imbriqués dans l'océan Indien

Prévision éolienne

Dans ces travaux de recherches, le choix de la production photovoltaïque a été fait comme énergie renouvelable à analyser et à prédire sur tout le territoire réunionnais. Toutefois, une autre ressource renouvelable, aussi en plein essor, aurait pu faire l'objet de ces travaux. En effet, la production d'énergie éolienne peut être étudiée en se basant sur les travaux effectués dans cette thèse en mettant dans un premier temps en lien les paramètres météorologiques impactant la ressource éolienne à La Réunion et sa production. Puis, en convertissant les données climatiques explicatives en données de production éolienne et enfin en prévoyant cette production sur le territoire. Les données récupérées étant aussi des séries temporelles, au même titre que les données de production PV, un modèle neuronal type LSTM, Bi-LSTM ou autre serait alors aussi judicieux. Le processus appliqué dans ces travaux de thèse à la production PV serait alors testé sur la production éolienne et ainsi des cartes de prévision de production éolienne de l'île de La Réunion pourraient être produites.

Optimisation du modèle de conversion PV de cointégration de Johansen

Le modèle de conversion PV de cointégration de Johansen est un modèle physique spécifique d'une centrale photovoltaïque et de ses spécificités en termes d'orientation, d'inclinaison, de la puissance crête installée ou encore de la technologie des panneaux utilisés.
Le but premier était de mettre en avant un modèle statistique encore jamais utilisé en PV. Toutefois, afin d'avoir un modèle de conversion plus reproductible sur l'île indépendamment des spécifiés de la centrale, un modèle de conversion PV reprenant ces différents paramètres pourrait être pertinent.

Prise en compte de l'incertitude et l'erreur sur le jeu de données WRF et de prévision

D'autres points qui nécessitent d'être évoqués sont l'incertitude et l'erreur sur les données météorologiques. Les données simulées par le modèle régional de climat WRF présentent des erreurs par rapport aux données climatiques satellitaires SARAH-2.1 et terrestre Météo-France. Cette erreur, certes quantifiée, n'a pas été prise en compte dans l'outil de conversion et dans la prévision de la production PV sur le territoire réunionnais. Ainsi, pour palier à cette erreur et incertitude, une prise en compte de celles-ci lors de la mise en place de l'algorithme d'apprentissage et de validation du modèle LSTM permettra d'améliorer considérablement la performance du modèle.

Optimisation des hyperparamètres neuronaux

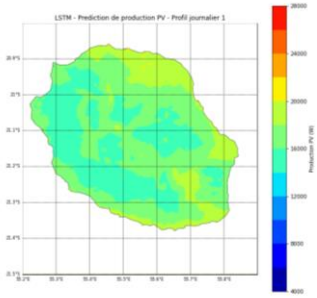
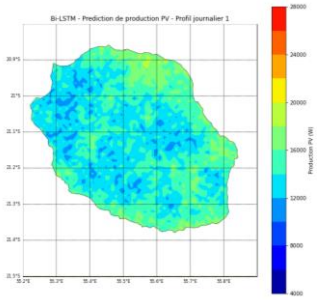
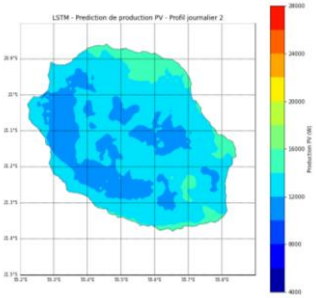
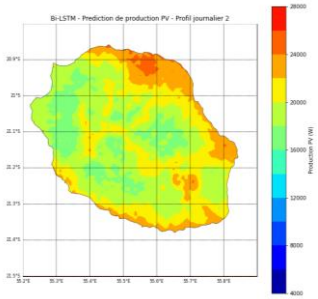
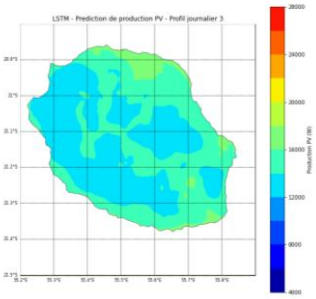
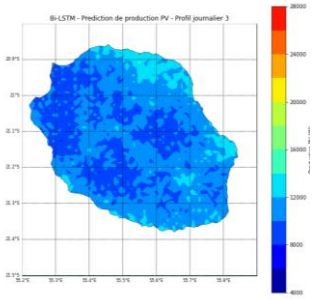
Dans les modèles neuronaux, les paramètres tels que les neurones, les époques et l'optimiseur sont en réalité des hyperparamètres qui, contrairement aux paramètres classiques, sont externes au processus d'entrainement du modèle, restent statiques et en définissent les propriétés. Leur optimisation a donc une influence majeure sur l'apprentissage des données et la performance du modèle. Toutefois, il existe énormément de combinaisons différentes quant à l'obtention de la bonne combinaison des hyperparamètres permettant une performance optimale

face au jeu de données utilisé. La méthode essais-erreurs est restée longtemps la seule approche pour optimiser les hyperparamètres (comme cela a été le cas dans ces travaux de recherche). Une boucle peut être mise en place au niveau extérieur de la boucle d'apprentissage et permettrait l'observation de l'influence de ces derniers. Il est possible d'utiliser certaines « toolboxes » récentes de *Keras*, notamment « *Talos* », qui afin d'effectuer une optimisation automatique, utilise des fonctions de perte afin de calculer les métriques de performances de notre modèle. Il peut utiliser des fonctions de perte de régression comme le Mean Squared Error (MSE) ou Mean Absolute Error (MAE) ou encore des fonctions de perte de qualification comme le Binary Cross Entropy ou le Multi-Class Cross Entropy. « *KerasTuner* » est un autre cadre d'optimisation des hyperparamètres évolutif qui résout les problèmes lié à la recherche d'hyperparamètres Il est livré avec les algorithmes d'optimisation bayésienne, d'hyperbande et de recherche aléatoire intégrés, et est également conçu pour être facilement extensible par les chercheurs afin d'expérimenter de nouveaux algorithmes de recherche. Il faut garder en tête que la puissance de calcul afin d'optimiser les hyperparamètres sont à renouveler face à chaque nouveau jeu de données et nécessite une puissance de calcul supplémentaire non négligeable en amont du modèle neuronal.

Annexes

Annexe A Prévision de production PV journalière en 2019

Dans le chapitre 5, nous n'avons présenté que les prévisions de production PV mensuelles sur l'année 2019 car au final, nous ne possédions pas de données au sol pour les analyser et valider. Ici, nous présentons les résultats de l'algorithme de prévision de production PV LSTM et Bi-LSTM avec un entrainement sur les données journalières par classe de journée en 2017 et un test sur les données journalières par classe en 2018 et une prévision sur les 6 profils journaliers en 2019.

Profil journalier	LSTM	Bi-LSTM
Profil 1		
Profil 2		
Profil 3		

Profil 4

Profil 5

Profil 6

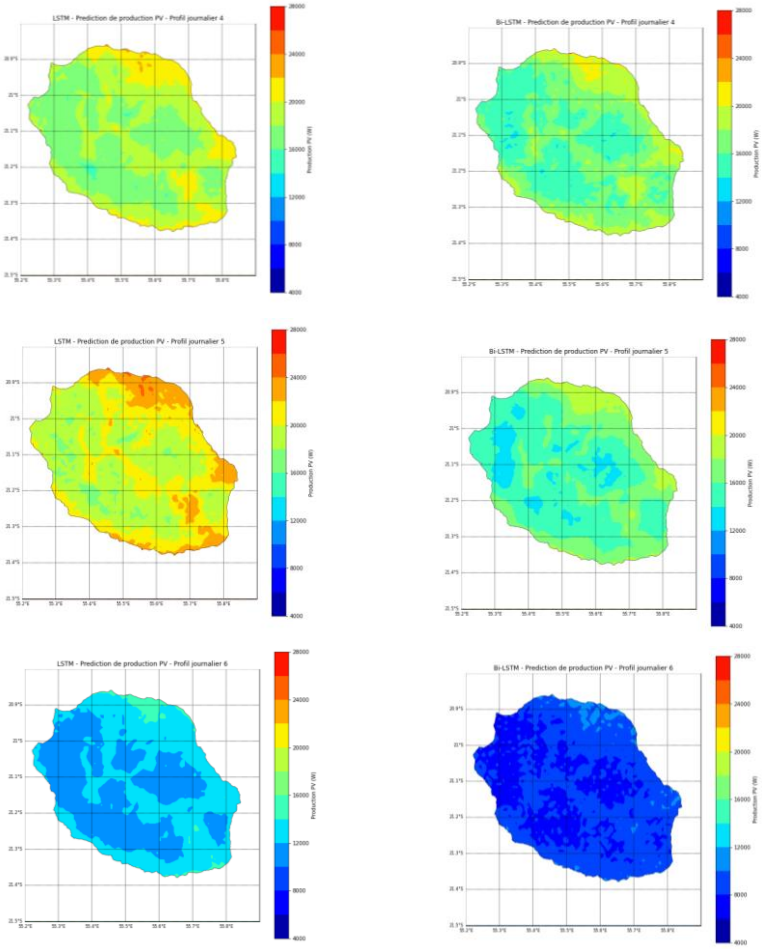


Figure A.15 Cartes de prévision de production PV journalière selon les 6 profils journaliers pour les modèles LSTM et Bi-LSTM (de gauche à droite) sur l'année 2019. L'échelle des couleurs de la production PV, comprise entre 4 kW et 28 kW

Annexe B Prévision de production PV horaire en 2018

Sont présentées dans cette annexe les cartes de prévision de production PV horaires simulées par le modèle LSTM et Bi-LSTM avec un entrainement sur les données à l'heure par classe journalière 2017 et testées sur les données horaires par profil journalier en 2018 pour les classes journalières 2 à 6 en 2019 (le profil journalier 1 apparait dans le chapitre5).

