Satellite-based Methods for Intra-day Solar Irradiation Forecast in French Guiana

Maha Salloum

Thèse

En vue de l'obtention du grade de

Docteur de l'université de Guyane

Spécialité : *Sciences de la terre et de l'univers*

Présentée et soutenue publiquement le ..... *2021* par :

maha SALLOUM

Satellite-based Methods for Intra-day Solar Irradiation Forecast in French Guiana

JURY

Ted SOUBDHAN	Premier MEMBRE	PU	Prés
Stéphane GRIEU	Deuxième MEMBRE	PU	
Sara ZERMANI	Troisième MEMBRE	MCF	
Tommy ALBARELO	Quatrième MEMBRE	DR	
Laurent LINGUET	Cinquième MEMBRE	PU	
Antoine PRIMEROSE	Sixième MEMBRE	MCF	Dire

Ecole Doctorale n°587 : *Diversités, Santé et développement en Amazonie*

“Everything that is or was, began with a dream….”

lavagirl

Abstract

Satellite-based Methods for Intra-Day Solar Irradiation Forecast in French Guiana

maha SALLOUM

Population growth and continuous demand for energy production have prompted electricity producers to focus increasingly on the use of local and clean resources. Renewable energy sources such as the sun and wind have shown an increasing interest. A problem with these power sources are their intermittent and random nature that drives grid operators to restrict their integration into the energy mix. This makes it imperative to combine different production systems to ensure grid stability and safety. French Guiana's electricity production mainly comes from local renewable energies (64%), compared to 36% from imported fossil fuels. To improve the integration of intermittent renewable energies, solar energy in this study, it is necessary to focus on resource forecasting. Knowing in advance the available power allows optimal management of the coupling between conventional and intermittent production systems.

The contribution of this thesis focuses on forecasting the Global Horizontal Irradiation (GHI) at different time horizons combining satellite-derived data instead of ground measurements and statistical methods. Satellite images offer the advantage of providing irradiance products over wide areas and with satisfactory accuracy. We are interested in forecasting the GHI as the generated PV power directly depends on the incoming GHI intensity. In this thesis, we have developed and studied ten solar radiation prediction models. Their methods are the Persistence, Scaled Persistence, AR, ARMA, Gaussian Process, Support Vector Machine, simple regression trees, bagged forests, WRF with Kalman filter, and an aggregation method. Satellites-derived data from the meteorological geostationary satellite GOES-13 are used as input of each model. These models were first developed to predict the GHI using satellite-derived data as input, and then to predict the GHI using ground measurements to be able to quantify discrepancies. The forecast horizons tested are from 1 to 6H per hour time step, to provide useful horizons for network managers. Data from six measurement sites located in French Guiana are used for GHI prediction. Five years of data from the six stations were used for the learning phase and one year for the validation phase.

First, we have characterized the data from each site to study their variability, which means their tendency to vary strongly or not with time. The main results of this work are that the forecast errors using the satellite-derived and ground measurements data as input are slightly different for intra-day forecasts; for all models. For a 1H time horizon, the model's performance using ground data is better than models using satellite-derived data as input. Horizons higher than 2H, models using data derived from the satellite are the same or better performing depending on the method used. It appears that the machine learning models tested in this work give similar results with the tested dataset and that the choice of one or the other will rather be made according to the technical and practical constraints

of the tools. Moreover, the uncertainties using machine learning models are better than the statistical models. Since the WRF model numerically solves equations describing the atmospheric processes, the WRF prediction contains information about the future state of the atmosphere. It was interesting for us to study the post-processing of the solar forecasts by WRF model using satellite-derived data. We have optimized the 24H forecast of WRF at hourly time steps using satellite-derived data and a Kalman filter. Using satellite-derived data as input for horizons ranging from 1H to 5H, we obtained better performance using WRF-kalman than the one obtained using only WRF. For the horizon of 6H, the Kalman filter is not considered suitable as a post-processing method. This is an interesting case because it will allow to reduce WRF bias and to maintain the advantage of WRF's spatial resolution in French Guiana. Finally, we continued the development and optimization of the forecasts by using an Aggregation method as a hybrid ensemble model. The models are used as members of the hybrid ensemble model. The comparison between the hybrid ensemble method and each member shows an improvement in the quality of the forecasts.

Keywords: forecasting, global radiation, satellite-derived data, Inter-Tropical Zone, Auto Regression model, machine learning models, WRF, Kalman filter, Ensemble Method.

Résumé

Prévision intra-journalière de la ressource solaire en Guyane française par Méthodes satellitaires

maha SALLOUM

La croissance démographique et la demande continue de production d'énergie ont incité les producteurs d'électricité à se concentrer de plus en plus sur l'utilisation de ressources locales et propres. Les sources d'énergie renouvelables telles que le soleil et le vent ont montré un intérêt croissant. Un problème avec ces sources d'énergie est leur caractère intermittent et aléatoire qui pousse les gestionnaires de réseau à restreindre leur intégration dans le mix énergétique. Il est donc impératif de combiner différents systèmes de production pour assurer la stabilité et la sécurité du réseau. La production d'électricité de la Guyane française provient principalement d'énergies renouvelables locales (64 %), contre 36 % d'énergies fossiles importées. Pour améliorer l'intégration des énergies renouvelables intermittentes, l'énergie solaire dans cette étude, il est nécessaire de se concentrer sur la prévision des ressources. Connaître à l'avance la puissance disponible permet une gestion optimale du couplage entre les systèmes de production conventionnels et intermittents.

La contribution de cette thèse se concentre sur la prévision de l'irradiation horizontale globale (GHI) à différents horizons temporels en combinant des données satellitaires au lieu de mesures au sol et méthodes statistiques. Les images satellitaires offrent l'avantage de fournir des produits d'éclairement sur de larges zones et avec une précision satisfaisante. Nous nous intéressons à la prévision du GHI car la puissance photovoltaïque générée dépend directement de l'intensité du GHI qui entre. Dans cette thèse, nous avons développé et étudié neuf modèles de prédiction du GHI. Leurs méthodes sont la persistance, Scaled persistance, AR, ARMA, processus gaussien, machine à vecteurs de support, arbres de régression simples, forêts ensachées, WRF avec filtre de Kalman et une méthode d'agrégation. Les données dérivées du satellite météorologique géostationnaire GOES-13 sont utilisées comme entrée de chaque modèle. Ces modèles ont d'abord été développés pour prédire le GHI en utilisant des données satellitaires en entrée, puis pour prédire le GHI en utilisant des mesures au sol pour pouvoir quantifier les écarts. Les horizons de prévision testés sont de 1 à 6H par pas de temps horaire, pour prévoir des horizons utiles aux gestionnaires de réseau. Les données de six sites de mesure situés en Guyane française sont utilisées pour la prédiction du GHI. Cinq années de données des six stations ont été utilisées pour la phase d'apprentissage du modèle et une année pour la phase de validation.

Dans un premier temps, nous avons caractérisé les données de chaque site et étudié leur variabilité, c'est-à-dire leur tendance à varier fortement ou non dans le temps. Les principaux résultats de ce travail sont que les erreurs de prévision utilisant les données satellitaires et les mesures au sol en entrée sont légèrement différentes pour les prévisions intra-journalières ; pour tous les modèles. Pour un horizon temporel de 1H, les

performances du modèle utilisant des données au sol sont meilleures que celles des modèles utilisant des données satellitaires en entrée. Horizons supérieurs à 2H, les modèles utilisant les données dérivées du satellite sont les mêmes ou mieux performants selon la méthode utilisée. Il apparaît que modèles d'apprentissage automatique testés dans ce travail donnent des résultats similaires avec le jeu de données testé et que le choix de l'un ou de l'autre se fera plutôt en fonction des contraintes techniques et pratiques des outils. De plus, les incertitudes utilisant les modèles d'apprentissage automatique sont beaucoup plus faibles que les modèles statistiques. Étant donné que le modèle WRF résout numériquement le système d'équations différentielles décrivant les processus physiques et chimiques de l'atmosphère, la prédiction WRF contient des informations sur l'état futur de l'atmosphère. Il était intéressant pour nous d'étudier le post-traitement des prévisions solaires par le modèle WRF à partir de données satellitaires. Nous avons optimisé la prévision de 24H de WRF à des pas de temps horaires à l'aide de données satellitaires et d'un filtre de Kalman. En utilisant des données satellitaires en entrée pour des horizons allant de 1H à 5H, nous avons obtenu de meilleures performances en utilisant WRF-kalman que celles obtenues en utilisant uniquement WRF. Pour l'horizon de 6H, le filtre de Kalman n'est pas considéré comme approprié comme méthode de post-traitement. C'est un cas intéressant car il permettra de réduire le biais de WRF et de conserver l'avantage de la résolution spatiale de WRF en Guyane française. Enfin, nous avons poursuivi le développement et l'optimisation des prévisions en utilisant une méthode d'agrégation comme modèle d'ensemble hybride. Les modèles sont utilisés comme membres du modèle d'ensemble hybride. La comparaison entre la méthode d'ensemble hybride et chaque membre montre une amélioration de la qualité des prévisions.

Mots clés : Prévisions, Rayonnement global, Données satellitaires, Zone Intertropicale, modèle de régression automatique, modèles d'apprentissage automatique, WRF, filtre de Kalman, méthode d'ensemble.

Acknowledgments

And who knew we would see the end?

It is at the end of my thesis that I would like to humbly thank all the people who have helped me directly or indirectly in this adventure.

I still remember my first day at French Guiana university. I literally had a little idea what PhD is about. Hardworking and not giving up were the key to achieving my goal. It was a challenging time for me, and I love challenges. I enjoyed every moment of my PhD despite lockdowns, stress, feeling down, being away from home, friends and family. All these circumstances make me feel so grateful to everyone who supported and helped me during this hard time.

It would not have been possible to write this master thesis without the help and support of the kind people around me, to only some of whom it is possible to give particular mention here. I am grateful to a number of people who offered helpful advice and comments.

Three years of our lives to make this dream become true turned into five. Doing this adventure with a self-financing was not an easy choice. So above all, I would like to thank my husband ***HAMAMAH Ghaleb*** for his personal, moral and financial support and his great patience at all times. My children ***Julia*** and ***Yann***, without your support, patience and understanding, i could not have completed my PhD journey.

Without my supervisors, I could not have achieved the PhD. Thank you my principal supervisor, ***Laurent LINGUET***. This PHD opportunity would not have been possible without your help, and support during my PhD time. Thank you my second supervisor ***Antoine PRIMEROSE.***

Also, I am so grateful to my examiners*,* ***Stephan GRIEU, Ted SOUBDHAN***, ***Sara ZERMANI*** and ***Tommy ALBARELO*** for their valuable input to my thesis and their time to read my work.

The good advice, support and friendship of my true friend ***Mouhammet DIALLO,*** for encouragement, discussions and advices, especially during some difficult times in the PhD.

To all those who participated in my studies, thank you for your input. Without your valuable input, I could not have achieved the thesis aims.

Big thanks also to my friends and colleagues in the **University** and **IRD laboratory**; ***Marie-line GOUBINDAS, Wilna LESPERENCE*** for their kindness, support, helping me all these years.

I would also like to thank ***UMR ESPACE DEV team*** in French Guiana. I could not have pursued my dream of getting a PhD and becoming a researcher without their support.

I also take this opportunity to express a deep sense of gratitude to ***Abdennebi OMRANE*** and ***Rachelle HO-COUI-YOUN*** for the valuable provided information, and their experiences.

I want to acknowledge my family love, support, and encouragement, particularly my father; ***Kasser***, mother; ***Ezdihar***, and my beautiful sisters and brothers; ***Yara, Mazen, Hasan, Bachar*** and **Walaa** have given me their unequivocal support throughout, as always, for which my mere expression of thanks likewise does not suffice.

I want to thank my supported, encouraged friends in France; ***hind TLIJANI, safa LABIDI, mouchira LABIDI,*** my lovely sisters in French Guiana ***Sara, Sophie, Koni, Sabrina, Rafika, Fatia, Andria*** for their constant encouragement without which this assignment would not be possible.

Finally, I would like to thank all those I did not mention explicitly for supporting me on my path, I am grateful for their cooperation during the period of my assignment.