- **Pour le profil journalier 2**

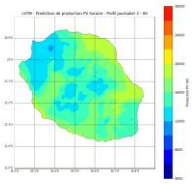
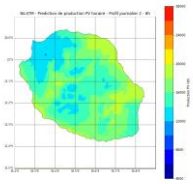
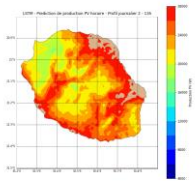
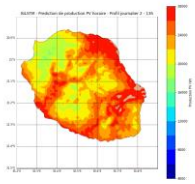
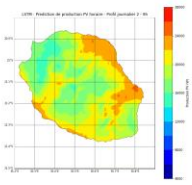
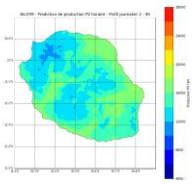
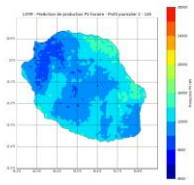
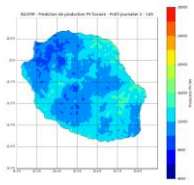
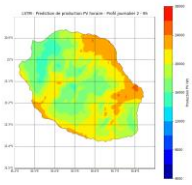
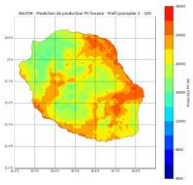
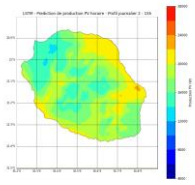
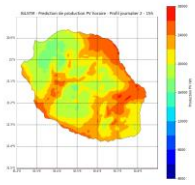
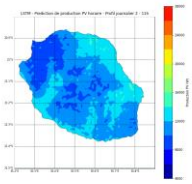
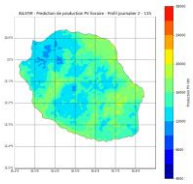
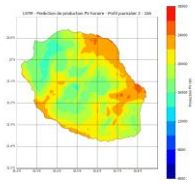
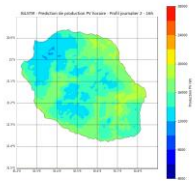
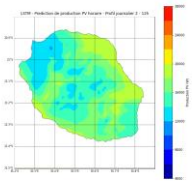
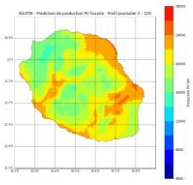
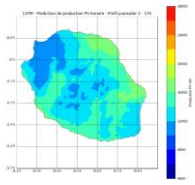
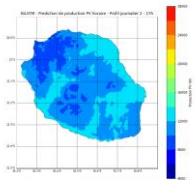
Heure	LSTM	Bi-LSTM	Heure	LSTM	Bi-LSTM
8h			13h		
9h			14h		
10h			15h		
11h			16h		
12h			17h		

Figure A.16 Cartes de prévision de production PV horaires sur une journée type de classe journalière 2 en 2018 pour les modèles LSTM et Bi-LSTM (de gauche à droite). L'échelle des couleurs de la production PV, comprise entre 4 kW et 28 kW

- **Pour le profil journalier 3**

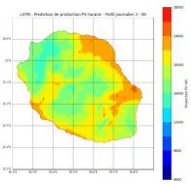
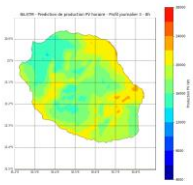
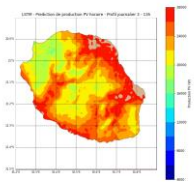
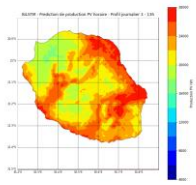
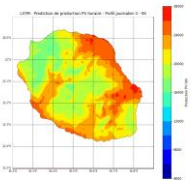
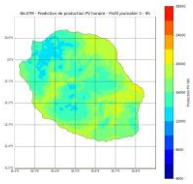
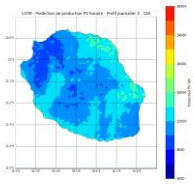
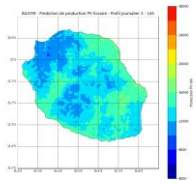
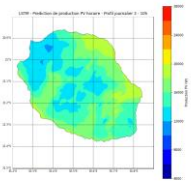
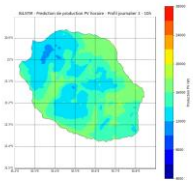
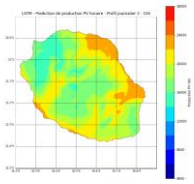
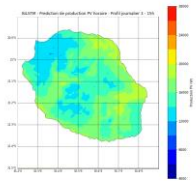
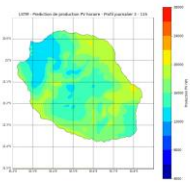
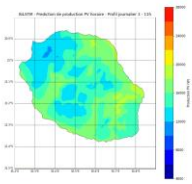
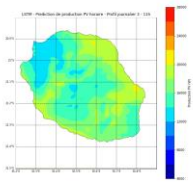
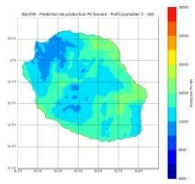
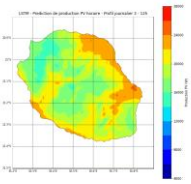
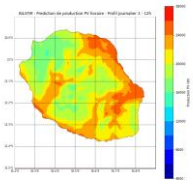
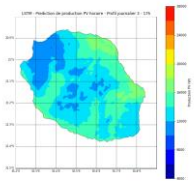
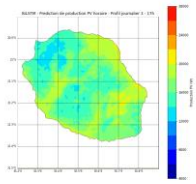
Heure	LSTM	Bi-LSTM	Heure	LSTM	Bi-LSTM
8h			13h		
9h			14h		
10h			15h		
11h			16h		
12h			17h		

Figure A.17 Cartes de prévision de production PV horaires sur une journée type de classe journalière 3 en 2018 pour les modèles LSTM et Bi-LSTM (de gauche à droite). L'échelle des couleurs de la production PV, comprise entre 4 kW et 28 kW

- **Pour le profil journalier 4**

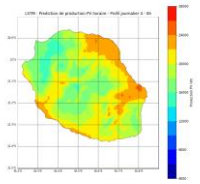
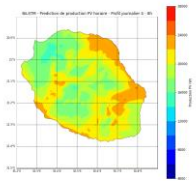
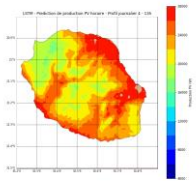
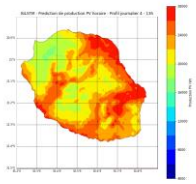
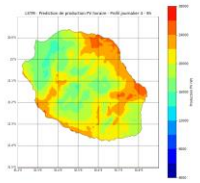
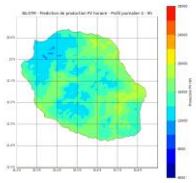
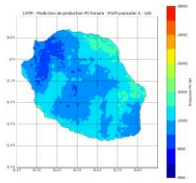
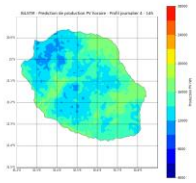
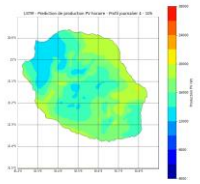
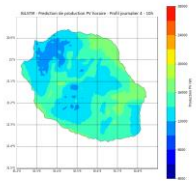
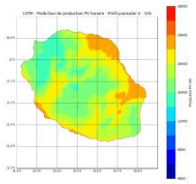
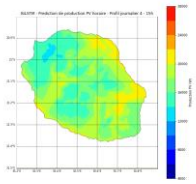
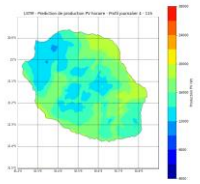
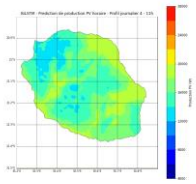
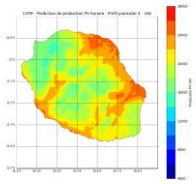
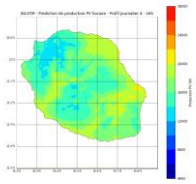
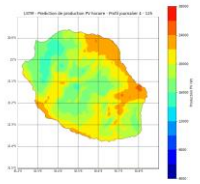
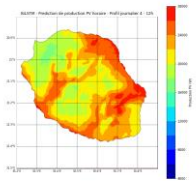
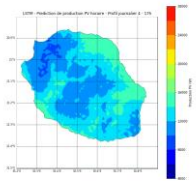
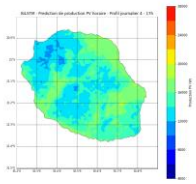
Heure	LSTM	Bi-LSTM	Heures	LSTM	Bi-LSTM
8h			13h		
9h			14h		
10h			15h		
11h			16h		
12h			17h		

Figure A.18 Cartes de prévision de production PV horaires sur une journée type de classe journalière 4 en 2018 pour les modèles LSTM et Bi-LSTM (de gauche à droite). L'échelle des couleurs de la production PV, comprise entre 4 kW et 28 kW

- **Pour le profil journalier 5**

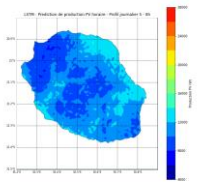
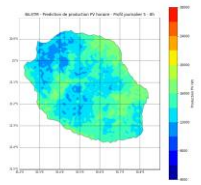
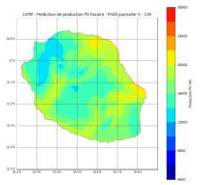
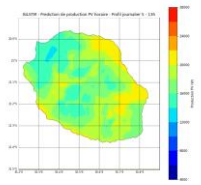
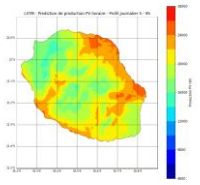
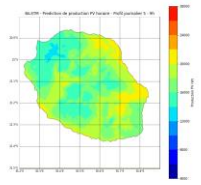
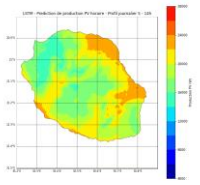
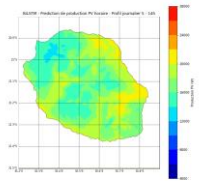
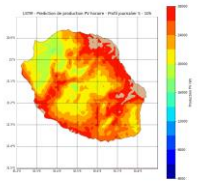
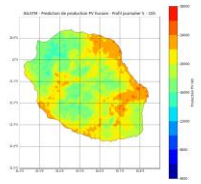
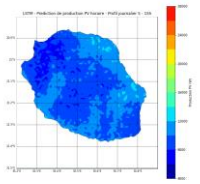
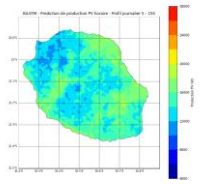
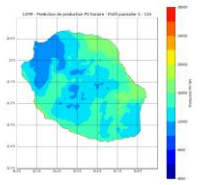
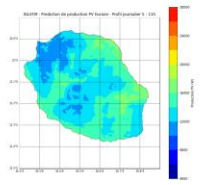
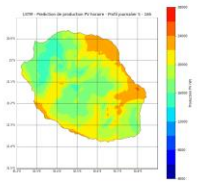
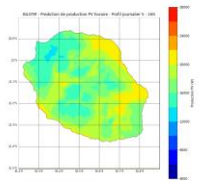
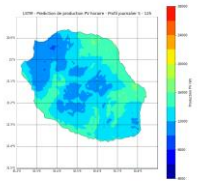
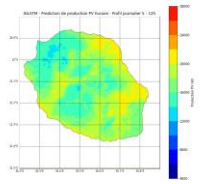
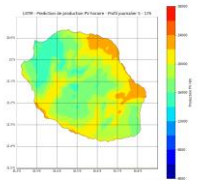
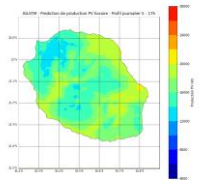
Heure	LSTM	Bi-LSTM	Heures	LSTM	Bi-LSTM
8h			13h		
9h			14h		
10h			15h		
11h			16h		
12h			17h		