Contents

List of Figures

List of tables

List of abbreviations

GHI	*Global horizontal irradiance,* W/m²
Gnd	*Ground Measured Data*
Sat	*Satellite-derived Data*
K_c	*Clear sky index*
K_t	*Clearness index*
G_C	*Clear sky model, W/m²*
G_{toa}	*Solar irradiance Arriving At The Top Of The Earth's Atmosphere, W/m²*
SZA	*Solar Zenith Angle*
AZA	*Solar Azimuth Angle*
ELV	*Solar elevation Angle*
GOES	*Geostationary Operational Environmental Satellite*
P	*Persistence*
SP	*Scaled Persistence*
AR	*Auto Regressive*
ARMA	*Auto Regressive Moving Average*
GPR	*Gaussian Process Regression*
SVM	*Support Vector Machine*
TRE	*Regression Trees*
ENsTRE	*Ensembles of Trees 'Bagging'*
WRF	*Weather Research and Forecasting model*
KAL	*Kalman filter*
RR	*Ridge regression method*
nRMSE	*Normalized Root Mean Squared Error*
nMAE	*Normalized Mean Absolut Error*
nMBE	*Normalized Mean Bias Error*
MK_C	*Mean clear sky index*

VK_c	*mean clear sky index variability*
SS	*Skill score*
aR2	*adjusted R square*
ESRA	European Solar Radiation Atlas
NWP	*Numerical Weather Prediction*
ITZ	*InterTropical Zone*
NOAA	*National Oceanic and Atmospheric Administration*
RO	*Rochambeau site*
SL	*Saint Laurent site*
MP	*Maripasoula site*
SG	*Saint Georges site*
IR	*Iles Royale site*
KR	*Kourou site*
H	*Hour*
GFS	*Global Forecast System*
QC	*Quality Check*

Chapter .I. Introduction general

I.1. Introduction

I.1.A. Renewable energies: issues and challenges

Renewable energies refer to energy sources that do not have undesirable consequences on the environment. The energy transition from fossil to renewable sources is a crucial point toward the environment and to reduce greenhouse gas emissions. Renewable energy sources are extendible and are considered "free" energy sources. These include biomass energy, wind energy, solar energy, geothermal energy, and hydroelectric energy sources. Combined with the use of recycling and clean alternative energies, such as the use of solar power systems, this will help ensure human survival in the 21st century and beyond. Solar energy is one of the fastest-growing sectors of the global energy market ('Snapshot of Global PV Markets' 2020). The growth of solar energy, since 2009, has been impressive. Solar energy has many basic advantages like its renewable nature, free of cost, abundant, and inexhaustible. The total amount of energy emitted by the sun at the earth's surface could generate enough energy to meet the increasing demand.

In 2019, solar PV stood for approximately 59% of the total renewable electricity production from new production assets ('Solar Power Europe' 2019). Nevertheless, due to the diurnal cycle and the occurrence of clouds, solar PV plant can't operate at full capacity every time. Yet, renewable resources such as biomass installations can virtually produce all day and all year-round.

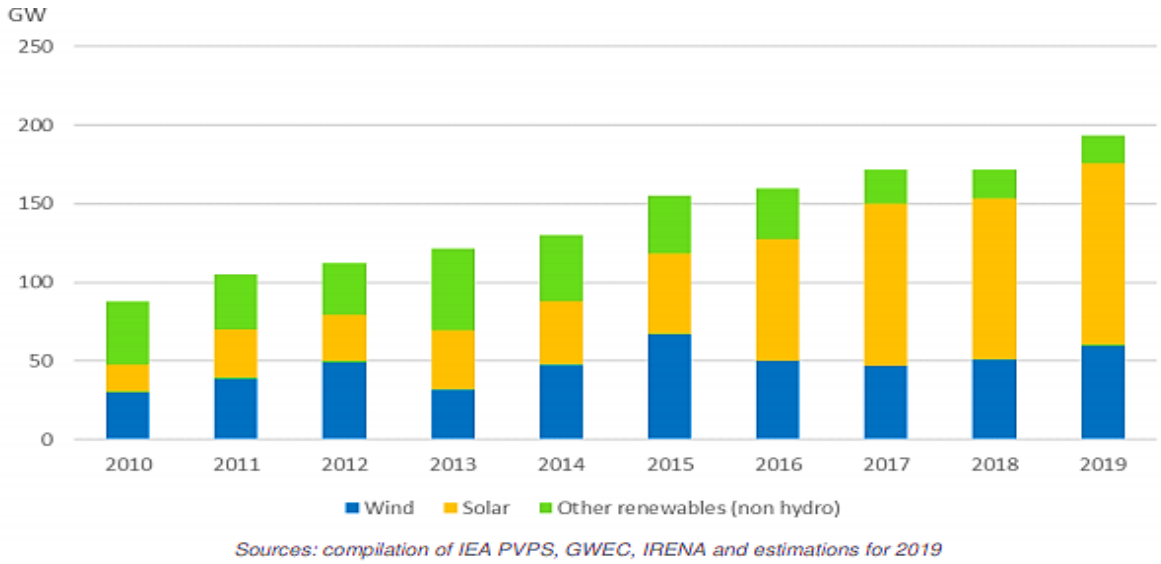


Figure 1: Evolution of renewable energy technologies (source 'Snapshot of Global PV Markets' 2020).

The same year, The global total installed PV capacity has exceeded 500 GW in 2019 (Figure 2) and is expected to exceed 1 TW in 2022. Additionally, global solar energy

demand is expected to increase by 7% to 169 GW in 2022, and 6% to 180 GW in 2023 ('Solar Power Europe' 2019).

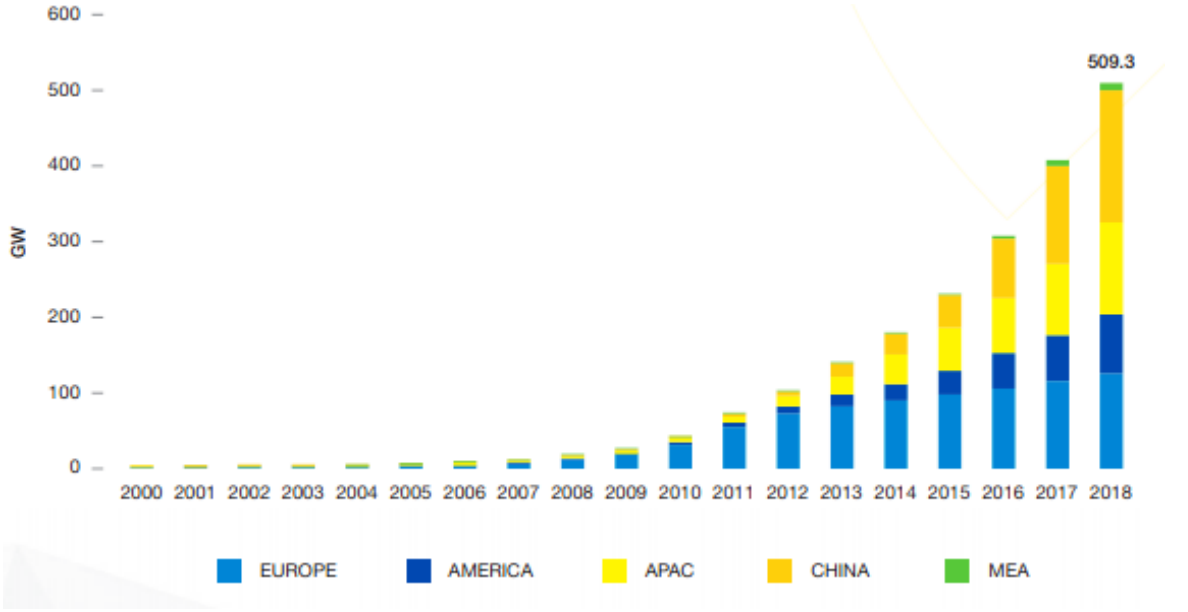


Figure 2: Evolution of the global total installed PV capacity in 2000-2018. Europe is shown in light blue, America in dark blue, China in orange, the Asia Pacific (APAC) states (excluding China) in yellow, and the Middle East with Africa (MEA) in green. ('Solar Power Europe' 2019).

A major drawback of exploiting solar power is its intermittent nature. Although solar energy can still be collected during cloudy and rainy days, the efficiency of the solar system decreases, so the generating energy is reduced. This phenomenon could cause damage to the operating systems. As the use of solar energy continues to grow, it becomes crucial to ensure grid stability and to enable optimal control of electricity and economic dispatch. Forecasting solar resource is important for grid management; however, the variability of solar resources is a major constraint. It is difficult to maintain the balance between electricity production and demand. As a consequence, operating optimally a solar PV installation requires planning, and management, including forecasting its production. Looking toward the integration of the increasing renewable energy sources into the electricity grid, these steps would guarantee the grid stability and cost-effectiveness (Voyant, Notton, et al. 2017). Forecasting solar resources gives the possibility to predict the resource availability and its variations. In the case of lower solar energy availability, this energy can be replaced by an alternative energy storage system used in off-the-grid solar systems. Consequently, the solar energy will be replaced without a significant decrease in volts.

I.1.B. Technologies for measuring solar radiation

The most used way to measure the irradiation is to install measurement station. This ground station include an instrument called a "pyranometer" that measures the overall radiation received on the ground (Chambers, 1977). The installation, implementation, and operation of the pyranometer impose some restrictions. First restrict, no shadows cast near the measuring instrument (buildings, vegetation). Then, the pyranometer should be regularly maintained: cleaning the pyranometer dome, checking and even replacing calibration of the equipment and materials, etc. All these upkeep and

maintenance operations require human intervention, which explains why there are no measuring stations at every point on the earth, (Figure 3). To make up for the lack of irradiation measurement in areas where there is no pyranometer installed, interpolation techniques can be used to produce data at all points.

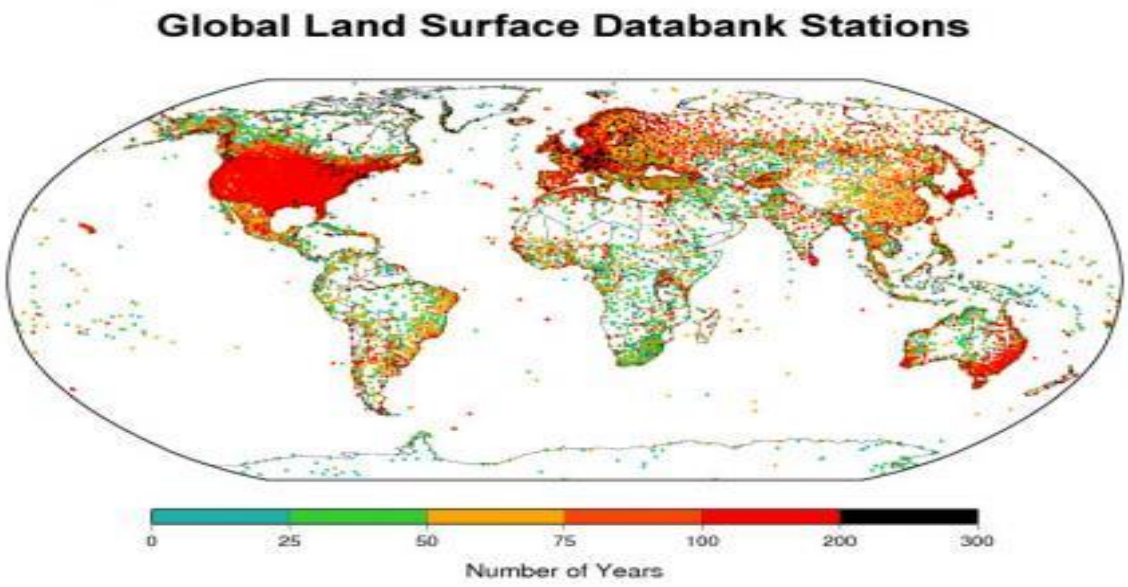


Figure 3: Location and period of record of approximately 32,000 stations in version one of the Global Historical Climatology Network stations in 2010 (NOAA, 2020)

The in situ data interpolation allows for mapping of missing solar irradiation, however, the interpolation of irradiation data can produce relevant and sufficiently accurate estimates that up to an average distance of 50 km between stations for daily average irradiation values and up to a distance of 34 km between stations for hourly average irradiation values (Perez et al. 1994). Beyond these distances, an alternative approach for producing solar irradiation data at any point is based on the use of observations and satellite images acquired by imaging radiometers. Regions located in the intertropical zone (ITZ) have higher solar potential (Trieb et al. 2009), The high variability of solar irradiation in this zone is due to the speedy cloud transformations (Aryaputera 2015). Consequently, solar energy arrived at the ground in this area is highly variable. This variability causes economical and technical challenges to fully exploit this resource.

I.1.C. Technologies for exploiting solar radiation

Solar panels depend on sunlight to produce solar energy. Consequently, solar energy cannot be collected during the night. Although solar energy can still be collected during cloudy days, the quantity of this energy is not the same as during clear sky. It can be used right away, or it can be stored in large batteries; Figure 4.

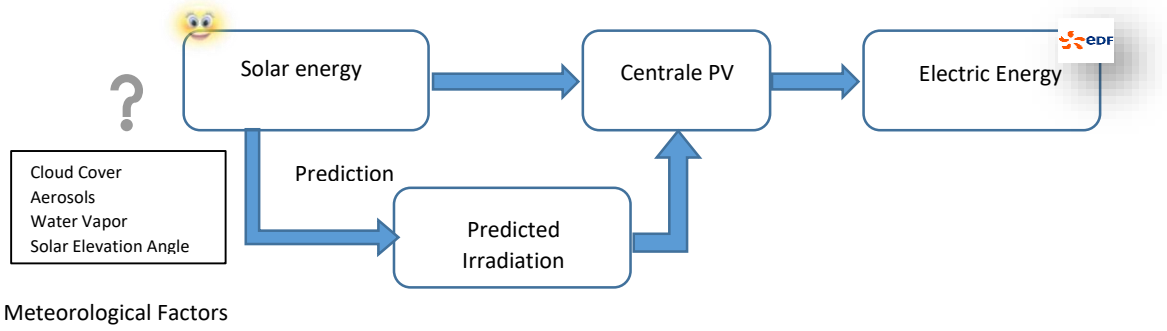


Figure 4: solar energy prediction system.

These batteries (used in off-the-grid solar systems) can be charged during the day so that the energy is used at night or when energy shortage occurs. This is a good solution for using solar energy all day long, but it is also an expensive one. Producing energy from the global solar radiation received on the ground can be more interesting by predicting the latter mentioned.