Figure A.19 Cartes de prévision de production PV horaires sur une journée type de classe journalière 5 en 2018 pour les modèles LSTM et Bi-LSTM (de gauche à droite). L'échelle des couleurs de la production PV, comprise entre 4 kW et 28 kW

- **Pour le profil journalier 6**

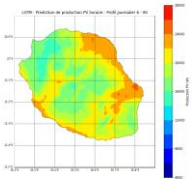
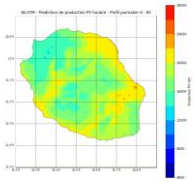
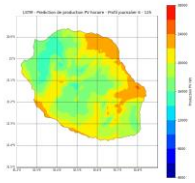
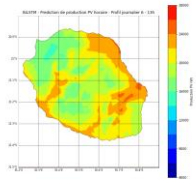
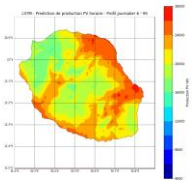
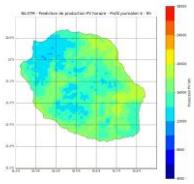
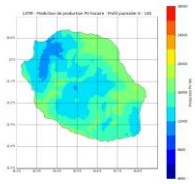
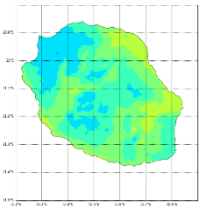
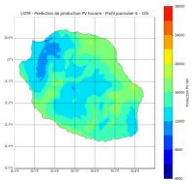
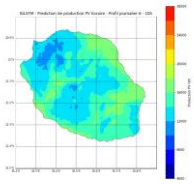
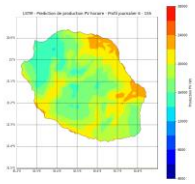
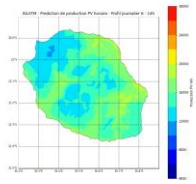
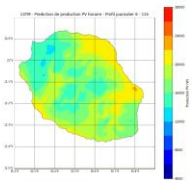
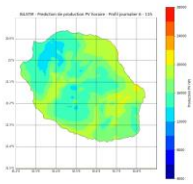
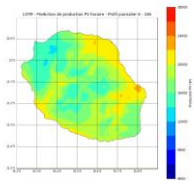
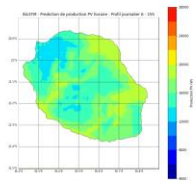
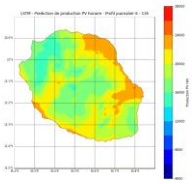
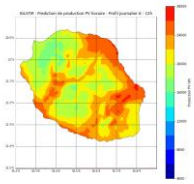
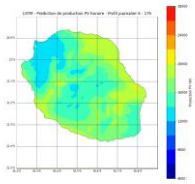
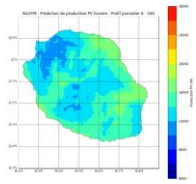
Heure	LSTM	Bi-LSTM	Heure	LSTM	Bi-LSTM
8h			13h		
9h			14h		
10h			15h		
11h			16h		
12h			17h		

Figure A.20 Cartes de prévision de production PV horaires sur une journée type de classe journalière 6 en 2018 pour les modèles LSTM et Bi-LSTM (de gauche à droite). L'échelle des couleurs de la production PV, comprise entre 4 kW et 28 kW

Références

Al- Sabounchi, A. M. (1998). Effect of ambient temperature on the demanded energy of solar cells at different inclinations. *Renewable Energy*, *14*(1), 149–155. https://doi.org/10.1016/S0960-1481(98)00061-5

Albioma. (2020). *Communiqué de presse , "Feu vert pour la conversion au 100 % biomasse de la centrale Albioma Bois-Rouge: Un pas décisif pour l'avenir énergétique de La Réunion."*

Al-Dahidi, S., Ayadi, O., Alrbai, M., & Adeeb, J. (2019). Ensemble Approach of Optimized Artificial Neural Networks for Solar Photovoltaic Power Prediction. *IEEE Access*, *7*, 81741–81758. https://doi.org/10.1109/ACCESS.2019.2923905

Al-Ezzi, A. S., & Ansari, M. N. M. (2022). Photovoltaic Solar Cells: A Review. *Applied System Innovation*, *5*(4), Article 4. https://doi.org/10.3390/asi5040067

Almosni, S., Delamarre, A., Jehl, Z., Suchet, D., Cojocaru, L., Giteau, M., Behaghel, B., Julian, A., Ibrahim, C., Tatry, L., Wang, H., Kubo, T., Uchida, S., Segawa, H., Miyashita, N., Tamaki, R., Shoji, Y., Yoshida, K., Ahsan, N., … Guillemoles, J.-F. (2018). Material challenges for solar cells in the twenty-first century: Directions in emerging technologies. *Science and Technology of Advanced Materials*, *19*(1), 336–369. https://doi.org/10.1080/14686996.2018.1433439

AlSkaif, T., Dev, S., Visser, L., Hossari, M., & van Sark, W. (2020). A systematic analysis of meteorological variables for PV output power estimation. *Renewable Energy*, *153*, 12–22. https://doi.org/10.1016/j.renene.2020.01.150

Amajama, J., Ogbulezie, J., Akonjom, N., & Onuabuchi, V. (2020). Impact of Wind on the Output of Photovoltaic panel and Solar Illuminance/Intensity. *International Journal Of Engineering Research and General Science*, *4*.

Amari, T., Luciani, J.-F., & Aly, J.-J. (2015). Small-scale dynamo magnetism as the driver for heating the solar atmosphere. *Nature*, *522*(7555), Article 7555. https://doi.org/10.1038/nature14478

Andrei, D., & Andrei, L. (2015). Vector Error Correction Model in Explaining the Association of Some Macroeconomic Variables in Romania. *Procedia Economics and Finance*, *22*, 568–576. https://doi.org/10.1016/S2212-5671(15)00261-0

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de-Pison, F. J., & Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. *Solar Energy*, *136*, 78–111. https://doi.org/10.1016/j.solener.2016.06.069

Arbizu-Barrena, C., Ruiz-Arias, J. A., Rodríguez-Benítez, F. J., Pozo-Vázquez, D., & Tovar-Pescador, J. (2017). Short-term solar radiation forecasting by advecting and diffusing MSG cloud index. *Solar Energy*, *155*, 1092–1103. https://doi.org/10.1016/j.solener.2017.07.045

ASTMG173-03. (2012). *Standard Tables for Reference Solar Spectral Irradiances: Direct Normal and Hemispherical on 37 Degree Tilted Surface*.

Azimi, R., Ghayekhloo, M., & Ghofrani, M. (2016). A hybrid method based on a new clustering technique and multilayer perceptron neural networks for hourly solar radiation forecasting. *Energy Conversion and Management*, *118*, 331–344. https://doi.org/10.1016/j.enconman.2016.04.009

Bacher, P., Madsen, H., & Nielsen, H. (2011). Online Short-term Solar Power Forecasting. In *Solar Energy* (Vol. 83). https://doi.org/10.1016/j.solener.2009.05.016

Badosa, J., Haeffelin, M., & Chepfer, H. (2013). Scales of spatial and temporal variation of solar irradiance on Reunion tropical island. *Solar Energy*, *88*, 42–56. https://doi.org/10.1016/j.solener.2012.11.007

Badosa, J., Haeffelin, M., Kalecinski, N., Bonnardot, F., & Jumaux, G. (2015). Reliability of day-ahead solar irradiance forecasts on Reunion Island depending on synoptic wind and humidity conditions. *Solar Energy*, *115*, 306–321. https://doi.org/10.1016/j.solener.2015.02.039

Bai, Y., Xie, J., Wang, D., Zhang, W., & Li, C. (2021). A manufacturing quality prediction model based on AdaBoost-LSTM with rough knowledge. *Computers & Industrial Engineering*, *155*, 107227. https://doi.org/10.1016/j.cie.2021.107227

Baldy, S., Ancellet, G., Bessafi, M., Badr, A., & Luk, D. L. S. (1996). Field observations of the vertical distribution of tropospheric ozone at the island of Reunion (southern tropics). *Journal of Geophysical Research*, *101*, 23,835-23,849. https://doi.org/10.1029/95JD02929

Barbieri, F., Rajakaruna, S., & Ghosh, A. (2017). Very short-term photovoltaic power forecasting with cloud modeling: A review. *Renewable and Sustainable Energy Reviews*, *75*, 242–263. https://doi.org/10.1016/j.rser.2016.10.068

Barman, P. P., & Boruah, A. (2018). A RNN based Approach for next word prediction in Assamese Phonetic Transcription. *Procedia Computer Science*, *143*, 117–123. https://doi.org/10.1016/j.procs.2018.10.359

Bashir, Z. A., & El-Hawary, M. E. (2007). Short-Term Load Forecasting Using Artificial Neural Network Based on Particle Swarm Optimization Algorithm. *2007 Canadian Conference on Electrical and Computer Engineering*, 272–275. https://doi.org/10.1109/CCECE.2007.74

Benmouiza, K., & Cheknane, A. (2013). Forecasting hourly global solar radiation using hybrid k-means and nonlinear autoregressive neural network models. *Energy Conversion and Management*, *75*, 561–569. https://doi.org/10.1016/j.enconman.2013.07.003

Bessa, R., Trindade, A., Silva, C., & Miranda, V. (2015). Solar power forecasting in smart grids using distributed information. *International Journal of Electrical Power & Energy Systems*, *72*. https://doi.org/10.1016/j.ijepes.2015.02.006

Bessafi, M., Carvalho, F., Charton, P., Delsaut, M., Despeyroux, T., Jeanty, P., Jean Daniel, L. S. L., Lechevallier, Y., Ralambondrainy, H., & Trovalet, L. (2015). *Clustering of Solar Irradiance.: In: Lausen B., Krolak-Schwerdt S., Böhmer M. (eds) Data Science, Learning by Latent Structures, and Knowledge Discovery. Studies in Classification, Data Analysis, and Knowledge Organization*.