I.1.D. Solar forecasting in the Inter Tropical Zone

Also called Equatorial Convergence Zone. It is an east-west-oriented low-pressure zone near the equator, where the northeast and southeast surface winds meet. The ITCZ moves over land and oceans near the equator, this zone is usually found within 350 km of the equator; about 80–300 km wide and 5° north and south from the equator. Figure 5 shows the northernmost and southernmost positions of the ITCZ.

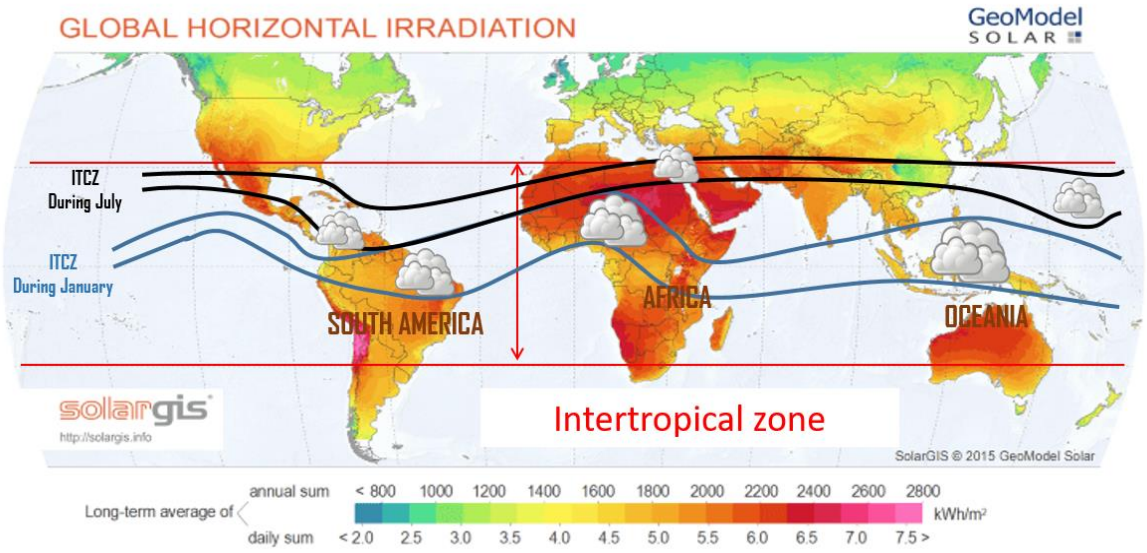


Figure 5: The northernmost and southernmost positions of the ITCZ in January (blue) and July (black) (source:Credit Solargis)

The position of the ITCZ varies seasonally, The ITCZ crosses French Guiana as it travels to the north from May to July, the ITCZ moves northward and it is located roughly at 5° N. Then again, as it moves south between November and January, it moves southward to be approximately located at 15° S (Albarelo et al. 2015a). The ITCZ is warm all year, averaging 25 to 28 degrees Celsius as this zone gets more exposure to the sun. because of

all that sun, the tropics do not experience the kind of season as other regions. Giving these countries two seasons: the wet season and the dry season.

From the meteorological point of view this shift from North to the south creates a seasonal cycle with 4 periods (Bovolo et al. 2012):

Dry season: The ITCZ lies north between July and November. Here, the sky is mostly clear although weak precipitation may occur (Albarelo et al. 2015a). During this period, cirrus and cumulus clouds are predominant.

Short rainy season: The ITCZ moves southward of French Guiana, starts from November to January, (Albarelo et al. 2015a)

Short dry season: is a period of transition before the ITCZ begins its northward motions again, starts from February to March (Albarelo et al. 2015a).

Rainy season: occurs when the ITCZ moves northward, starts from April to May. During this period cumulonimbus clouds are predominant.

Rising temperatures, changes in precipitation, sea-level rise, an increase in the number of extreme events, everyone has heard of climate change and its many consequences for the entire planet. From the solar irradiance point of view, the ITCZ shift from North to the south create a high variability of the incoming solar resource. Its regional impact is rarely examined. So we focus on French Guiana in the international trade area.

I.1.E. Energy issues in French Guiana

French Guiana is considered an islanded microgrid since it is non-interconnected to its neighbors Brazil and Suriname. A unique electricity network supplies the coastline from East to West, and the rest of the territory remains isolated and highly dependent on imported oil products (Lesperance et al. 2018). Fossil resources make up about 80% of French Guiana's energy supply. Local resources represent only 20% of French Guiana's energy supply and more than 89% consist of hydraulic energy. French Guiana's electricity production (at 862.6 GWh in 2014), comes mainly from local renewable energies (64%), compared to 36% from imported fossil fuels (ADEME 2017). Between 2009 and 2014, the population in French Guiana increased by 27,000 inhabitants, representing a growth of 12%. This growth must be followed by an increase in electricity production and consumption. Consequently, French Guiana aims to increase the share of renewable energy in the electricity mix to 80% by 2023 ('Albioma_DDR' 2018).

Table 1: French Guiana: Development objectives for electricity production from renewable energies (Source: Rapport PPE French Guiana (mars 2017))

ENR	**POWER INSTALLER PARAPORT OF 2015**	
	2018 (MW)	**2023 (MW)**
Hydraulic energy	4,5	16,5
Pv energy without storage energy	8	26
Pv energy with storage energy	15	25
Wind energy with storage energy	10	20
Biomass energy	15	40
Waste to energy	0	8

French Guiana had 45 MWp of PV panels installed at the end of 2016. In 2016, this park produced 55.4 GWh of energy or 6% of the energy supplied to the grid. The park is made up of 35 MWp of panels without storage and 9.8 MWp of PV panels with storage capacity (ADEME 2017). Table 1 shows the renewable energy production to be achieved, by 2018 and 2023 in French Guiana.

I.2. Basic terminologies used in solar energy forecast

The production of electrical energy by a photovoltaic system depends on several factors. Slope, technology and PV temperature affect the production, but the received solar radiation remains the main factor of variation. Solar radiation assessment is an important part of sizing and planning of a photovoltaic power plant. Incoming hourly solar radiation data on the surface of a region include both deterministic and stochastic behaviors. The deterministic part comes from the solar geometry, whereas the stochastic part has occurred due to random atmospheric events such as the motion of clouds. In this section, we will define different basic concepts used in solar energy production.

I.2.A. Solar Radiation

Solar radiation is made up of electromagnetic waves emitted by the Sun. It includes wavelengths ranging from infrared to ultraviolet. Most of the solar radiation entering the earth's atmosphere will be absorbed, scattered and reflected (Messenger and Ventre 2004). The three components of solar radiation are the following:

Direct **N**ormal radiation (DN): Direct beam radiation comes in a direct line from the sun to the earth's surface.

Diffuse **H**orizontal radiation (DH): Diffuse beam radiation, scattered out of the direct beam by molecules, cloud droplets, ice crystals, and aerosols (Wald, 2007).

Albedo radiation: Sunlight that is reflected from the ground is known as albedo radiation.

The sum of the direct beam diffuse, and ground-reflected radiation arriving on a surface horizontal to the surface of the earth is called **G**lobal **H**orizontal radiation (GH)

Most of the energy emitted by the Sun is gathered at wavelengths of the visible spectrum (0.39 μm to 0.76 μm). The global irradiation energy received on the ground at any time per unit area is called global irradiation, and its unit is watthour per square meter ($Wh.m^{-2}$). We usually define two causes of GHI variability:

- The variability linked to cloud cover and the movement of clouds.
- The variability traceable to the annual and daily variability induced by the solar cycle.

Atmospheric factors such as aerosols and water vapor also influence GHI variability. The expression between these three solar components is given using the following equation:

GHI = DHI + DNI * cos (**SZA**). 1. 1

These three types of surface shortwave irradiation can be obtained using various models such as general circulation models, reanalysis products, satellite-to-irradiation models and ground-based radiometers. Different instruments are used to measure the different quantities of the components of solar radiation, where a pyrheliometer on a solar tracker is used to measure the direct normal radiation, to measure the global horizontal radiation a pyranometer is used, and a pyranometer (shaded) is used to measure the diffuse horizontal radiation (World Meteorological Organization 2008).

I.2.B. Solar Position

The position of the sun as seen from a particular place on the surface of the earth is an important input needed to forecast solar radiation and model PV System performance. This geometric position varies from day to day and hour to hour in the year. At any given instance, the sun's position can be fixed by three angles, altitude, zenith and azimuth (Solstice et al. 2013), represented in Figure 6.

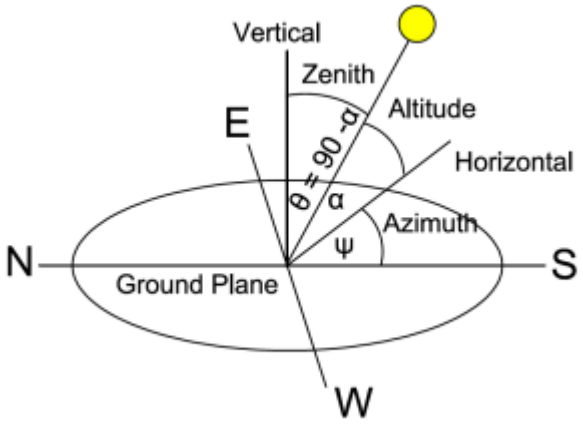


Figure 6: Zenith, altitude and azimuth angles for northern latitude (Shakya 2016).

Solar Zenith Angle (θz): is the angle between the connecting line between the observations on the ground and the sun and the vertical above the observation (directly overhead),(Duffie and Beckman 2013).

$$cos\theta z = sin\delta\ sin\varphi + cos\delta\ cos\varphi\ cos\omega \qquad 1.2$$

Where: ω the hour angle; is expressed as difference between noon and the desired time of day in terms of a 360° rotation in 24 hours. δ the declination angle; depends upon the revolution of earth and remains constant for a day. φ the latitude; is the angular location north or south of the equator. The view of zenith angle matrix values from 0 to 90° is shown in Figure 7.

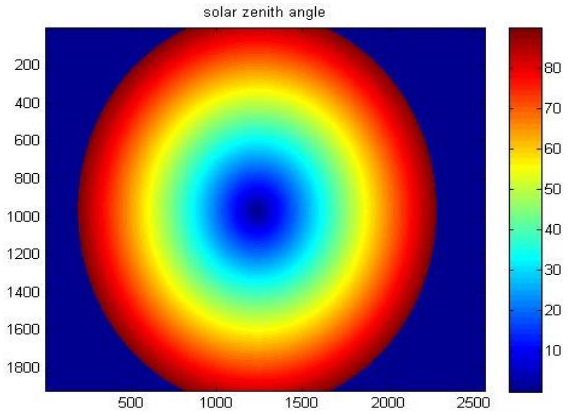


Figure 7: Solar Zenith Angle matrix in MATLAB.

Solar Azimuth Angle (ψ): It is the angle between the horizontal direction of the sun and a reference direction (usually North 0°), Figure 6. Briefly, it is the angle around the horizon (Duffie and Beckman 2013). The SAA value increases in the clockwise direction (North = 0°, east = 90°, south=180°, west=270°).

Solar altitude or elevation angle (α): It is the angle between the line to the sun and the horizontal (Duffie and Beckman 2013). It is denoted by α and is given by:

$$\alpha = 90° - \theta z \qquad 1.3$$

It varies throughout the day, from 0° at sun rise and sunset to 90° when the sun is directly overhead.

I.2.C. Clear Sky Index (Kc)

For the clear-sky index calculation, a clear-sky model is required. A clear sky model is used to estimate the GHI under a cloudless sky, which called a clear-sky irradiation G_c; it includes information about the solar position and the cloudless atmospheric conditions. An overview of different models is presented by (Antonanzas-Torres et al. 2019) presented an overview of different models. Clear-sky models range from empirical models to radiative transfer-based calculations. Information on the state of the atmosphere is needed as input (Elke Lorenz, Hammer, and Heinemann 2004). (Beyer et al. 1996) introduced a clear sky model to better account for the irradiation dependence on solar zenith angle, atmospheric aerosol and water vapor content. One of the clear sky models is the European Solar Radiation Atlas (ESRA), (Rigollier, Bauer, and Wald 2000). It has been applied in forecasting solar studies due to its simplicity and robustness. Kc is defined as the ratio of irradiation measured by the ground station GHI to the irradiation of a defined clear-sky model G_c,

$$K_c = \frac{GHI}{G_c} \qquad 1.4$$

I.2.D. Clearness Index (Kt)

Describes the overall extinction of clouds and atmospheric compounds with extraterrestrial irradiation. The clearness index k_t is defined as the ratio of irradiation

measured by the ground station GHI to extraterrestrial irradiation (Top Of the Atmosphere hourly irradiation) G_{TOA}. The clearness index defined as follows:

$$K_t = \frac{GHI}{G_{TOA}} \qquad 1.5$$

Where G_{TOA} is the intensity (power) of the sun at the top of the Earth's atmosphere. Calculated as a function of solar geometry for the region and day of the year using a solar constant of 1366 W m^{-2}.

I.3. Scientific questions and objectives

Although ground-based irradiation measurements are the most reliable in terms of accuracy; yet they are expensive and have regular need for calibration. The number of measurement sites are still relatively scarce worldwide. They are known to be instruments that need regular calibration. Sub-optimal data quality control practices may significantly reduce the quality of these ground data. For these reasons, researchers often resort to the reanalysis of modeled irradiation data or remote-sensed data from instruments onboard geostationary satellites. In this regard, satellite-derived irradiance is the most promising data for real forecasting applications at the intra-day horizon. To best address our study's issue, the research questions and objectives of this thesis are proposed as follows:

Global Research question: What is the impact of using estimated solar irradiance from satellite data on intraday GHI forecasts in tropical environments?

As this section shows, the interest of our study is to test the robustness of different models using satellite-derived data in an ITZ. Consequently, the first research question arises:

Research question 1: Can statistical method using satellite-derived irradiance achieve comparable accuracy than statistical method using ground data in the ITZ?

Satellite-derived data have a broader coverage than ground-based irradiation measurement; yet, they give information only at the top of cloud processes. For information on the cloud depth and cloud overlap, multiple spectral channel must be considered which may not be available on onboard instruments. NWP models are dynamic models representing the whole atmosphere and a part of the Earth's surface. Therefore, the second research question that arises is

Research question 2: What impact do high-resolution satellite-derived irradiance have on improving GHI forecasts from a numerical model in French Guiana?

Nevertheless, these GHI forecasts must be improved as PV managers have financial incentives to produce more accurate forecasts of electricity production. The third major research question to be answered is

Research question 3: How the quality of GHI forecasts deriving from geostationary satellites can be improved in French Guiana?

Answering these questions will provide high spatiotemporal resolution solar irradiation data forecasting with good precision over a part of the South American continent,

considering the climatic specificity of this region. Consequently, three main scientific objectives are treated in this thesis. The first objective of this thesis is the following:

Objective 1: The relevance of using geostationary satellite data with different prediction methods to make solar intraday predictions in tropical environments.

A comparison of the results obtained with geostationary satellite data with those with measured station data using the same methods must be done. For that, we develop the problem according to:

- Statistical methods.

- Machine learning models.

Once this first objective is achieved, the next step is to quantify the impact on irradiation forecast models, leading to the second and third ones:

Objective 2: Improving predictions of numerical methods using satellite-derived irradiance while maintaining a high spatial resolution of these methods.

Objective 3: The improvement of satellite-based models in French Guiana using a hybrid model.

Three research questions, to be answered during the thesis, are raised. Moreover, three main objectives are defined. The overall objective of this thesis consists in implementing a solar irradiance forecasting method that uses geostationary satellite data and meteorological information from a weather prediction model available in French Guiana an ITZ.

I.4. Outline of the thesis

This thesis report is performed and presented gradually. It is structured in five chapters:

Chapter 1: it is an overview of the issue of energy production globally and particularly in our study area 'French Guiana'. The fundamentals of solar radiation, and the objectives of this work, are discussed.

Chapter 2: describes the data used in this study. The preprocessing protocol of ground-based irradiation measurement, geostationary satellite data, and numerical data is discussed.

Chapter 3: presents the state of the art of forecasting methods with an emphasis on their use in the context of solar energy forecasting is discussed. The bibliographic study made it possible to determine the areas of performance according to the spatial and temporal resolution and the forecast horizon. Classical algorithms and their framework of applicability are also presented here.

Chapter 4: deals with intra-day forecasts of global solar irradiation. This chapter presents the forecasting method performance. Statistical methods using ground-measured irradiance performance are compared to methods using satellite data as input. The first step is dedicated to forecast global solar irradiation using statistical methods and

satellite-derived data. Then, using four machine learning models and satellite-derived irradiance to forecast global solar irradiation over 12 months of 2016. The WRF model was configured considering the specificity of French Guiana environment to produce GHI forecasts. Then, the systematic error identified in the predictions made by the numerical model was corrected by the Kalman filter. This study is carried out over a period of three months in 2016. The development of solar energy systems requires solar radiation data sufficiently precise to simulate, design, manage and optimize the system's operation. We propose to improve the solar irradiation forecasts from statistical, machine learning and one hybrid physical statistical prediction model by linearly combining forecasts of five models with weights varying by models.

Lastly, Chapter 5: briefly presenting what is discussed in this report, discuss the results obtained and their potential applications and usability in the ITZ. Then, we conclude by developing perspectives aimed at ensuring the continuity of the work.

Publications

Parts of this research work have been presented in four publications in an international congress:

1) European Photovoltaic Solar Energy Conference and Exhibition-2019: Improving Hour Ahead Solar Irradiation Forecast Using Ensemble Method in French Guiana

2) European Photovoltaic Solar Energy Conference and Exhibition-2021: Day-Ahead Solar Irradiation Forecasting Using Model Output Statistics Trained with Satellite Data in French Guiana

3) International Conference on time series and forecasting-2021: Machine Learning for Very Short-term Solar Irradiation Forecasting in French Guiana

4) European Photovoltaic Solar Energy Conference and Exhibition-2021: Intraday Post-processing of Solar Irradiation Forecasts from WRF Model Using Satellite-Derived Data in French Guiana.

Chapter .II. Data collection and Data pre-processing

II.1. Introduction

The data obtained will be analyzed to generate a valid dataset for testing the models developed later. This chapter presents the data used throughout this manuscript and their treatments for modeling the solar radiation. It introduces the formalism, hypotheses, and approximations that were made.

II.2. Time series

In this manuscript, a time series refers to a finite series of data indexed by time. These data can be measured or simulated. Time series can be expressed in second, minute, hour, day, or year depending on the study case. Figure 9 illustrates a time series of hourly solar irradiation over île Royale station located in French Guiana. We assume that the time interval from ground-based irradiation measurement and numerical model is regular and that there is no modification in the acquisition protocol. Yet, for measured time series, it is difficult to obtain reliable series of measurements over long periods. This difficulty may be due to possible sensor drift or maintenance of the measurement installations.

The foundations of time series analysis are based on decomposition, analysis of subseries and then re-composition of the chronicle. This approach, by decomposition, assumes that the time series can be separated into simpler elements that can be modeled, in order then to be recomposed to give the forecast. Historical studies on the subject (Bourbonnais and Terraza 2010) have made it possible to standardize the decomposition of time series into three components:

- **Autocorrelation** represents the degree of similarity between time series at different instances. The time difference between the two instances is referred to as the time lag.
- **Seasonality,** the presence of variations occurs at specific regular intervals (less than a year: weekly, monthly, or quarterly). Seasonality can also be derived from an autocorrelation plot if it has a sinusoidal shape. Looking at the period gives the length of the season.
- **Stationarity,** an important characteristic of time series. A time series is considered stationary if its statistical properties do not change over time. In other words, it has a constant mean and variance over time, and a covariance independent of time.

Having accurate and long-term irradiation is a technical challenge, as discussed in section 2. Due to the various operational assumptions, the integrity of the dataset must be analyzed. Figure 1 and Figure 9 show two representations of the time series dataset. Figure 8 and Figure 9 show two representations of global horizontal solar radiation measurements at île Royale station, French Guiana. This representation allows having the first qualitative evaluation and flagging suspicious value that needs to be investigated.

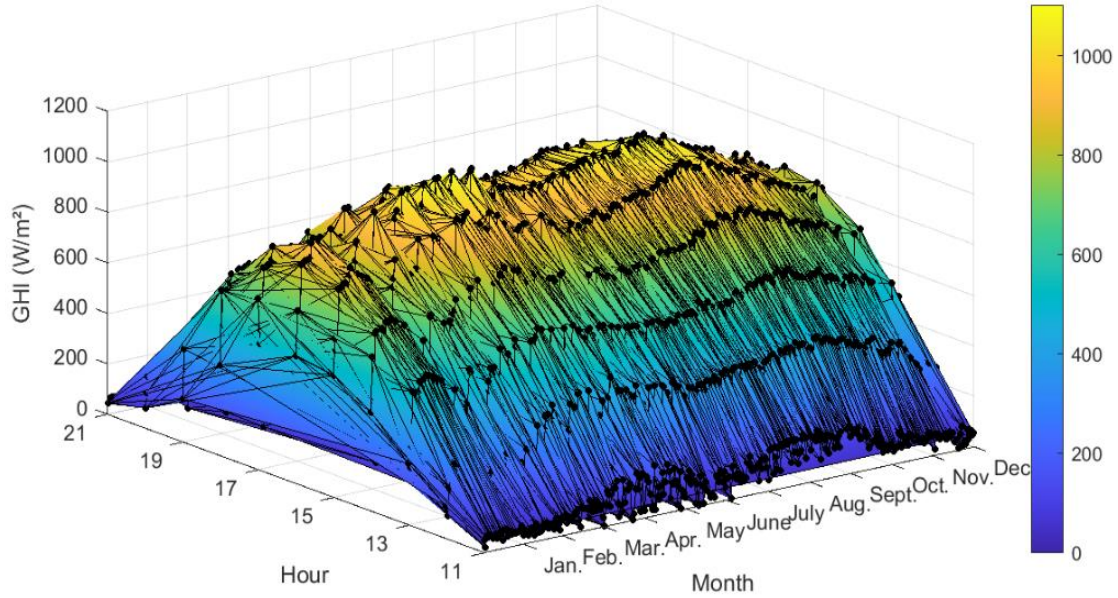


Figure 8: Graphic representation of the time series of global solar radiation measured to illustrate the frequency in station IR. Data acquired with an hourly resolution.

This overview of the measured data permits the first qualitative evaluation of the dataset and in particular, makes it possible to have an overview of the orders of magnitude, to judge the presence of "holes" and to evaluate the periodicity (double in the case of solar radiation: 12H and 366 days) Figure 9 with the presence of daily and annual intervals.

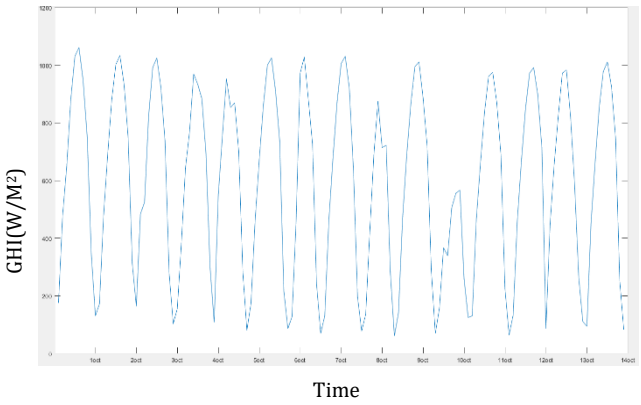


Figure 9: Graphic representation of the time series of global solar irradiation measured in IR to illustrate the daily frequency.

Indeed, there are different time scales: legal time (that of the clock time), universal time (Coordinated universal time UTC), Greenwich Mean Time (GMT), mean solar time... we have decided to index the data according to UTC time, Regarding French Guiana: local time = UTC - 3H. This makes it possible to standardize the scales of the data sets and to overcome the legal constraints specific to each country (Iqpal 1983). Further analysis of the time series will be described in detail in section II.4. The following part is devoted to the different data set used.

II.3. Data and Material

This section presents the datasets that were used to train and validate the models for assessing solar irradiation.

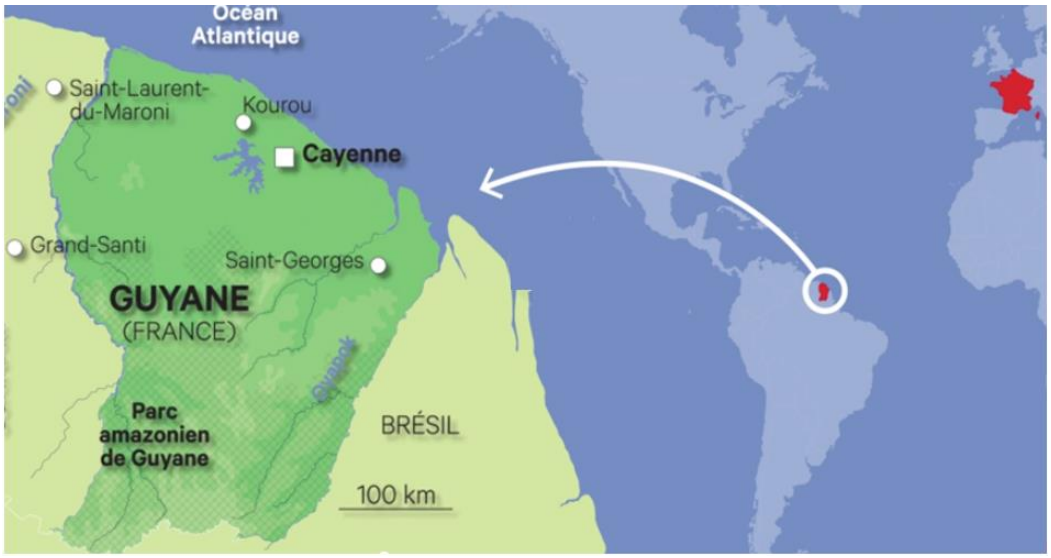


Figure 10: Political map of French Guiana in the world (Les Echos)

Our study focuses on French Guiana. This territory is located near the equator and is subjected to high cloud cover as part of a regular and important cycle of evaporation and precipitation due to the intertropical Convergence Zone (ICTZ) or monsoon trough. The ICTZ passes twice over the Guiana Shield, causing heavy rainfall and defining two main seasons: the dry season from July to October, when the ICTZ is located north of the Guiana Shield and the sky is mostly clear, and the rainy season from November to June, described in detail in section I.1.D.

II.3.A. Experimental sites

Measurements of global radiation used in this study are collected from six ground stations managed by the French national meteorological agency (Météo France). As of 2016, Météo France provides data of hourly global horizontal irradiation at each station.

II.3.A.i. Geographic properties

The geographical location of the different sites studied shows in Figure 11.