Bessafi, M., Oree, V., Khoodaruth, A., Jumaux, G., Bonnardot, F., Jeanty, P., Delsaut, M., Chabriat, J.-P., & Dauhoo, M. Z. (2018). Downscaling solar irradiance using DEM-based model in young volcanic islands with rugged topography. *Renewable Energy*, *126*, 584–593. https://doi.org/10.1016/j.renene.2018.03.071

Bin Shams, M., Haji, S., Salman, A., Abdali, H., & Alsaffar, A. (2016). Time series analysis of Bahrain's first hybrid renewable energy system. *Energy*, *103*, 1–15. https://doi.org/10.1016/j.energy.2016.02.136

Bouzerdoum, M., Mellit, A., & Massi Pavan, A. (2013). A hybrid model (SARIMA–SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant. *Solar Energy*, *98*, 226–235. https://doi.org/10.1016/j.solener.2013.10.002

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

British Petroleum. (2022). *Statistical Review of World Energy 2022*.

Britt, J., & Ferekides, C. (1993). Thin-film CdS/CdTe solar cell with 15.8% efficiency. *Applied Physics Letters*, *62*(22), 2851–2852. https://doi.org/10.1063/1.109629

Buber, E., & Diri, B. (2019). Web Page Classification Using RNN. *Procedia Computer Science*, *154*, 62–72. https://doi.org/10.1016/j.procs.2019.06.011

Casola, & Wallace. (2007). *Identifying Weather Regimes in the Wintertime 500-hPa Geopotential Height Field for the Pacific–North American Sector Using a Limited-Contour Clustering Technique in: Journal of Applied Meteorology and Climatology Volume 46 Issue 10 (2007)*. https://journals.ametsoc.org/view/journals/apme/46/10/jam2564.1.xml

Chandra, S., Agrawal, S., & Chauhan, D. S. (2018). Effect of Ambient Temperature and Wind Speed on Performance Ratio of Polycrystalline Solar Photovoltaic Module: An Experimental Analysis. *International Energy Journal*, *18*(2), Article 2. http://www.rericjournal.ait.ac.th/index.php/reric/article/view/1698

Cheng, X., & Wallace, J. M. (1993). Cluster Analysis of the Northern Hemisphere Wintertime 500-hPa Height Field: Spatial Patterns. *Journal of the Atmospheric Sciences*, *50*(16), 2674–2696. https://doi.org/10.1175/1520-0469(1993)050<2674:CAOTNH>2.0.CO;2

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. https://doi.org/10.3115/v1/D14-1179

Choi, H., Cho, K., & Bengio, Y. (2018). Fine-grained attention mechanism for neural machine translation. *Neurocomputing*, *284*, 171–176. https://doi.org/10.1016/j.neucom.2018.01.007

Chu, Y., Urquhart, B., Gohari, S. M. I., Pedro, H., Kleissl, J., & Coimbra, C. (2015). Short-term reforecasting of power output from a 48 MWe solar PV plant. *Solar Energy*, *112*, 68–77. https://doi.org/10.1016/j.solener.2014.11.017

COI. (2017). *Rapport Annuel COI 2017* (p. 118) [Rapport annuel]. Commission de l'Océan Indien.

Crétat, J., Macron, C., Pohl, B., & Richard, Y. (2011). Quantifying internal variability in a regional climate model: A case study for Southern Africa. *Climate Dynamics*, *37*(7–8), 1335–1356. https://doi.org/10.1007/s00382-011-1021-5

Crétat, J., & Pohl, B. (2012). How Physical Parameterizations Can Modulate Internal Variability in a Regional Climate Model. *Journal of the Atmospheric Sciences*, *69*(2), 714–724. https://doi.org/10.1175/JAS-D-11-0109.1

Cretat, J., Pohl, B., & Richard, Y. (2011). *Les modèles climatiques régionaux: Outils de décomposition des échelles spatio-temporelles*. 11.

Crétat, J., Pohl, B., Richard, Y., & Drobinski, P. (2012). Uncertainties in simulating regional climate of Southern Africa: Sensitivity to physical parameterizations using WRF. *Climate Dynamics*, *38*(3–4), 613–634. https://doi.org/10.1007/s00382-011-1055-8

Cuthbertson, K., Hall, S. G., & Taylor, M. P. (1992). *Applied Econometric Techniques*. Philip Allan. https://books.google.com/books?id=iha7AAAAIAAJ

Daniel, M. (2017). *Villes, climat urbain et climat régional sur la France: Étude par une approche de modélisation climatique couplée* [Phdthesis, Université Paul Sabatier - Toulouse III]. https://theses.hal.science/tel-01955973

Das, U. K., Tey, K. S., Seyedmahmoudian, M., Mekhilef, S., Idris, M. Y. I., Van Deventer, W., Horan, B., & Stojcevski, A. (2018). Forecasting of photovoltaic power generation and model optimization: A review. *Renewable and Sustainable Energy Reviews*, *81*, 912–928. https://doi.org/10.1016/j.rser.2017.08.017

Datas, A., & Martí, A. (2017). Thermophotovoltaic energy in space applications: Review and future potential. *Solar Energy Materials and Solar Cells*, *161*, 285–296. https://doi.org/10.1016/j.solmat.2016.12.007

Davidson, R., & MacKinnon, J. G. (2009). *Econometric Theory and Methods*.

De Meij, A., Vinuesa, J.-F., & Maupas, V. (2018). GHI calculation sensitivity on microphysics, land- and cumulus parameterization in WRF over the Reunion Island. *Atmospheric Research*, *204*, 12–20. https://doi.org/10.1016/j.atmosres.2018.01.008

Diagne, H. M. (2015). *Gestion intelligente du réseau électrique réunionnais. Prévision de la ressource solaire en milieu insulaire*. 153.

Diagne, M., David, M., Boland, J., Schmutz, N., & Lauret, P. (2014). Post-processing of Solar Irradiance Forecasts from WRF Model at Reunion Island. *Energy Procedia*, *57*, 1364–1373. https://doi.org/10.1016/j.egypro.2014.10.127

Diagne, M., David, M., Lauret, P., Boland, J., & Schmutz, N. (2013). Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renewable and Sustainable Energy Reviews*, *27*, 65–76. https://doi.org/10.1016/j.rser.2013.06.042

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, *74*(366a), 427–431. https://doi.org/10.1080/01621459.1979.10482531

Diebold, F. X. (2001). *Elements of Forecasting* (2nd ed.). Southern Western Publishing. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.571.5408&rep=rep1&type=pdf

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In G. Goos, J. Hartmanis, & J. van Leeuwen (Eds.), *Multiple Classifier Systems* (Vol. 1857, pp. 1–15). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

Directive européenne 2009/28/EC. (2009). Directive 2009/28/EC of 23 April 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC. In Directive européenne 2009/28/EC, *Core Statutes on Company Law* (pp. 757–759). https://doi.org/10.1007/978-1-137-54507-7_21

Dozat, T. (2015). *Incorporating Nesterov Momentum into Adam*. 6.

Dubey, S., Sarvaiya, J., & Seshadri, B. (2013). Temperature Dependent Photovoltaic (PV) Efficiency and Its Effect on PV Production in the World – A Review. *Energy Procedia*, *33*, 311–321. https://doi.org/10.1016/j.egypro.2013.05.072

Dudhia, J. (1988). Numerical Study of Convection Observed during the Winter Monsoon Experiment Using a Mesoscale Two-Dimensional Model. *Journal of the Atmospheric Sciences*, *46*(20), 3077–3107. https://doi.org/10.1175/1520-0469(1989)046<3077:NSOCOD>2.0.CO;2

Dufresne, J.-L., & Salas y Mélia, D. (2017). 3. Besoins en modélisation numérique. In C. Jeandel & R. Mosseri (Eds.), *Le climat à découvert* (pp. 249–250). CNRS Éditions. https://doi.org/10.4000/books.editionscnrs.11502

Duncan, R. A. (1981). Hotspots in the Southern Oceans—An absolute frame of reference for motion of the Gondwana continents. *Tectonophysics*, *74*(1), 29–42. https://doi.org/10.1016/0040-1951(81)90126-8

Dutta, S., Li, Y., Venkataraman, A., Costa, L. M., Jiang, T., Plana, R., Tordjman, P., Choo, F. H., Foo, C. F., & Puttgen, H. B. (2017). Load and Renewable Energy Forecasting for a Microgrid using Persistence Technique. *Energy Procedia*, *143*, 617–622. https://doi.org/10.1016/j.egypro.2017.12.736

El-Baz, W., Tzscheutschler, P., & Wagner, U. (2018). Day-ahead probabilistic PV generation forecast for buildings energy management systems. *Solar Energy*, *171*, 478–490. https://doi.org/10.1016/j.solener.2018.06.100

El-Damak, D., & Chandrakasan, A. (2015). *Solar energy harvesting system with integrated battery management and startup using single inductor and 3.2nW quiescent power*. C280–C281. https://doi.org/10.1109/VLSIC.2015.7231290

Enders, W. (1995). *Applied Econometric Time Series*. https://www.wiley.com/en-us/Applied+Econometric+Time+Series%2C+4th+Edition-p-9781118808566

Engle, R., & Granger, C. (1991). *Long-Run Economic Relationships: Readings in Cointegration* [OUP Catalogue]. Oxford University Press. https://econpapers.repec.org/bookchap/oxpobooks/9780198283393.htm

Fanchette, Y., Ramenah, H., Casin, P., Benne, M., Tanougast, C., & Adjallah, K. (2019). Predictive Causality of Granger Test for Long Run Equilibrium to Photovoltaic System. *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 942–946. https://doi.org/10.1109/IDAACS.2019.8924303

Fanchette, Y., Ramenah, H., Tanougast, C., Benne, M., 1 LE^{2}P—Energy-Lab, University of Reunion Island, 97744 Saint-Denis, France, & 2 LCOMS, University of Lorraine, 57070 Metz, France. (2020). Applying Johansen VECM cointegration approach to propose a forecast model of photovoltaic power output plant in Reunion Island. *AIMS Energy*, *8*(2), 179–213. https://doi.org/10.3934/energy.2020.2.179