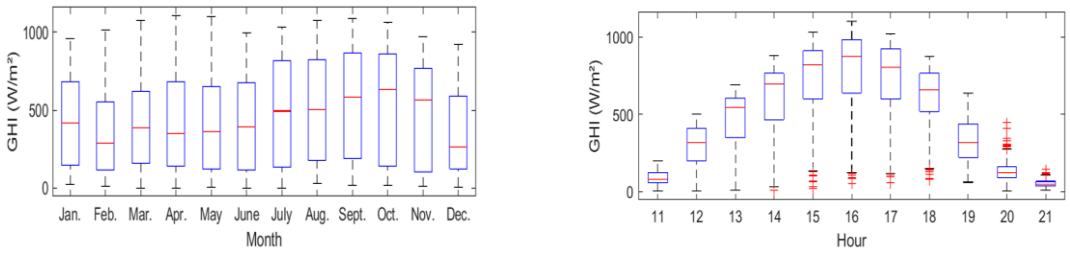
Figure 11: Meteorological stations in French Guiana (Salloum et al. 2019)

Three stations are located inland between 30 and 230 km from the coast to the interior of the country named, Maripasoula, Saint Laurent, Saint Georges. The other three stations are located on the Atlantic coast, namely: Kourou (situated on the coast), île Royale (situated about 7 km offshore from the coast), Rochambeau (situated about 13 km from the Atlantic Ocean). These six stations are characterized by different microclimates caused by their difference in heights. The list of the stations and their latitude, longitude and altitude is given in Table 2.

Table 2: Latitude, longitude and altitude of the ground meteorological stations understudy in French Guiana.

Station	***Location***	***Latitude***	***Longitude***	***Altitude***	***Ground Data points***	***Satellite Data points***
RO (Rochambeau)	Cayenne Felix Eboue Airport, Matoury	4.81	-52.37	4	18270	36540
SG (Saint-Georges)	Saint George of Oyapock	3.88	-51.80	6	19046	38092
MP (Maripasoula)	Maripasoula	3.63	-54.03	104	17373	34746
SL (Saint-Laurent)	Saint-Laurent-du-Maroni aerodrome	5.48	-53.90	4	18655	37310
KR (Kourou)	Guyanese Space Center	5.12	-52.44	12	16570	33140
IR (Ile Royale)	Kourou	5.28	-52.58	48	18156	36312

All of the ground stations are located in plain areas with effects. Their geographic positions are illustrated in Figure 11. Rochambeau and Saint Laurent stations are located at an altitude of 4 meters above sea level on a plain area. Saint Georges station is located at an altitude of 6 meters in the municipalities of the same name. Maripasoula station is located at an altitude of 104 meters on a plain area inside the territory, which is 230 km from the ocean. Kourou station is located at an altitude of 12 meters in the municipalities of the same name. The last station named île Royale is located at an altitude of 48 meters on the island of the same name located in the Atlantic Ocean 14 km from the coast of French Guiana.



a. By month *b. During the day*

Figure 12: Hourly GHI-measured data distribution for IR in 2016.

Figure 12, a) shows the distribution of ground values depending on the month for île Royale station in 2016. It can be observed that measured data have a similar tendency throughout the year. Comparably, Figure 12, b) exhibits the distribution of the data during the day for the same station in the same year. Data distribution shows its maximum values at midday and following the expected pattern, increasing in the mornings and decreasing in the afternoons.

II.3.A.ii. Ground-based irradiation measurement material

As of 2016, French weather services in French Guiana do not measure the direct normal (DNI) and the diffuse horizontal irradiation parts (DHI). Therefore, in this study, only the global horizontal (GHI) component of solar radiation is used. Different instruments can measure GHI. A pyranometer with a hemispherical (180°) view angle is the most common instrument used to measure GHI. The response of the pyranometer to a beam of light is proportional to the cosine of the beam's incident angle. Most pyranometers use a thermopile sensor to sense incoming light; GHI may also be measured using a photovoltaic reference cell, which will have spectral sensitivity and will not exhibit true cosine response.

In French Guiana, The weather stations are equipped with a pyranometer of CMP6 class B manufactured by Kipp & Zonen, Figure 13. This sensor is the first class pyranometer as defined by the World Meteorological Organization. It is suitable for measuring solar radiation on a plane surface (W/m2). Additionally, CMP11 also manufactured by Kipp & Zonen, is a secondary-standard pyranometer that monitors solar radiation for the full solar spectrum range and performs high-precision measurements with an uncertainty of less than 3% on daily cumulative radiation (World Meteorological Organization 2008).

Figure 13: CM6B pyrometer (WMO 2008)

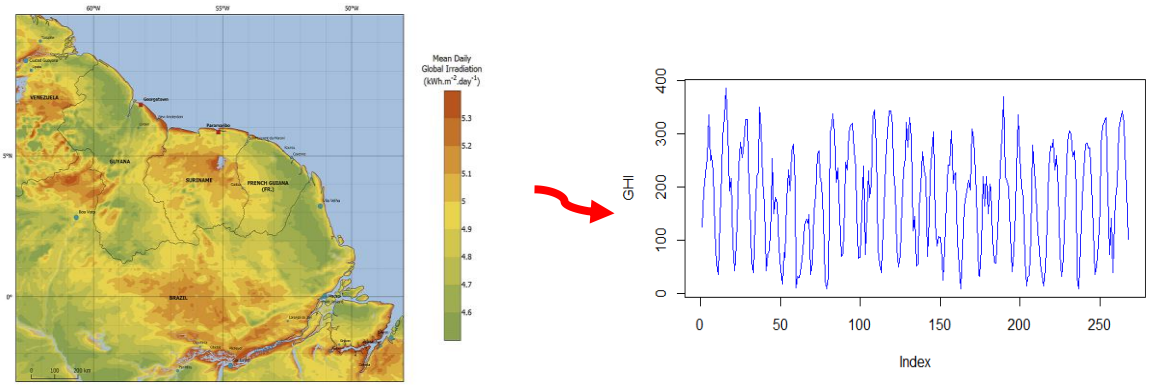
Both are equipped with a ventilation fan. Preventive maintenance is performed every two months. Standard exchange of the pyranometer is systematically performed every two years. Before its final installation, each pyranometer is calibrated in the Radiometry National Center of Météo France located in Carpentras (France). Once installed, the coefficients of the new pyranometer are then entered into the data acquisition system of the in situ stations (Albarelo 2017).

II.3.B. Satellite-derived data

Satellite images have broader coverage than ground-based irradiation measurements. Satellite images have many advantages, including:

- ✓ Continuous geographical coverage
- ✓ approx. 3+ km Spatial resolution
- ✓ 15 and 30 minutes Frequency of measurements
- ✓ Spatial and temporal consistency Calibration stability
- ✓ High availability (gaps are filled)
- ✓ Up to 27+ years history − variability of weather

In terms of forecasting, satellite-derived data are mainly used in two ways. The first approach directly uses the gridded data to generate areal forecasts. The second approach assimilates the satellite-derived irradiance into numerical weather models. Overview of past, current and future perspectives for estimating GHI from satellites in (Huang et al. 2019). A time series of images satellite used in this study are retrieved from the meteorological geostationary satellite GOES-13, orbiting at 74.5° W (longitude). 16-bit image data are collected from the visible channel (0.55 µm – 0.75 µm) and taken every 30 min (Fillol et al. 2017), with a spatial resolution of 0.6 km in longitude and 1 km in latitude, and a ground spot of approximately 1 km in diameter, more details can be found in section III.3.B.



Information extracted from the image

Creation of time series

Figure 14: Satellite-derived data time series.

The selected images are from January 2011 to December 2016 (6 years of satellite image). These images were provided by the National Oceanic and Atmospheric Administration (NOAA) via the Comprehensive Large Array-data Stewardship System (CLASS) catalog. The day was divided into half-hour slots, as suggested by (Rigollier, Bauer, and Wald 2000). The period of the day for this study lasts from slot 18 (8h45 UTC) to slot 45 (22h45 UTC). Satellite data related to ground elevation were also used in this study. The elevation map is obtained from the Shuttle Radar Topography Mission (SRTM) digital elevation model downloaded at a resolution of 90 m (Jarvis et al. 2008). Figure 14 shows the GHI map of the study zone.

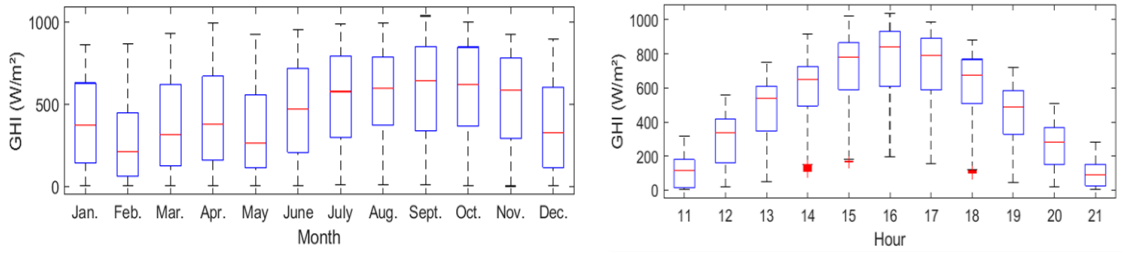
Regarding image acquisition, GHI prediction processing from satellite imagery can be divided into two main steps:

- Conversion: the raw values of the pixels of satellite image are calibrated into radiance values then converted into an estimates of GHI (Fillol et al. 2017).

- Forecast: This involves anticipating the movement of the cloud masses observed.

The main interest of geostationary satellite imagery for GHI forecasting lies in its ability to describe in time and space the cloud cover over large areas with a spatial resolution of 3 to 10 km and a resolution time of 15 min. Cloud cover represents the main factor of intraday variability of the GHI for most sites.

The distribution of satellite-derived data depending on the month for île Royale station for 2016 is shown in Figure 15, a).



a. By month

b. During the day

Figure 15: Hourly GHI satellite data distribution for IR in 2016.

Visually comparing measured and satellite data in Figure 12 with Figure 15, respectively. It can be observed that both data have a similar tendency throughout the year. Comparably, Figure 15, b) exhibits the distribution of satellite data during the day for the same station in the same year. Both measured and satellite sets behave alike for visual observation, showing their maximum values at midday and following the expected pattern, increasing in the mornings and decreasing in the afternoons.

II.3.C. Weather Research and Forecast Model (WRF) data

Weather and research forecast (WRF) is a numerical model that forecasts the state of a portion of the earth's atmosphere on a discretized grid. Sub-grid scale (i.e unresolved) physical process is approximated using physical parameterization. Radiative transfer, hence, the GHI is obtained using the radiative transfer parameterization scheme. For tropical locations, solar irradiance at time horizons of 12H or more can only be predicted with the help of numerical weather prediction (NWP) models. Numerical Weather Prediction (NWP) models simulate the temporal evolution of atmospheric processes important to the prediction of solar irradiation.

In this manuscript, the work of (Diallo 2019) to calibrate WRF is used to forecast GHI in French Guiana. where a set of 36H forecasts from WRF is used. The duration of WRF simulations was set to 48H with output data at 1H intervals. The first 12H of each simulation were discarded as coinciding with the model's spin-up period and evaluation was conducted using the remaining 36H of model data. Table 3 shows the physical parameterization used in this study.

Table 3: WRF physical parameterization (Diallo 2018)

Physical Parameterization	Scheme used
Planetary boundary layer scheme (PBL)	The local second-order MYNN3 (Nakanishi and Niino, 2009) scheme
Microphysics scheme (MP)	Thompson aerosol-aware (Thompson and Eidhammer, 2014)
Land Surface Scheme (LSM)	Noah (Tewari et al., 2004)
Cumulus Scheme (Cu)	(Grell and Dévényi, 2002) scheme
Radiation Scheme Longwave Radiations (RAD LW)	RRTMG (Iacono et al., 2008)
Radiation Scheme Shortwave Radiations (RAD SW)	RRTMG (Iacono et al., 2008)
Horizontal Resolution	1 km
Vertical Resolution	100 levels
Time Step	40 seconds
Radiative Effect	3 minutes
Numerical Weather Prediction (NWP) Model	Global Forecast System (GFS) model

Figure 16. shows the computation domain of WRF, only solar forecast of the domain D03 was considered.

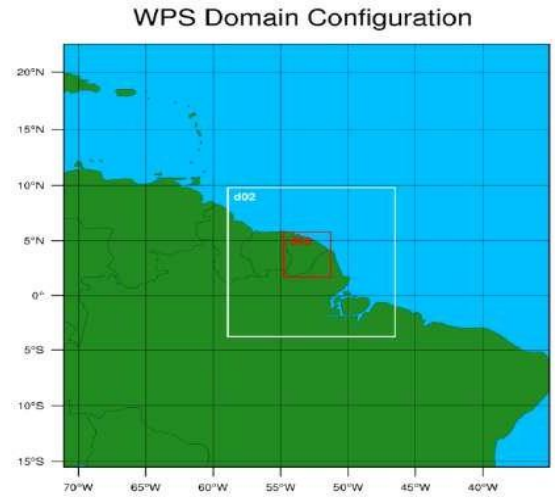


Figure 16: WRF nested domain ways, A large domain, the black square D01, with a resolution of 27 km, then the white one D02 with a horizontal resolution of 9 km the last domain is the red one D03 with a horizontal resolution of 3 km.

Three datasets used in this study have been presented, in the following paragraph, the different stages of the preprocessing of solar radiation data will be shown.