Fu, T., Tang, X., Cai, Z., Zuo, Y., Tang, Y., & Zhao, X. (2020). Correlation research of phase angle variation and coating performance by means of Pearson's correlation coefficient. *Progress in Organic Coatings*, *139*, 105459. https://doi.org/10.1016/j.porgcoat.2019.105459

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202. https://doi.org/10.1007/BF00344251

Gao, M., Li, J., Hong, F., & Long, D. (2019). Day-ahead power forecasting in a large-scale photovoltaic plant based on weather classification using LSTM. *Energy*, *187*, 115838. https://doi.org/10.1016/j.energy.2019.07.168

Gaviria, J. F., Narváez, G., Guillen, C., Giraldo, L. F., & Bressan, M. (2022). Machine learning in photovoltaic systems: A review. *Renewable Energy*, *196*, 298–318. https://doi.org/10.1016/j.renene.2022.06.105

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

Gill, H. S., & Khehra, B. S. (2021). An integrated approach using CNN-RNN-LSTM for classification of fruit images. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.06.016

Giorgi, F. (2019). Thirty Years of Regional Climate Modeling: Where Are We and Where Are We Going next? *Journal of Geophysical Research: Atmospheres*, 2018JD030094. https://doi.org/10.1029/2018JD030094

Goldani, M. H., Momtazi, S., & Safabakhsh, R. (2021). Detecting fake news with capsule neural networks. *Applied Soft Computing*, *101*, 106991. https://doi.org/10.1016/j.asoc.2020.106991

Govender, P. (2017). *Clustering analysis for classification and forecasting of solar irradiance in Durban, South Africa*. 162.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, *37*(3), 424–438. https://doi.org/10.2307/1912791

Granger, C. W. J., Hyung, N., & Jeon, Y. (2001). Spurious regressions with stationary series. *Applied Economics*, *33*(7), 899–904. https://doi.org/10.1080/00036840121734

Granger, C. W. J., & Weiss, A. A. (1983). TIME SERIES ANALYSIS OF ERROR-CORRECTION MODELS. In S. Karlin, T. Amemiya, & L. A. Goodman (Eds.),

Studies in Econometrics, Time Series, and Multivariate Statistics (pp. 255–278). Academic Press. https://doi.org/10.1016/B978-0-12-398750-1.50018-8

Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks* (Vol. 385). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-24797-2

Green, M. A., Emery, K., King, D. L., Igari, S., & Warta, W. (2005). SHORT COMMUNICATION: Solar cell efficiency tables (version 25). *Progress in Photovoltaics: Research and Applications*, *13*(1), 49–54. https://doi.org/10.1002/pip.598

Gujarati, D. (2004). *Basic Econometrics*. http://dspace.agu.edu.vn:8080/handle/AGU_Library/3789

Hamisultane, H. (2002). *MODELE A CORRECTION D'ERREUR (MCE) ET APPPLICATIONS*.

Hamzah, A. H., & Go, Y. I. (2023). Design and assessment of building integrated PV (BIPV) system towards net zero energy building for tropical climate. *E-Prime - Advances in Electrical Engineering, Electronics and Energy*, *3*, 100105. https://doi.org/10.1016/j.prime.2022.100105

Hassan, M. A., Bailek, N., Bouchouicha, K., & Nwokolo, S. C. (2021). Ultra-short-term exogenous forecasting of photovoltaic power production using genetically optimized non-linear auto-regressive recurrent neural networks. *Renewable Energy*, *171*, 191–209. https://doi.org/10.1016/j.renene.2021.02.103

Hassani, H., & Yeganegi, M. R. (2019). Sum of squared ACF and the Ljung–Box statistics. *Physica A: Statistical Mechanics and Its Applications*, *520*, 81–86. https://doi.org/10.1016/j.physa.2018.12.028

Hertz, J., Krogh, A., Palmer, R. G., & Horner, H. (1991). Introduction to the Theory of Neural Computation. *Physics Today*, *44*, 70. https://doi.org/10.1063/1.2810360

Hinton, G., Srivastava, N., & Swersky, K. (2012). Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited On*, *14*(8), 2.

Hochreiter, B., Frasconi, Schmidhuber, & Bengio. (2009). Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies. In J. F. Kolen, *A Field Guide to Dynamical Recurrent Networks*. IEEE. https://doi.org/10.1109/9780470544037.ch14

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoffman, J. I. E. (2015). Chapter 6—Normal Distribution. In J. I. E. Hoffman (Ed.), *Biostatistics for Medical and Biomedical Practitioners* (pp. 101–119). Academic Press. https://doi.org/10.1016/B978-0-12-802387-7.00006-8

Hong, S.-Y., Noh, Y., & Dudhia, J. (2006). A New Vertical Diffusion Package with an Explicit Treatment of Entrainment Processes. *Monthly Weather Review*, *134*(9), 2318–2341. https://doi.org/10.1175/MWR3199.1

Hornik, K. (1989). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, *Vol. 2*, 359–366.

Huang, J., Korolkiewicz, M., Agrawal, M., & Boland, J. (2013). Forecasting solar radiation on an hourly time scale using a Coupled AutoRegressive and Dynamical System (CARDS) model. *Solar Energy*, *87*, 136–149. https://doi.org/10.1016/j.solener.2012.10.012

Huang, R., Wei, C., Wang, B., Yang, J., Xu, X., Wu, S., & Huang, S. (2022). Well performance prediction based on Long Short-Term Memory (LSTM) neural network. *Journal of Petroleum Science and Engineering*, *208*, 109686. https://doi.org/10.1016/j.petrol.2021.109686

Hüsken, M., & Stagge, P. (2003). Recurrent neural networks for time series classification. *Neurocomputing*, *50*, 223–235. https://doi.org/10.1016/S0925-2312(01)00706-8

Inman, R. H., Pedro, H. T. C., & Coimbra, C. F. M. (2013). Solar forecasting methods for renewable energy integration. *Progress in Energy and Combustion Science*, *39*(6), 535–576. https://doi.org/10.1016/j.pecs.2013.06.002

Insee. (2023). *Part des énergies renouvelables dans l'Union européenne*. https://www.insee.fr/fr/statistiques/4318263

International Energy Agency. (2022). *Renewables 2022 Analysis and forecast to 2027*.

IRENA. (2022). *RENEWABLE CAPACITY STATISTICS 2022*.

Jaeger, H. (n.d.). *Adaptive Nonlinear System Identification with Echo State Networks*. 8.

Jalil, A., & Rao, N. H. (2019). Chapter 8—Time Series Analysis (Stationarity, Cointegration, and Causality). In B. Özcan & I. Öztürk (Eds.), *Environmental Kuznets Curve (EKC)* (pp. 85–99). Academic Press. https://doi.org/10.1016/B978-0-12-816797-7.00008-4

Jayawardena, K. D. G. I., Rozanski, L. J., Mills, C. A., Beliatis, M. J., Nismy, N. A., & Silva, S. R. P. (2013). 'Inorganics-in-Organics': Recent developments and outlook for 4G polymer solar cells. *Nanoscale*, *5*(18), 8411. https://doi.org/10.1039/c3nr02733c

Jha, K., & Shaik, A. G. (2023). A comprehensive review of power quality mitigation in the scenario of solar PV integration into utility grid. *E-Prime - Advances in Electrical Engineering, Electronics and Energy*, *3*, 100103. https://doi.org/10.1016/j.prime.2022.100103

Jiang, X., Pang, Y., Sun, M., & Li, X. (2018). Cascaded Subpatch Networks for Effective CNNs. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(7), 2684–2694. https://doi.org/10.1109/TNNLS.2017.2689098

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, *12*(2), 231–254. https://doi.org/10.1016/0165-1889(88)90041-3

Johansen, S. (1991). Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica*, *59*(6), 1551–1580. https://doi.org/10.2307/2938278

Johansen, S. (2015). Time Series: Cointegration. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)* (pp. 322–330). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.42086-6

Joseph, A., & Oku, D. E. (2016). *Wind versus UHF Radio signal*. *5*(2).

Jumaux, G., Quetelard, H., & Roy. (2011). *Atlas climatique de la Réunion*. METEO FRANCE. Sainte-Clotilde. https://side.developpement-durable.gouv.fr/Default/doc/SYRACUSE/214962/atlas-climatique-de-la-reunion

Jung Xu. (2012). *Arrays of CdSe sensitized ZnO/ZnSe nanocables for efficient solar cells with high open-circuit voltage*. https://pubs.rsc.org/en/content/articlelanding/2012/JM/c2jm31970e

Kain, J. S. (2004). The Kain–Fritsch Convective Parameterization: An Update. *Journal of Applied Meteorology*, *43*(1), 170–181. https://doi.org/10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2

Kaldellis, J., Kapsali, M., & Kavadias, K. (2014). Temperature and wind speed impact on the efficiency of PV installations. Experience obtained from outdoor measurements in Greece. *Renewable Energy*, *66*, 612–624. https://doi.org/10.1016/j.renene.2013.12.041

Kamadinata, J. O., Ken, T. L., & Suwa, T. (2019). Sky image-based solar irradiance prediction methodologies using artificial neural networks. *Renewable Energy*, *134*, 837–845. https://doi.org/10.1016/j.renene.2018.11.056

Katircioglu, S. T. (2009). Revisiting the tourism-led-growth hypothesis for Turkey using the bounds test and Johansen approach for cointegration. *Tourism Management*, *30*(1), 17–20. https://doi.org/10.1016/j.tourman.2008.04.004

Kavlak, G., McNerney, J., & Trancik, J. E. (2018). Evaluating the causes of cost reduction in photovoltaic modules. *Energy Policy*, *123*, 700–710. https://doi.org/10.1016/j.enpol.2018.08.015

Keis, K., Magnusson, E., Lindstro, H., Lindquist, S.-E., & Hagfeldt, A. (2002). A 5% efficient photoelectrochemical solar cell based on nanostructured ZnO electrodes. *Solar Energy Materials*.

Kennewell, J., & McDonald, A. (2015). *The solar constant*.

Ketjoy, N., & Konyu, M. (2014). Study of Dust Effect on Photovoltaic Module for Photovoltaic Power Plant. *Energy Procedia*, *52*. https://doi.org/10.1016/j.egypro.2014.07.095

Kingma, D. P., & Ba, J. (2017). *Adam: A Method for Stochastic Optimization* (arXiv:1412.6980). arXiv. https://doi.org/10.48550/arXiv.1412.6980

Kitamura, Y. (1998). LIKELIHOOD-BASED INFERENCE IN COINTEGRATED VECTOR AUTOREGRESSIVE MODELS: By Søren Johansen, Oxford University Press, 1995. *Econometric Theory*, *14*(4), 517–524. https://doi.org/10.1017/S0266466698144067

Koop, G. (2005). *Analysis of economic data*. John Wiley & Sons.