II.4. Data pre-processing

Before using the measured and simulated data to make a forecast, it is necessary to perform several preprocessing steps. The purpose of these steps is to make the data sets usable by the methods that we implement during the modeling. If the analyzed data comply with quality control, they can then be sent to the preprocessing to feed the models and make a forecast. Otherwise, they may, in certain cases, be corrected before use or even deleted.

Data pre-processing protocol will be presented step by step, responding to the problems encountered. We will start with quality control, management of missing data, then filtering procedure.

II.4.A. Missing data

It is common that the ground-based irradiation measurements or satellite images contain missing or faked data points, due to communication or sensor failures, or due to rejection by data quality control. Concerning the treatment of missing data in our study, in the case of missing hours, these data are not taken into account. If a day is missed, it is not taken into account for the study. The data are from January 2011 to December 2016. Data's gaps are shown in Table 4 for every site.

II.4.B. Quality control

The purpose of data quality control (QC, or quality check) is to verify their presence and physical consistency and to establish limits within data is to be accepted. Since the study is concerned with the quality of the global irradiation measured and estimated, only tests applied directly to both the GHI measured and the GHI estimated are presented in this section. Within the framework of the European Geosciences Union conference,(Espinar et al. 2012) produced a report based in particular on the work of

(Muneer and Fairooz 2002) as well as (Geiger et al. 2002) on the quality control of meteorological data. For our study, we performed a Quality Check process based on extreme values, following the method proposed by (Geiger et al. 2002) and the SoDa website. Hourly GHI (Wh.m-2) was considered valid when it respected the following condition:

$$G_{MIN,t} \leq GHI_t \leq \min[G_{max1}, G_{max2,t}] \quad (2.\ 1)$$

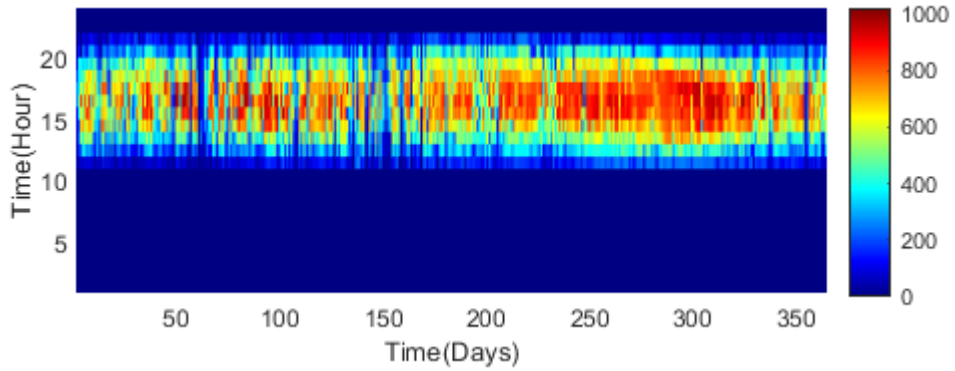
Where:

$G_{MIN,t} = 0.03 * G_{TOA,t}$

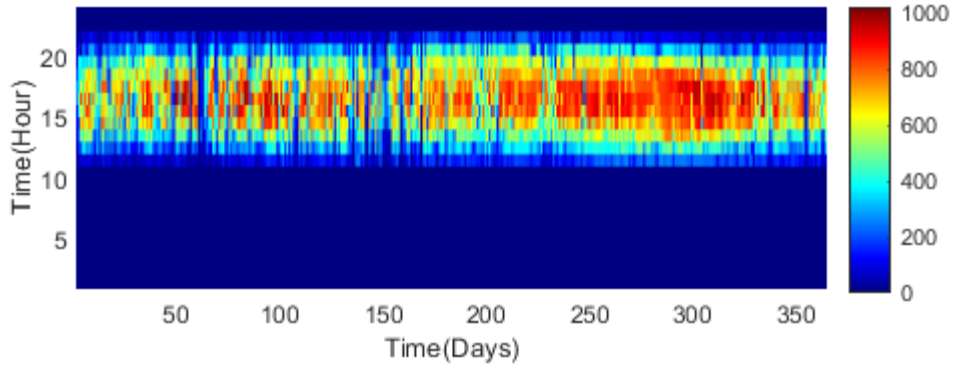
$G_{max1} = 1.2 * I_0$

$G_{max2,t} = 1.5 * I_0 * \cos(\theta_t)^{1.2} + 100$

Where the conditions of validity at time t are defined from the radiation outside the atmosphere GHI_{TOA}, and the solar zenith angle θ with the solar constant (I_0 =1367 W.m-2). A simple example in Figure 17, for one year at île Royale site, we get 4792 ground values, where 30 ground values were rejected by the QC test.



(a) measured data before QC test



(b) measured data after QC test

Figure 17: Global Solar radiation at île Royale site with and without QC.

Figure 17a and Figure 17b respectively present an overview of the overall radiation available on the île Royale site before and after the application of the QC test. The small difference between these two figures shows the high-quality of the measurements made at the experimental sites.

A final check of the data concerns their consistency with the time base. Table 4 gathers the results obtained after the data QC test, knowing that the total number of daytime points for the six years is 21960.

Table 4: Data and quality control

STATION	***Number of points***	***Outliers Data***	***Missing Data***	***Time Scale***
Rochambeau	18270	4%	16%	UTC-3
Saint Laurent	18652	4%	15%	UTC-3
Maripasoula	17373	5%	20%	UTC-3
Saint Georges	19046	3%	13%	UTC-3
Iles Royale	18153	5%	17%	UTC-3
Kourou	16563	5%	24%	UTC-3

The quality of the data has been verified, and we can continue preprocessing steps. which is to make the solar radiation data stationary after modeling solar radiation under clear sky conditions, this is the subject of the next paragraph.

II.4.C. Data filtering

Data filtration overcomes two important issues:

- The night hours, during which the solar irradiation is equal to 0, there is, therefore, no need to overload the models with all these very easily predictable data.

- Data measured at sunrise and sunset are not always reliable. The incident rays on the sensor arrive at a low zenith angle, which induces an error in the measurement; the response of the sensor is distorted by all the undesirable effects that occur in the measurement dome of the pyranometer, this measurement is even more degraded if a standard cell is used (Muneer and Fairooz 2002). We then consider that the solar height threshold, beyond which the measurement is considered correct is 6 ° (zenith solar angle = 84 °) (Muneer and Fairooz 2002). Additionally, for low solar heights, solar illumination can be disturbed by the direct environment of the place of the measurement: the presence of mountains, trees or buildings around the measurement station can generate errors. In the data related to the shadow of these obstacles.

In our study, we decided to apply refinement for a limit of 10 ° of solar height, (solar zenith angle = 80°) in order not to consider the night measurements.

With this refinement should remember that even with a low level of sunlight, the output of a high-power photovoltaic plant can be high. In addition to checking the quality of the measurements, it is important to identify the type of sunlight at the experimental sites. In this context, a solar classification procedure is put in place to characterize the sites.

II.5. Site characterization

This paragraph is devoted to the mathematical characterization of the meteorological conditions of the data. In addition to materialization, it is important to identify the type of sunshine at the experimental sites. In this context, a solar classification procedure is implemented to characterize the sites. To estimate, quantitatively, the degree of variation of the solar radiation data, a coefficient of variability will be calculated for each meteorological site studied. We will thus estimate the impact of this variability on the performance of our forecasts.

II.5.A. Classification of sky conditions

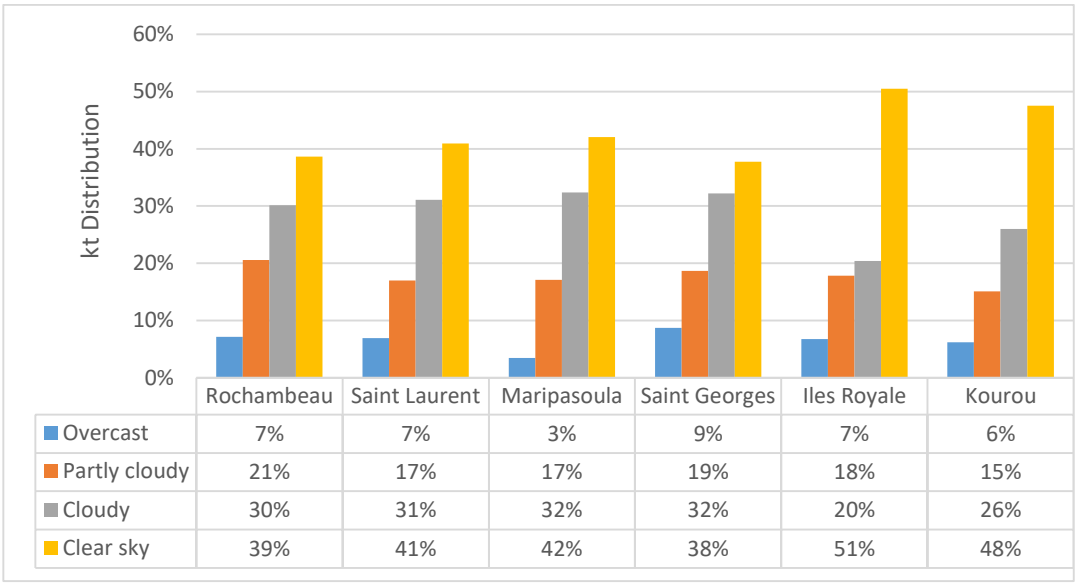
The classification of daily sky conditions is based on two criteria, namely, the type of sky and the variability. The daily mean value, $\bar{k}^*_{t,day}$ a day of the clear sky index k^*_t, defines the type of sky. In this context, four classes of day have been identified by (Fouilloy 2019):

- Clear Sky where the sky is clear and $\bar{k}^*_{t,day}$ is greater than 0.8, with 10%-30% of cloud in the sky.
- Cloudy Sky where the sky is mixed and $\bar{k}^*_{t,day}$ is from 0.6 to 0.8, with 30%-60% of cloud in the sky. Here, there will be more clouds than sunshine.
- Partly Cloudy where the sky is cloudy and $\bar{k}^*_{t,day}$ is between 0.2 and 0.6, with 60%-70% of cloud in the sky.
- Over Cast where the entire sky is cloud covered and $\bar{k}^*_{t,day}$ is less than 0.2 with 70%-90% of cloud in the sky.

In this subsection, the distribution of the different clear sky index classes over the 6 ground-based irradiation measurement stations was determined for 5 years' span. This involves computing the number of data for several intervals of clear sky indexes;

$$\bar{k}^*_{t,day} = \frac{1}{N}\sum_i k^*(t) \qquad (2.\ 2)$$

With N the number of daily measurements available after filtering and with respect to the solar zenith angle. We chose to distinguish four classes (Fouilloy 2019), from 0 to 0.25, from 0.25 to 0.5, from 0.5 to 0.75, from 0.75 to 1 while clear sky indexes greater than 1 (data for which the irradiation measured solar exceeds that calculated by clear sky) has been excluded. The classification of the daily sunshine of each site is illustrated in Figure 18 and summarized in the table.



	Rochambeau	Saint Laurent	Maripasoula	Saint Georges	Iles Royale	Kourou
Overcast	7%	7%	3%	9%	7%	6%
Partly cloudy	21%	17%	17%	19%	18%	15%
Cloudy	30%	31%	32%	32%	20%	26%
Clear sky	39%	41%	42%	38%	51%	48%

Figure 18: Distribution in percentage of measurements according to the quality of sunlight from 2011 to 2016.

If all days were governed by very low (clear sky) or very heavy (overcast) cloudiness, we would get a high percentage of data in the last or first class, and therefore would be representative of a low variability climate. In contrast, a very variable climate would see a more homogeneous distribution of the clear sky index in the different classes. It is therefore possible to relate the classification to a proportion of clear sky indexes with the variability of the data. First, all stations have a much more homogeneous distribution of clear sky indexes in the different classes, this means that the weather conditions on the six stations are similar. Second, the clear sky class [0.75; 1[is predominant, grouping between 40% and 50% of the data for each site, which means that these stations have more clear sky days than other types of the sky. In the next section, we try to find the relation between the variability within the meaning of the MALR and the distribution by clear sky index classes.

II.5.B. The variability of the data

To study weather characteristics on each site we quantify the variability of the time series for each one; such a study was carried out by (Voyant et al. 2015) who compared 20 parameters to qualify the variability of different datasets on several stations spread across the globe. They concluded that the most relevant parameter was the "mean absolute log return" (MALR), defined by Equation (2. 3):

$$mean(abs\ logr) = E[|\ln(K(t)) - \ln(K(t-1))|] \qquad (2.\ 3)$$

With E the Expected value and K the clear sky index for the component considered. The results obtained for each dataset are presented in Figure 19. To judge whether the site variability is high or low, we calculated the MALR for a time series of constant values (Const) and randomly generated values.

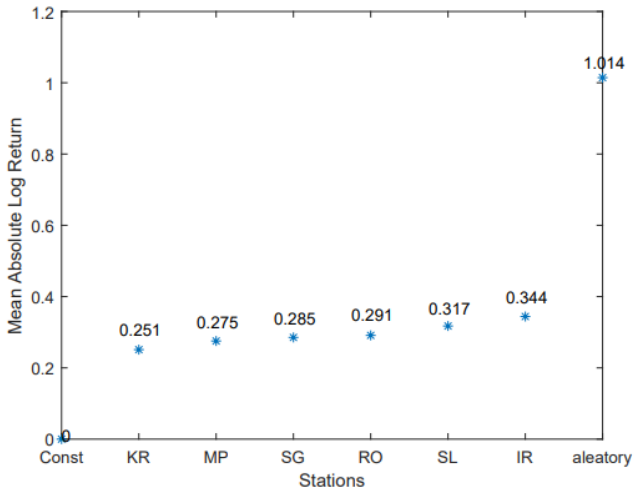


Figure 19: Mean Absolute Log Return for the different measurement locations.