Kostylev, V., & Pavlovski, A. (n.d.). *Solar Power Forecasting Performance – Towards Industry Standards*.

Krause, B., Murray, I., Renals, S., & Lu, L. (2017). *MULTIPLICATIVE LSTM FOR SEQUENCE MODELLING*. 9.

Kumar Dubey, A., Kumar, A., García-Díaz, V., Kumar Sharma, A., & Kanhaiya, K. (2021). Study and analysis of SARIMA and LSTM in forecasting time series data. *Sustainable Energy Technologies and Assessments*, *47*, 101474. https://doi.org/10.1016/j.seta.2021.101474

Kumari, P., & Toshniwal, D. (2021). Deep learning models for solar irradiance forecasting: A comprehensive review. *Journal of Cleaner Production*, *318*, 128566. https://doi.org/10.1016/j.jclepro.2021.128566

Kumler, A., Xie, Y., & Zhang, Y. (2019). *A Physics-based Smart Persistence model for Intra-hour forecasting of solar radiation (PSPI) using GHI measurements and a cloud retrieval technique*. https://doi.org/10.1016/j.solener.2018.11.046

Kushwaha, V., & Pindoriya, N. M. (2017). Very short-term solar PV generation forecast using SARIMA model: A case study. *2017 7th International Conference on Power Systems (ICPS)*, 430–435. https://doi.org/10.1109/ICPES.2017.8387332

Kuzmiakova, A. (2017). *Short-term Memory Solar Energy Forecasting at University of Illinois*. 6.

Labouret, A., & Villoz, M. (2006). *Energie solaire photovoltaïque* (Vol. 3). Dunod Paris.

Lakatos, L., Hevessy, G., & Kovacs, J. (2011). *Advantages and Disadvantages of Solar Energy and Wind-Power Utilization*. https://www.tandfonline.com/doi/full/10.1080/02604020903021776

Laronde, R., Charki, A., & David, B. (2010). Reliability of photovoltaic modules based on climatic measurement data. In *International Journal of Metrology and Quality Engeneering* (Vol. 1). https://doi.org/10.1051/ijmqe/2010012

Le, Q. V., Jaitly, N., & Hinton, G. E. (2015). A Simple Way to Initialize Recurrent Networks of Rectified Linear Units. *ArXiv:1504.00941 [Cs]*. http://arxiv.org/abs/1504.00941

Lecun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time-series. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks*. MIT Press.

Lénat, J.-F., Bachèlery, P., & Merle, O. (2012). Anatomy of Piton de la Fournaise volcano (La Réunion, Indian Ocean). *Bulletin of Volcanology*, *74*(9), 1945–1961. https://doi.org/10.1007/s00445-012-0640-y

Lesouëf, D. (2010). *Étude numérique des circulations locales à la Réunion: Application à la dispersion de polluants*.

Lesouëf, D., Gheusi, F., Delmas, R., & Escobar, J. (2011). Numerical simulations of local circulations and pollution transport over Reunion Island. *Annales Geophysicae*, *29*(1), 53–69. https://doi.org/10.5194/angeo-29-53-2011

Li, L.-L., Wen, S.-Y., Tseng, M.-L., & Wang, C.-S. (2019). Renewable energy prediction: A novel short-term prediction model of photovoltaic output power. *Journal of Cleaner Production*, *228*, 359–375. https://doi.org/10.1016/j.jclepro.2019.04.331

li, P., morel, B., bessafi, M., solmon, F., & Chiacchio, M. (2013). *The radiation budget in the regional climate model RegCM4: Simulation results from two different radiative schemes over the south-western Indian Ocean*. EGU2013-723. https://ui.adsabs.harvard.edu/abs/2013EGUGA..15..723L

Li, Q., Miloud, B., Delage, O., Chabriat, J.-P., & Li, P. (2015, May 11). *INTERMITTENCY STUDY OF GLOBAL SOLAR RADIATION ON REUNION ISLAND USING HILBERT-HUANG TRANSFORM*.

Li, W., Guo, D., & Fang, X. (2018). Multimodal architecture for video captioning with memory networks and an attention mechanism. *Pattern Recognition Letters*, *105*, 23–29. https://doi.org/10.1016/j.patrec.2017.10.012

li, Y., Su, Y., & Shu, L. (2014). An ARMAX model for forecasting the power output of a grid connected photovoltaic system. *Renewable Energy*, *66*, 78–89. https://doi.org/10.1016/j.renene.2013.11.067

Liang, T., Zhao, Q., Lv, Q., & Sun, H. (2021). A novel wind speed prediction strategy based on Bi-LSTM, MOOFADA and transfer learning for centralized control centers. *Energy*, *230*, 120904. https://doi.org/10.1016/j.energy.2021.120904

Lin-Kwong-Chon, C. (n.d.). *Approches neuronales adaptatives pour le contrôle tolérant aux défauts de systèmes pile à combustible*. 210.

Lin-Kwong-Chon, C. (2020). *Approches neuronales adaptatives pour le contrôle tolérant aux défauts de systèmes pile à combustible*. 210.

Liu, D., & Sun, K. (2019). Random forest solar power forecast based on classification optimization. *Energy*, *187*, 115940. https://doi.org/10.1016/j.energy.2019.115940

Liu, Y., Gong, C., Yang, L., & Chen, Y. (2020). DSTP-RNN: A dual-stage two-phase attention-based recurrent neural network for long-term and multivariate time series prediction. *Expert Systems with Applications*, *143*, 113082. https://doi.org/10.1016/j.eswa.2019.113082

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303. https://doi.org/10.1093/biomet/65.2.297

Lo Brano, V., Ciulla, G., & Di Falco, M. (2014). Artificial Neural Networks to Predict the Power Output of a PV Panel. *International Journal of Photoenergy*, *2014*, 193083. https://doi.org/10.1155/2014/193083

Lorenz, E., Scheidsteger, T., Hurka, J., Heinemann, D., & Kurz, C. (2011). Regional PV power prediction for improved grid integration. *Progress in Photovoltaics: Research and Applications*, *19*(7), 757–771. https://doi.org/10.1002/pip.1033

Luo, Y., Zhang, L., Liu, Z., Wu, J., Zhang, Y., Zhenghong, W., & He, X. H. (2017). Performance analysis of a self-adaptive building integrated photovoltaic thermoelectric wall system in hot summer and cold winter zone of China. *Energy*, *140*. https://doi.org/10.1016/j.energy.2017.09.015

Ma, M., Tang, J., Ou, T., & Zhou, P. (2023). High-resolution climate projection over the Tibetan Plateau using WRF forced by bias-corrected CESM. *Atmospheric Research*, *286*, 106670. https://doi.org/10.1016/j.atmosres.2023.106670

Magnone, L., Sossan, F., Scolari, E., & Paolone, M. (2017). Cloud Motion Identification Algorithms Based on All-Sky Images to Support Solar Irradiance Forecast. *2017 IEEE*

44th Photovoltaic Specialist Conference (PVSC), 1415–1420. https://doi.org/10.1109/PVSC.2017.8366102

Marcinkiewicz, E. (2014). Some Aspects of Application of VECM Analysis for Modeling Causal Relationships Between Spot and Futures Prices. *Optimum. Studia Ekonomiczne*, *5*, 114–125. https://doi.org/10.15290/ose.2014.05.71.09

Maresca, B., & Dujin, A. (2014). La transition énergétique à l'épreuve du mode de vie. *Flux*, *96*(2), 10–23. https://doi.org/10.3917/flux.096.0010

Mathijsen, D. (2021). The role of composites in getting the solar car to our driveways: Lightyear one. *Reinforced Plastics*, *65*(4), 178–187. https://doi.org/10.1016/j.repl.2021.06.001

Mellit, A., & Kalogirou, S. A. (2008). Artificial intelligence techniques for photovoltaic applications: A review. *Progress in Energy and Combustion Science*, *34*(5), 574–632. https://doi.org/10.1016/j.pecs.2008.01.001

Mellit, A., Pavan, A. M., & Lughi, V. (2021). Deep learning neural networks for short-term photovoltaic power forecasting. *Renewable Energy*, *172*, 276–288. https://doi.org/10.1016/j.renene.2021.02.166

Mialhe, P. (2018). *Variabilité spatiale et temporelle du rayonnement solaire global sur une topographie à relief marqué et complexe. Cas de l'île de La Réunion*. 253.

Mialhe, P., Pohl, B., Morel, B., Trentmann, J., Jumaux, G., Bonnardot, F., Bessafi, M., & Chabriat, J.-P. (2020). On the determination of coherent solar climates over a tropical island with a complex topography. *Solar Energy*, *206*, 508–521. https://doi.org/10.1016/j.solener.2020.04.049

Mills, T. C. (2019). Chapter 14—Error Correction, Spurious Regressions, and Cointegration. In T. C. Mills (Ed.), *Applied Time Series Analysis* (pp. 233–253). Academic Press. https://doi.org/10.1016/B978-0-12-813117-6.00014-4

Ministère de la Transition écologique. (2015). *Loi de transition énergétique pour la croissance verte*. Ministère de la Transition écologique. https://www.ecologie.gouv.fr/loi-transition-energetique-croissance-verte

Ministère de la Transition écologique. (2023). Stratégie française pour l'énergie et le climat—Programmation pluriannuelle de l'énergie. *2023*. https://www.ecologie.gouv.fr/programmations-pluriannuelles-lenergie-ppe

Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J., & Clough, S. A. (1997). Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *Journal of Geophysical Research: Atmospheres*, *102*(D14), 16663–16682. https://doi.org/10.1029/97JD00237

Monteiro, C., Santos, T., Fernandez-Jimenez, L. A., Ramirez-Rosado, I. J., & Terreros-Olarte, M. S. (2013). Short-Term Power Forecasting Model for Photovoltaic Plants Based on Historical Similarity. *Energies*, *6*(5), 2624–2643. https://doi.org/10.3390/en6052624

Morel, B., Pohl, B., Richard, Y., Bois, B., & Bessafi, M. (2014). Regionalizing Rainfall at Very High Resolution over La Réunion Island Using a Regional Climate Model. *Monthly Weather Review*, *142*(8), 2665–2686. https://doi.org/10.1175/MWR-D-14-00009.1

Muzaffar, S., & Afshari, A. (2019). Short-Term Load Forecasts Using LSTM Networks. *Energy Procedia*, *158*, 2922–2927. https://doi.org/10.1016/j.egypro.2019.01.952

Observatoire Énergie Réunion. (2020). *Bilan énergétique de la Réunion 2019—Édition 2020*.

Observatoire Énergie Réunion. (2022). *Bilan électrique de La Réunion*.

OMM. (2014). *Siting classification for surface observing stations on land. Technical report*.

Omubo-Pepple, V., Israel-Cookey, C., & Alaminiokuma, G. (2009). Effects of Temperature, Solar Flux and Relative Humidity on the Efficient Conversion of Solar Energy to Electricity. *European Journal of Scientific Research*, *35*, 173–180.

Pedro, H. T. C., & Coimbra, C. F. M. (2012). Assessment of forecasting techniques for solar power production with no exogenous inputs. *Solar Energy*, *86*(7), 2017–2028. https://doi.org/10.1016/j.solener.2012.04.004

Pedro, H. T. C., Coimbra, C. F. M., David, M., & Lauret, P. (2018). Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts. *Renewable Energy*, *123*, 191–203. https://doi.org/10.1016/j.renene.2018.02.006

Pelland, S., Galanis, G., & Kallos, G. (2013). Solar and photovoltaic forecasting through post-processing of the Global Environmental Multiscale numerical weather prediction model. *Progress in Photovoltaics: Research and Applications*, *21*(3), 284–296. https://doi.org/10.1002/pip.1180

Perez, R., Kivalov, S., Schlemmer, J., Hemker, K., Renné, D., & Hoff, T. E. (2010). Validation of short and medium term operational solar radiation forecasts in the US. *Solar Energy*, *84*(12), 2161–2172. https://doi.org/10.1016/j.solener.2010.08.014

Perez, R., Moore, K., Wilcox, S., Renné, D., & Zelenka, A. (2007). Forecasting solar radiation – Preliminary evaluation of an approach based upon the national forecast database. *Solar Energy*, *81*(6), 809–812. https://doi.org/10.1016/j.solener.2006.09.009

Peumans, P., Yakimov, A., & Forrest, S. R. (2003). Small molecular weight organic thin-film photodetectors and solar cells. *Journal of Applied Physics*, *93*(7), 3693–3723. https://doi.org/10.1063/1.1534621

Pfeifroth, U., Kothe, S., Trentmann, J., Hollmann, R., Fuchs, P., Kaiser, J., & Werscheck, M. (2019). *Surface Radiation Data Set—Heliosat (SARAH)—Edition 2.1* (2.1, p. 7.5 TiB) [NetCDF-4]. Satellite Application Facility on Climate Monitoring (CM SAF). https://doi.org/10.5676/EUM_SAF_CM/SARAH/V002_01

Pfeifroth, Uwe, Kothe, Steffen, Drücke, Jaqueline, Trentmann, Jörg, Schröder, Marc, Selbach, Nathalie, & Hollmann, Rainer. (2023). *Surface Radiation Data Set—Heliosat (SARAH)—Edition 3* (3.0, p. 12.3 TiB) [NetCDF-4]. Satellite Application Facility on Climate Monitoring (CM SAF). https://doi.org/10.5676/EUM_SAF_CM/SARAH/V003

Pohl, B., Morel, B., Barthe, C., & Bousquet, O. (2016). Regionalizing Rainfall at Very High Resolution over La Réunion Island: A Case Study for Tropical Cyclone Ando. *Monthly Weather Review*, *144*(11), 4081–4099. https://doi.org/10.1175/MWR-D-15-0404.1

Powers, J. G., Klemp, J. B., Skamarock, W. C., Davis, C. A., Dudhia, J., Gill, D. O., Coen, J. L., Gochis, D. J., Ahmadov, R., Peckham, S. E., Grell, G. A., Michalakes, J., Trahan, S., Benjamin, S. G., Alexander, C. R., Dimego, G. J., Wang, W., Schwartz, C. S., Romine, G. S., … Duda, M. G. (2017). The Weather Research and Forecasting Model: Overview, System Efforts, and Future Directions. *Bulletin of the American Meteorological Society*, *98*(8), 1717–1737. https://doi.org/10.1175/BAMS-D-15-00308.1

Praene, J. P., David, M., Sinama, F., Morau, D., & Marc, O. (2012). Renewable energy: Progressing towards a net zero energy island, the case of Reunion Island. *Renewable and Sustainable Energy Reviews*, *16*(1), 426–442. https://doi.org/10.1016/j.rser.2011.08.007

Projet ENERGIES-COI. (2019). *LES ÉNERGIES RENOUVELABLES AU SERVICE DES COMMUNAUTÉS : Zoom sur les projets concrets de valorisation des énergies renouvelables cofi nancés dans le cadre du Programme ENERGIES* (p. 56P.) [Etude]. COI.

Qasem, H., Betts, T., Müllejans, H., AlBusairi, H., & Gottschalg, R. (2014). Dust Induced Shading on Photovoltaic Modules. *Progress in Photovoltaics*, *22*, accepted for publication. https://doi.org/10.1002/pip.2230

Qing, X., & Niu, Y. (2018). Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. *Energy*, *148*, 461–468. https://doi.org/10.1016/j.energy.2018.01.177

Qu, J., Qian, Z., & Pei, Y. (2021). Day-ahead hourly photovoltaic power forecasting using attention-based CNN-LSTM neural network embedded with multiple relevant and target variables prediction pattern. *Energy*, *232*, 120996. https://doi.org/10.1016/j.energy.2021.120996

Rahman, A., Srikumar, V., & Smith, A. D. (2018). Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. *Applied Energy*, *212*, 372–385. https://doi.org/10.1016/j.apenergy.2017.12.051

Rajagukguk, R. A., Ramadhan, R. A. A., & Lee, H.-J. (2020). A Review on Deep Learning Models for Forecasting Time Series Data of Solar Irradiance and Photovoltaic Power. *Energies*, *13*(24), 6623. https://doi.org/10.3390/en13246623

Ramalingam, K., & Indulkar, C. (2017). Solar Energy and Photovoltaic Technology. In *Distributed Generation Systems* (pp. 69–147). Elsevier. https://doi.org/10.1016/B978-0-12-804208-3.00003-0

Ramenah, H., Casin, P., Ba, M., Benne, M., & Tanougast, C. (2017). *Accurate determination of parameters relationship for photovoltaic power output by augmented dickey fuller test and engle granger method*.

Ramenah, H., Khoodaruth, A., Oree, V., Coya, Z., Murdan, A., Miloud, B., & Doseeah, D. (2023). Johansen model for photovoltaic a very short term prediction to electrical power grids in the Island of Mauritius. *Clean Technologies and Recycling*, *3*, 107–118. https://doi.org/10.3934/ctr.2023007

Randall, D. A. (2000). *General Circulation Model Development: Past, Present, and Future*. Elsevier.

Rawat, W., & Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, *29*(9), 2352–2449. https://doi.org/10.1162/neco_a_00990

Raza, M. Q., Nadarajah, M., & Ekanayake, C. (2016). On recent advances in PV output power forecast. *Solar Energy*, *136*, 125–144. https://doi.org/10.1016/j.solener.2016.06.073

Rehman, A. U., Malik, A. K., Raza, B., & Ali, W. (2019). A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis. *Multimedia Tools and Applications*, *78*(18), 26597–26613. https://doi.org/10.1007/s11042-019-07788-7

Reikard, G., & Hansen, C. (2019). Forecasting solar irradiance at short horizons: Frequency and time domain models. *Renewable Energy*, *135*, 1270–1290. https://doi.org/10.1016/j.renene.2018.08.081

Ren, Y., Suganthan, P. N., & Srikanth, N. (2015). Ensemble methods for wind and solar power forecasting—A state-of-the-art review. *Renewable and Sustainable Energy Reviews*, *50*, 82–91. https://doi.org/10.1016/j.rser.2015.04.081

Réseau de Transport d'électricité (RTE). (2023). *Bilan électrique 2022*.

Rifkin, J. (n.d.). *Leading the Way to the Third Industrial Revolution:* 36.

Riihelä, A., Kallio, V., Devraj, S., Sharma, A., & Lindfors, A. V. (2018). Validation of the SARAH-E Satellite-Based Surface Solar Radiation Estimates over India. *Remote Sensing*, *10*(3), Article 3. https://doi.org/10.3390/rs10030392

Roche, S., Bellemare, L., & Ferrari, S. (2018). Rayonner par la technique: Des îles d'Outre-mer au cœur de la transition énergétique française ? *Norois, n° 249*(4), 61–73. https://www.cairn.info/revue-norois-2018-4-page-61.htm

Roger, C. (2014). *Developpement de cellules photovoltaïques à base de CIGS sur substrats métalliques.*

Royo, A. E. G. (n.d.). *SOLAR IRRADIANCE FORECASTING USING NEURAL NETWORKS*. 82.

Saha, S. (2015). Materials Research and Opportunities in Solar (Photovoltaic) Cells. *Proceedings of the Indian National Science Academy*, *81*. https://doi.org/10.16943/ptinsa/2015/v81i4/48309

Schmutz, W. K. (2021). Changes in the Total Solar Irradiance and climatic effects. *Journal of Space Weather and Space Climate*, *11*, 40. https://doi.org/10.1051/swsc/2021016

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. https://doi.org/10.1109/78.650093

Selosse, S., Garabedian, S., Ricci, O., & Maïzi, N. (2018). The renewable energy revolution of reunion island. *Renewable and Sustainable Energy Reviews*, *89*, 99–105. https://doi.org/10.1016/j.rser.2018.03.013

Shafiullah, M., Ahmed, S. D., & Al-Sulaiman, F. A. (2022). Grid Integration Challenges and Solution Strategies for Solar PV Systems: A Review. *IEEE Access*, *10*, 52233–52257. https://doi.org/10.1109/ACCESS.2022.3174555

She, D., & Jia, M. (2021). A BiGRU method for remaining useful life prediction of machinery. *Measurement*, *167*, 108277. https://doi.org/10.1016/j.measurement.2020.108277

Shukla, P. R., Skea, J., Reisinger, A., & Slade, R. (2022). *Climate Change 2022 Mitigation of Climate Change*.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., & Powers, J. G. (2008). *A description of the Advanced Research WRF version 3. NCAR Technical note -475+STR*.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., Wang, W., Powers, J. G., Duda, M. G., Barker, D. M., & Huang, X.-Y. (2019). *A Description of the Advanced Research WRF Model Version 4*. 162.

Skoplaki, E., & Palyvos, J. A. (2009). Operating temperature of photovoltaic modules: A survey of pertinent correlations. *Renewable Energy*, *34*, 23–29. https://doi.org/10.1016/j.renene.2008.04.009

Smets, A., Jäger, K., Isabella, O., van Swaaij, R., & Zeman, M. (2016). Solar cell parameters and equivalent circuit. *Sol. Energy Phys. Eng. Photovolt. Conversion, Technol. Syst*, 113–121.