In KR, the solar irradiation data are the least volatile, followed by MP, SG, RO then SL and finally IR. Figure 18 and Figure 19 show that for IR and KR, clear sky class is predominant for the two sites around 50% of the data; but we can note that IR has a distribution in the other three classes that is less homogeneous than KR, which may justify the fact that the variability within the meaning of the MALR is higher than for KR. Other stations have a much more homogeneous distribution of clear sky indexes, consequently, we can relate to the approximately logical way, the variability within the meaning of the MALR, and distribution by clear sky index classes.

It is possible to make an estimate of the variability (var) in percentage from MALRmin (constant series;0) and MALRmax (white noise; 1.014) using the following formula:

$$var\% = \frac{\text{MALR}}{\text{MALRmax} - \text{MALRmin}} * 100 \qquad (2.\ 4)$$

This allowed us to classify the datasets as in Table 5,

Table 5: variability calculations for different datasets.

	CTE	**KR**	**MP**	**SG**	**RO**	**SL**	**IR**	**ALEAT**
MALR	0	0,251	0,275	0,285	0,291	0,317	0,344	1.014
var	0	24,75	27,12	28,11	28,70	31,26	33,93	100

We have presented the data used, then studied the sites. Now, to evaluate the models' performance, we have at our disposal several tools, mathematics or graphics, which will be detailed in the following part.

II.6. Model evaluation

The purpose of the model evaluation step is to measure their ability to make a good forecast. It is important to perform this assessment using currently recognized criteria to have an objective assessment. In the case of modeling by models, whose relationships between inputs and outputs are very complex to interpret, we cannot use an uncertainty

propagation technique. The models are then evaluated, point by point, by comparing the results of the forecast with the corresponding measured variable.

It is difficult to assess the performance of models among themselves, specially to compare them with the results of the literature. There is a set of performance criteria to evaluate model accuracy in training and validation sets and to make comparisons between different models. According to (Sobri, Koohi-Kamali, and Rahim 2018), (David, Diagne, and Lauret 2012). Indeed, the data sets are very different, the periods and time steps of measurements are also different and finally, there are several error indications, which sometimes even, with the same name, have different definitions.

Error indicators are used to quantify the differences between the values obtained from the observation and the values obtained during the forecast. Two-time series are compared with each other, the first one is made up of the observations and the second one is made up of the corresponding forecasts. Since there is an error made in the observations (fitting, acquisition, etc.) we do not exactly assess the difference between forecast and reality, the difference is therefore assessed between forecasting and observing reality. Systematic error or bias is measured by some indicators we have selected the most widely used valuation indicators in the literature these metrics are defined as follows:

- Root Mean Squared Error (RMSE)

A measure of the standard deviation of the residuals. It corresponds to the square root of the Mean Squared Error (RMSE), computed as the average squared difference between the forecasted and measured values (MSE). RMSE tends to penalize large errors in a square order by giving them more weight, it is used when significant errors are undesirable:

$$RMSE = \sqrt{\frac{1}{N} * \sum_{i} (y_i - x_i)^2} \qquad (2.5)$$

Where x_i is the measured variable; y_i is the forecasted one that corresponds to ground measures; i is the time index, N is the total number of data.

For an evaluation of precision, MSE's does not have the same unit as that of the variable studied. It is, for this reason, the RMSE has the same dimension as the variables being studied. In the statistical field, the RMSE and its standardized version are widely used because they allow good readability as well as a good evaluation of the accuracy of the models. Authors do not always define RMSE, some authors divide RMSE by the maximum difference between the data, or even the maximum value from the data. In our case, we have chosen the "classic" version of the normalization for which we divide the RMSE by the mean of the data:

$$nRMSE = \frac{\sqrt{\frac{1}{N}\sum_i (y_i - x_i)^2}}{\frac{1}{N}\sum_i x_i} * 100 \qquad (2.6)$$

- Mean Absolute Error (MAE)

Measures of the average of the absolute errors between the forecasted and the measured values. MAE is less sensitive to large errors because all differences have equal weights. MAE combines systematic error and random error. It allows the accuracy of a model to be judged, in addition to that its unit is the same as the quantity measured

$$MAE = \frac{1}{N}\sum_i |y_i - x_i| \qquad (2.\ 7)$$

Its normalized form is obtained by dividing the MAE by the mean value of the measurements to obtain a percentage error:

$$nMAE = \frac{\sum_i |y_i - x_i|}{\sum_i x_i} * 100 \qquad (2.\ 8)$$

- Mean Bias Error (MBE)

A measure of the average bias in the forecasting. It is not a tool for measuring the accuracy of a model but only for indicating if the model overestimates or underestimates the actual value, the bias defines as follows:

$$MBE = \frac{1}{N}\sum_i (y_i - x_i) \qquad (2.\ 9)$$

We obtained relative values by normalization to the mean value of the ground-measured GHI for the considered period:

$$nMBE = \frac{\sum_i (y_i - x_i)}{\sum_i x_i} * 100 \qquad (2.\ 10)$$

For RMSE, MAE, MBE, Better predictions are indicated by smaller values (closer to zero), which indicates that both forecast and real values were similar.

- Adjusted R-square (aR^2)

Used to compare models with different numbers of independent variables. It should be used while selecting important predictors as input for the model.

$$aR^2 = 1 - [\frac{(1 - R^2) * (T - 1)}{(T - K - 1)}] \qquad (2.\ 11)$$

Where T represents the number de output variables and K represents the number of predictors, R^2 is the correlation coefficient that computes the correlation between the forecasted and the measured values.

$$R^2 = 1 - \frac{SSres}{SStot} \qquad (2.\ 12)$$

SSres represents the **S**um of **S**quares of **res**iduals and SStot represents the **tot**al **S**um of **S**quares. R^2 and aR^2 values closer to one, which indicates a perfect fit.

- Mean clear sky index (MK_C) and mean clear sky index variability (VK_C)

We use two indexes based on clear sky Index K_C (eq 1. 2 in Chapter .I) to quantify the GHI variability. The sky conditions were considered as clear when K_C> 0.65, cloudy when 0.4 <K_C< 0.65 and overcast when K_C< 0.4, (Aryaputera 2015).

The first index is the hourly variability of the clear sky Index VK_C (Marquez and Coimbra 2011) defined as (2.13). The second index is the hourly mean clear sky Index MK_C (Marquez and Coimbra 2011), defined as (2.14):

$$VK_c = \sqrt{\frac{1}{N}\sum_{t}^{t+N}(K_c[t+\Delta t) - K_c[t))^2} \quad (2.\ 13)$$

$$MK_c = \frac{1}{N}\sum_{t}^{t+N} K_c(k) \quad (2.\ 14)$$

Where dt is the time step of ground-based irradiation measurements, i.e. hourly. For a model with hourly GHI outputs N=1.

- Skill Score (SS)

The Forecast Score (FS) is an indicator that is increasingly used, it makes it possible to compare the performance of the model to a well-known reference model as the persistence model. It is interesting to identify very quickly whether a model is correctly configured or not. In the results phase, this type of parameter makes it possible to determine whether a complex model has much better performance than a simple naive model.

$$ss = 1 - \frac{MSE_{model}}{MSE_{reference}} \quad (2.\ 15)$$

A negative score indicates that the model is less good than the reference one, a zero score indicates that the model is as good. Furthermore, a positive score indicates that one is better than the reference model. Finally, a score of one means that the forecast is perfect.

We have chosen to limit the indicators to these because they are widely used in the field of forecasting and provide an accurate idea of the performance of the models.

II.7. Summary

In this chapter, we have presented the formalism of time series, which is particularly important because it is the basis of the simulations which will be conducted. We then detailed all the available data (ground-based irradiation measurement data, satellite-derived data, or numerical model data) as well as the meteorological characteristics of the measurement sites. We then listed data preparation (or preprocessing) phases with their quality control, management of missing data, and management of outliers.

Chapter .III. Methodology and forecasting models

III.1. Introduction

This chapter presents the different methods used for intraday GHI forecasting with satellite-derived data as observation instead of ground measurement one. Ten models were tested in this study. Ground measured data then satellite-derived data were used as observations to study the impact of using satellite observations instead of ground one. We will detail the models used to conduct this study in this section.

III.2. Solar global irradiation forecast

There are several ways to forecast solar irradiation. Forecasting methods Approaches, with their limitations and precision, can be found in the literature (Elliston and MacGill 2010); (Ahmed et al. 2020); (Heinemann 2006); (Diagne et al. 2013). Comparative studies of multiple approaches have assessed the accuracy of solar irradiation predictions (Mihalakakou, Santamouris, and Asimakopoulos 2000); (Perez et al. 2010); (Remund, Perez, and Lorenz 2009). The choice of one or another method depends on the user's needs and the forecast horizon. Indeed, all models do not have the same precision depending on the type of forecast desired, the location or the time horizon, or even the number of data available. In addition, the metrics used to estimate the error are often different, making it even more complex to compare models between them; some parameters such as correlation coefficient, root mean square error are often used, but not always suitable for comparing model performance. The period used to evaluate the precision varies considerably from one article to another: some of them estimate the precision of the model over a period of one or more years, others over a period of a few weeks introducing a potential seasonal bias. Under these conditions, it is not easy to make comparisons of the results presented, the various literature studies must therefore be carefully analyzed.

Existing solar forecasting methods can be categorized depending on the forecast horizons or on the selected approach to obtain solar irradiation forecasts as in Figure 20.

For classification based on the forecast horizon, we may distinguish:

- The very short-term (intra-hour), models based on sky imaging are used (Kurtz, Mejia, and Kleissl 2017), or even statistical models based on the time series formalities and the persistence of cloudiness. Other models based on satellite images can also be used in these forecast horizons.

- The short-term (intra-day), models use the time series formalities with machine learning methods or even satellite images (Elke Lorenz, Hammer, and Heinemann 2004). The forecast using satellite images is based on a cloud motion vector approach. They show good performance for the temporal range from 30 minutes up to 6 hours.

- The long term (day-ahead), based on the resolution of the equations of the atmosphere, called NWP for "Numerical Weather Prediction", are the most relevant (Perez et al. 2010). These forecasting models are sometimes associated with post-processing modules to improve the NWP forecast at a short time horizon (intra-day). The NWPs are supplemented with information from satellites (Lara-Fanego et al. 2012).

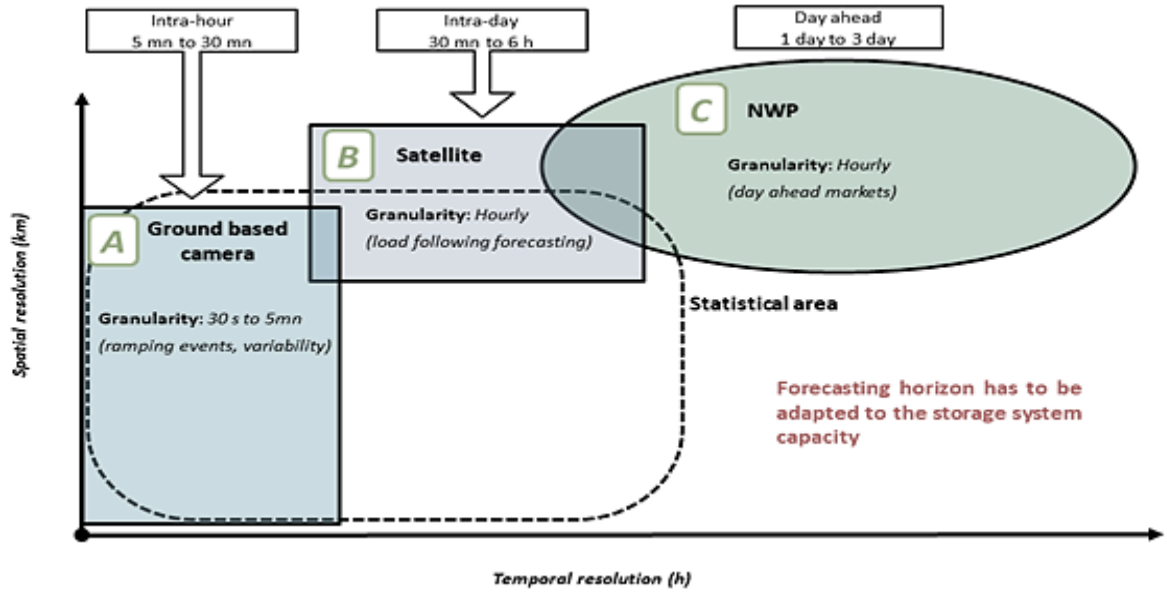


Figure 20: Temporal resolution and time horizon for the different forecasting methods.