Smirnov, D., & Nguifo, E. M. (n.d.). *Time Series Classification with Recurrent Neural Networks*. 80.

Smith, R. P., Hwang, A. A.-C., Beetz, T., & Helgren, E. (2018). Introduction to semiconductor processing: Fabrication and characterization of p-n junction silicon solar cells. *American Journal of Physics*, *86*(10), 740–746. https://doi.org/10.1119/1.5046424

Soares de Araujo. (2020). *Combination of WRF Model and LSTM Network for Solar Radiation Forecasting—Timor Leste Case Study*. 37.

Soares de Araujo, J. (2021). Improvement of Coding for Solar Radiation Forecasting in Dili Timor Leste—A WRF Case Study. *Journal of Power and Energy Engineering*, *09*, 7–20. https://doi.org/10.4236/jpee.2021.92002

Sobri, S., Koohi-Kamali, S., & Rahim, N. Abd. (2018). Solar photovoltaic generation forecasting methods: A review. *Energy Conversion and Management*, *156*, 459–497. https://doi.org/10.1016/j.enconman.2017.11.019

Solar Impulse completes historic round-the-world flight. (2016). *Reinforced Plastics*, *60*(6), 339–340. https://doi.org/10.1016/j.repl.2016.10.013

Song, Z., Li, C., Chen, L., & Yan, Y. (2022). Perovskite Solar Cells Go Bifacial—Mutual Benefits for Efficiency and Durability. *Advanced Materials*, *34*(4), 2106805. https://doi.org/10.1002/adma.202106805

Tang, C., Morel, B., Wild, M., Pohl, B., Abiodun, B., & Bessafi, M. (2019). Numerical simulation of surface solar radiation over Southern Africa. Part 1: Evaluation of regional and global climate models. *Climate Dynamics*, *52*(1–2), 457–477. https://doi.org/10.1007/s00382-018-4143-1

Tapiador, F. J., Navarro, A., Moreno, R., Sánchez, J. L., & García-Ortega, E. (2020). Regional climate models: 30 years of dynamical downscaling. *Atmospheric Research*, *235*, 104785. https://doi.org/10.1016/j.atmosres.2019.104785

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183–7192. https://doi.org/10.1029/2000JD900719

Team Keras, K. (n.d.). *Keras documentation: Keras API reference*. Retrieved May 1, 2023, from https://keras.io/api/

Tewari, M., Chen, F., Wang, W., Dudhia, J., LeMone, M. A., Gayno, G., Wegiel, J., & Cuenca, R. H. (2004). *14.2A IMPLEMENTATION AND VERIFICATION OF THE UNIFIED NOAH LAND SURFACE MODEL IN THE WRF MODEL*. 6.

United Nations. (2015). *United Nations Sustainable Development Summit 2015*. https://sustainabledevelopment.un.org/post2015/summit

U.S. Department of Energy, P. (2006). *Benefits of Demand Response in Electricity Markets and Recommendations for Achieving Them*.

Verbois, H., Rusydi, A., & Thiery, A. (2018). Probabilistic forecasting of day-ahead solar irradiance using quantile gradient boosting. *Solar Energy*, *173*, 313–327. https://doi.org/10.1016/j.solener.2018.07.071

Verduci, R., Romano, V., Brunetti, G., Yaghoobi Nia, N., Di Carlo, A., D'Angelo, G., & Ciminelli, C. (2022). Solar Energy in Space Applications: Review and Technology Perspectives. *Advanced Energy Materials*, *12*(29), 2200125. https://doi.org/10.1002/aenm.202200125

Voyant, C. (2011). *Prédiction de séries temporelles de rayonnement solaire global et de production d'énergie photovoltaïque à partir de réseaux de neurones artificiels*. 258.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, *105*, 569–582. https://doi.org/10.1016/j.renene.2016.12.095

Wan, C., Zhao, J., Song, Y., Xu, Z., Lin, J., & Hu, Z. (2015). Photovoltaic and solar power forecasting for smart grid energy management. *CSEE Journal of Power and Energy Systems*, *1*, 38–46. https://doi.org/10.17775/CSEEJPES.2015.00046

Wang, F., Xuan, Z., Zhen, Z., Li, K., Wang, T., & Shi, M. (2020a). A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework. *Energy Conversion and Management*, *212*, 112766. https://doi.org/10.1016/j.enconman.2020.112766

Wang, F., Xuan, Z., Zhen, Z., Li, K., Wang, T., & Shi, M. (2020b). A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework. *Energy Conversion and Management*, *212*, 112766. https://doi.org/10.1016/j.enconman.2020.112766

Wang, F., Yu, Y., Zhang, Z., Li, J., Zhen, Z., & Li, K. (2018). Wavelet Decomposition and Convolutional LSTM Networks Based Improved Deep Learning Model for Solar Irradiance Forecasting. *Applied Sciences*, *8*(8), 1286. https://doi.org/10.3390/app8081286

Wang, H., Yi, H., Peng, J., Wang, G., Liu, Y., Jiang, H., & Liu, W. (2017). Deterministic and probabilistic forecasting of photovoltaic power based on deep convolutional neural network. *Energy Conversion and Management*, *153*, 409–422. https://doi.org/10.1016/j.enconman.2017.10.008

Wang, W., Bruyère, C., Duda, M., Dudhia, J., & Zhang, X. (2012). *ARW Version 3.4 Modelling System User's Guide*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.386.3535

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Widén, J., Wäckelgård, E., Paatero, J., & Lund, P. (2010). Impacts of distributed photovoltaics on network voltages: Stochastic simulations of three Swedish low-voltage distribution grids. *Electric Power Systems Research*, *80*(12), 1562–1571. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-132904

Wild, M., Ohmura, A., Schär, C., Müller, G., Folini, D., Schwarz, M., Hakuba, M. Z., & Sanchez-Lorenzo, A. (2017). The Global Energy Balance Archive (GEBA) version 2017: A database for worldwide measured surface energy fluxes. *Earth System Science Data*, *9*(2), 601–613. https://doi.org/10.5194/essd-9-601-2017

Worku, M. Y., & Abido, M. A. (2015). Grid-connected PV array with supercapacitor energy storage system for fault ride through. *2015 IEEE International Conference on Industrial Technology (ICIT)*, 2901–2906. https://doi.org/10.1109/ICIT.2015.7125526

Yang, T., Li, B., & Xun, Q. (2019). LSTM-Attention-Embedding Model-Based Day-Ahead Prediction of Photovoltaic Power Output Using Bayesian Optimization. *IEEE Access*, *7*, 171471–171484. https://doi.org/10.1109/ACCESS.2019.2954290

Yang, Z., & Wang, J. (2018). A hybrid forecasting approach applied in wind speed forecasting based on a data processing strategy and an optimized artificial intelligence algorithm. *Energy*, *160*, 87–100. https://doi.org/10.1016/j.energy.2018.07.005

Yona, A., Senjyu, T., Funabashi, T., & Kim, C.-H. (2013). Determination Method of Insolation Prediction With Fuzzy and Applying Neural Network for Long-Term Ahead PV Power Output Correction. *IEEE Transactions on Sustainable Energy*, *4*(2), 527–533. https://doi.org/10.1109/TSTE.2013.2246591

Yu, C., Li, Y., Bao, Y., Tang, H., & Zhai, G. (2018). A novel framework for wind speed prediction based on recurrent neural networks and support vector machine. *Energy Conversion and Management*, *178*, 137–145. https://doi.org/10.1016/j.enconman.2018.10.008

Yule, G. U. (1926). Why do we Sometimes get Nonsense-Correlations between Time-Series?—A Study in Sampling and the Nature of Time-Series. *Journal of the Royal Statistical Society*, *89*(1), 1–63. https://doi.org/10.2307/2341482

Zamo, M., Mestre, O., Arbogast, P., & Pannekoucke, O. (2014a). A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: Deterministic forecast of hourly production. *Solar Energy*, *105*, 792–803. https://doi.org/10.1016/j.solener.2013.12.006

Zamo, M., Mestre, O., Arbogast, P., & Pannekoucke, O. (2014b). A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. *Solar Energy*, *105*, 804–816. https://doi.org/10.1016/j.solener.2014.03.026

Zang, H., Liu, L., Sun, L., Cheng, L., Wei, Z., & Sun, G. (2020). Short-term global horizontal irradiance forecasting based on a hybrid CNN-LSTM model with spatiotemporal correlations. *Renewable Energy*, *160*, 26–41. https://doi.org/10.1016/j.renene.2020.05.150

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0

Zhang, X., Liang, S., Wang, G., Yao, Y., Jiang, B., & Cheng, J. (2016). Evaluation of the Reanalysis Surface Incident Shortwave Radiation Products from NCEP, ECMWF,

GSFC, and JMA Using Satellite and Surface Observations. *Remote Sensing*, *8*(3), Article 3. https://doi.org/10.3390/rs8030225

Zhang, Y., Qin, C., Srivastava, A. K., Jin, C., & Sharma, R. K. (2020). Data-Driven Day-Ahead PV Estimation Using Autoencoder-LSTM and Persistence Model. *IEEE Transactions on Industry Applications*, *56*(6), 7185–7192. https://doi.org/10.1109/TIA.2020.3025742

Zhao, J., Guo, Z.-H., Su, Z.-Y., Zhao, Z.-Y., Xiao, X., & Liu, F. (2016). An improved multi-step forecasting model based on WRF ensembles and creative fuzzy systems for wind speed. *Applied Energy*, *162*, 808–826. https://doi.org/10.1016/j.apenergy.2015.10.145

Zhen, H., Niu, D., Wang, K., Shi, Y., Ji, Z., & Xu, X. (2021). Photovoltaic power forecasting based on GA improved Bi-LSTM in microgrid without meteorological information. *Energy*, *231*, 120908. https://doi.org/10.1016/j.energy.2021.120908

Zhou, H., Zhang, Y., Yang, L., Liu, Q., Yan, K., & Du, Y. (2019). Short-Term Photovoltaic Power Forecasting Based on Long Short Term Memory Neural Network and Attention Mechanism. *IEEE Access*, *7*, 78063–78074. https://doi.org/10.1109/ACCESS.2019.2923006

Zhu, D., & Ooka, R. (2023). WRF-based scenario experiment research on urban heat island: A review. *Urban Climate*, *49*, 101512. https://doi.org/10.1016/j.uclim.2023.101512