For classification based on the approach used for deriving solar irradiation, there are three main methods (Gupta, Gupta, and Saroha 2021):

- Statistical method, analyzes the trend of a time series to assess future patterns. They required long-term and high-quality observations over the area under study to train the model. Satellite images offer the advantages of providing long-term irradiation observations; yet, its sensor needs to be calibrated and their accuracy decreases with time. Examples of direct time series models are Auto-Regressive (AR) and Auto Regressive Moving Average (ARMA) models. Furthermore, artificial neural networks (ANN), (Prado 2020), (Voyant et al. 2013).
- Physical method, based on NWP for "Numerical Weather Prediction" and or radiative transfer models, as in (Mathiesen and Kleissl 2011), (Diallo 2018). These forecasting models are sometimes associated with post-processing modules and are supplemented with information from satellites (Lara-Fanego et al., 2012).
- The hybrid method combines both physical and statistical methods. It gets an optimized forecast more specific and more accurate than each member of the models as in (Ren, Suganthan, and Srikanth 2015), (Du et al. 2018).

These methods are illustrated in Figure 21. We focus in Table 6 on solar forecasting methods using satellite data. We found only a few studies focused on obtaining GHI forecasts in the ITZ without the use of ground-based irradiation measurements.

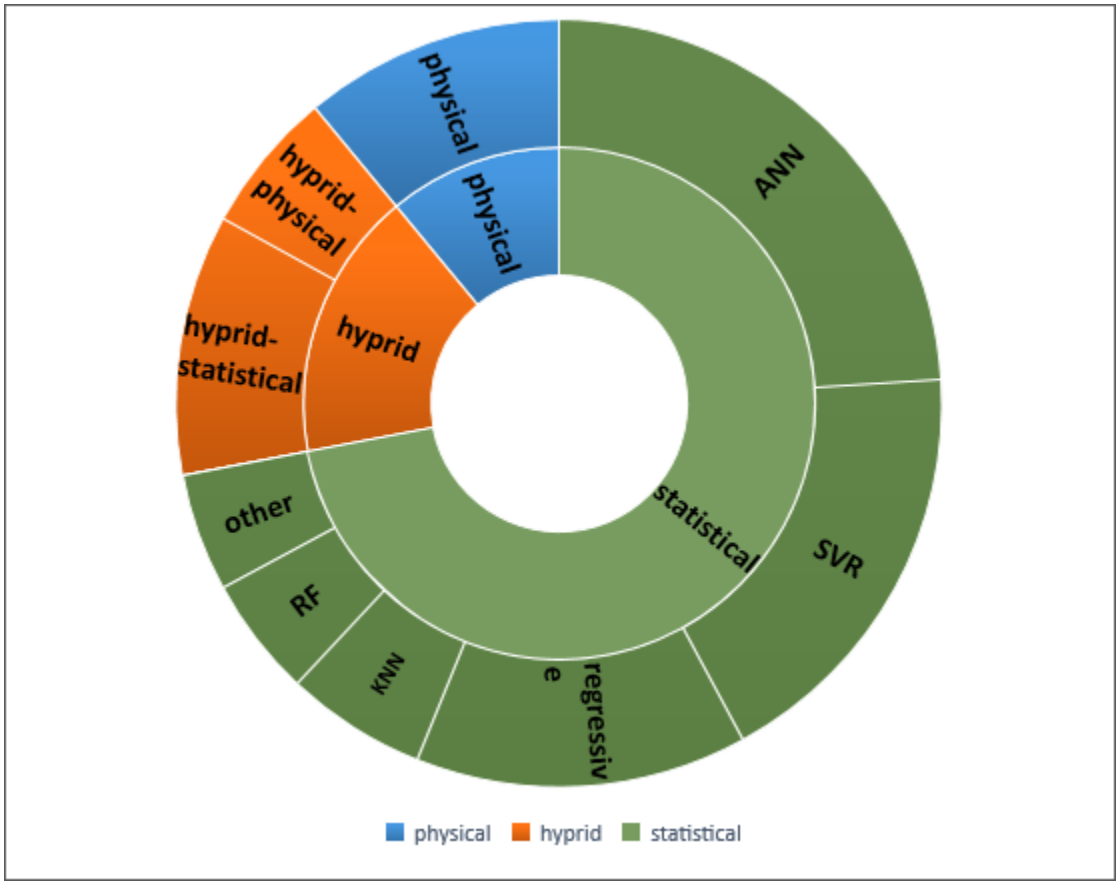


Figure 21: Distribution of studies with respect to the technique used.

Solar forecasting methods using satellite data and measured one are presented in Table 6. The highlighted lines expose the studies where only satellite data was used to make the forecast.

Table 6: Information revealed by the Non-exhaustive literature review of forecasting solar irradiation using satellite data until 2020.

AUTEUR	*LOCATION*	*HORIZON*	*SATELLITE*	*RESULTS*	
(Miller et al. 2018)	USA	Intra-day	Goes	-	Errors = 8.5-17.2%.
(Mazorra Aguiar et al. 2019)	Canary island	Intra-day	Helioclim-3	-	ANN (satellite pixels + ghi_{past}) > ANN (ghi_{past}) > smart persistence model.
(Kallio-Myers et al. 2020).	Finland,	4 h	Solis-heliosat	-	Solis-heliosat > persistence.
(Rodríguez-Benítez et al. 2020)	Spain	Intra-day		-	rRMSE = 25% to 70%.
(Voyant 2014)	Corsica island	1h	Helioclim-3	-	nRMSE (scaled persistence and ANN) = 16.5% .
(Marquez, Pedro, and Coimbra 2013)	USA	Intra-hour Intra-day	_	-	Improvements over persistence $_{(h=1)}$ = 5-19% and 10-25% for multiple time-step forecasts.
(André et al. 2019)	Caribbean islands	1h	_	-	Spatio-temporal vector autoregressive > cloud motion vector.

(Cornejo-Bueno et al. 2019)	Spain	_	Heliosat-2,	-	Extreme learning machines > heliosat-2 and cams.
(Laguarda et al. 2020)	South America	_	heliosat-4 method (goes-east satellite)	-	rRMSD $_{mcclear}$ = 2.8%.
				-	In the presence of clouds: model $_{clear\text{-}sky}$ < ESRA $_{clear\text{-}sky.}$
(Yagli 2020)	**USA**	**Intra-hour Intra-day**	**National solar radiation database**	-	**Improves the forecast accuracy.**
				-	**Leads to high-quality predictive distributions.**
(Bright 2019).	several	1h	Solcast website	-	$rRMSE_{clear=}$3.4%, $rRMSE_{clouded}$= 25.6%, $rRMSE_{all\text{-}sky}$= 16.9 .
(Guijo-Rubio 2020)	Spain	_	_	-	Sigmoid unit-product unit with evolutionary training > SVR and > extreme-learning machines.
(Dong et al. 2014)	Singapore	1h	_	-	The hybrid model has better performance than other forecasting models.
(Aguiar et al. 2016)	Gran canary island.	Intra-hour	Helioclim-3	-	(NN + ECMWF + SAT) > (NN +SAT) > (NN +ECMWF).
(Mazorra Aguiar et al. 2015)	Gran canary island.	Intra-day	Helioclim-3	-	(ANN + Gnd + SAT) > (ANN + Gnd).
(Yagli 2020)	**several**	**1h**	**_**	-	**Forecasts generated using bias-corrected satellite-derived data > ground-based data.**
(Alonso-Suárez et al. 2020)	South America.	Intra-hour Intra-day	NOAA's surfrad solar radiation network	-	The addition of satellite information further improves the quality of the probabilistic forecasts.
(Marchesoni-Acland and Alonso-Suárez 2020)	South America	Intra-day	_	-	Averaging window size is an important parameter.
				-	Satellite lags are of limited utility and spatial averages are more useful than weighted time averages.
(Singh Doorga et al. 2019)	Island of Mauritius	5 days	Eumetsat's geostationary meteosat satellites	-	Double exponential smoothing model > ARMA and > NAR-neural network.
(Deo 2017)	Queensland	Monthly seasonaly	_	-	ANN > MLR and > ARIMA.
(Deo 2019)	Australia	_	Moderate resolution imaging spectroradiometer	-	Universally trained ELM > RF, M5 tree, and MARS.
				-	Models register a legates & mccabe's index of ELM = 0.555–0.896, RF = 0.411–0.858, M5 tree = 0.434–0.811, MARS = 0.113–0.868.

				- $rRMSE_{ELM}$ = 3.715–7.191% vs. $rRMSE_{RF}$ = 4.907–10.784%, $rRMSE_{M5tree}$ = 7.111–11.169%, $rRMSE_{MARS}$= 4.591–18.344%.
(H. Jiang et al. 2019)	China	_	Multi-functional Transport Satellite (MTSAT-1R)	- $RMSE_{(spatial\ pattern,\ point\ information)}$ of - Hourly = 84.18 w/m2, daily total= 0.30 mj/m2, monthly total = 1.92 mj/m2 and scales= 1.08 mj/m2.
(H. Jiang et al. 2020)	China	_	Multi-functional Transport Satellite (MTSAT-1R)	- Exploring spatial scale effects using remote sensing data.
(Verbois 2020)	Singapore	_	_	- A holistic approach to evaluate forecasts instead of RMSE.
(Verbois et al. 2018)	Singapore	Day-ahead	Global Forecasting System (GFS)	- (WRF + statistical learning method) > smart persistence and > a climatological forecast and GFS. - RMSE 23% lower than smart persistence.
(Mazorra Aguiar et al. 2019)	Canary islands		Cm saf and mcclear model	- Linear Regression > clear and > cloudy sky conditions.
(Huva, Verbois, and Walsh 2020)	Singapore	Day-ahead	_	- rRMSE (post-processing 4DVAR) = 37%. - Models + post-processing > persistence ensemble and > climatological references.
(Zambrano and Giraldo 2020)	Colombia	Intra-day Day-ahead	_	- Learned metric > measurements from the whole set of available sites.
(Yeom et al. 2020)	North Korea	_	_	- Pyranometer and satellite-sourced solar radiation. - Instantaneous: RMSE=87.90 $W.m^2$, MBE=16.84 $W.m^2$. - Daily 'all sky conditions' RMSE= 624.98 $Wh.m^2$,MBE=13.89 $Wh.m^2$.
(Salazar et al. 2020)	Brazil	Intra-hour monthly	_	- Hourly: cams> MERRA-2. - Long-term mean-monthly: with cams, ceres and NASA-power > MERRA-2.
(Benamrou et al. 2020)	Morocco	1 h	_	- Combining the ground measurement and the most relevant surrounding satellite-derived GHI improve forecasts for time lags 1.2.3 and 4 hours.
(He et al. 2020)	China	_	_	- Input meteorological factors combinations were different in the four different climatic zones. - Sunshine hours, extraterrestrial radiation, and air temperature → greater impacts. Wind speed had little influence on solar radiation estimation.

				- SVM > angstrom-prescott formula and > multiple linear regression method.
(Olomiyesan and Oyedum 2016)	Nigeria	_	_	- *R*2 = 0.922 to 0.961.
(J. Thorey et al. 2015)	France	Intra-day	_	- Tigge ensembles are under-dispersed but rather different from one to another. - Aggregation decreases the forecast error by 20%, with a more realistic spatial pattern of predicted irradiation.

This bibliographic study provided us a piece of useful information on the various models used with satellite data. The achieved state of the art then made it possible to orient our study towards the models chosen in this study. GHI forecasting approaches may be categorized according to the input data used, which also determine the forecast horizon (Diagne et al. 2013).

In the following part, we present the different methods for forecasting solar irradiation. Additionally, we will detail the state of the art carried out to orient our work.

III.3. Thesis methodologies

Some models are more widely developed than others in solar forecasting. Nevertheless, some of the slightly used methods are still interesting to study. In this section, we will present different approaches to forecast solar irradiation that have been considered in this thesis with their advantages and disadvantages. We start with statistical methods in section III.3.A. then describes sky imagery and satellite-based method in section III.3.B. We proceed in section III.3.C. with the description of the mesoscale weather prediction model and conclude in III.3.D. with hybrid methods. Section III.4. summaries of the approaches selected for this thesis and the reason for those choices.

III.3.A. Statistical method

Methods based on linear regression (ARMA, ARIMA, SARIMA) are widely used, although they are slightly less efficient than methods based on learning prediction, they can be recommended in some cases. Learning models have also been widely used in the field of solar irradiation prediction, they have been studied in many regions of the world and researchers have demonstrated the ability of these techniques to achieve a prediction based on the use of time series (Benamrou et al. 2020) (Lauret et al. 2015) (Yagli 2019). Forecasting methods based on the statistical method fall into two categories: linear models and non-linear ones (Diagne et al. 2013).

III.3.A.i. Linear models

➢ *Persistence reference P model*

The most common reference model and simplest way to perform in the solar forecasting community for short-term forecasting is the persistence model. The persistence estimates the model variability between t and $t + \Delta t$, This last point is the key for time series forecasting (Diagne et al. 2013). Baseline forecasts with the persistence model indicate quickly whether you can do significantly better or not. So this model assumes that the global irradiation value at time t is equal to the global irradiation value at time $t + \Delta t$. For the comparative study equation (3. 1) was used to predict the GHI values with $\Delta t(1, 2, 3 \ldots . hours$ ahead):

$$\hat{X}^P(t + \Delta t) = X(t) \qquad (3.\ 1)$$

where $\hat{X}^P(t + \Delta t)$ is the forecasted solar irradiation at time t+Δt. Persistence forecast accuracy decreases with forecast horizon according to the evolution of cloud cover. Persistence should be used only as a baseline forecast for comparison to more advanced techniques (Diagne et al. 2013).

➢ *Scaled Persistence SP model*

The Persistence model is widely used due to the persistence of weather conditions. This model can be improved by coupling with a knowledge model taking into account the variation in the position of the sun and therefore in the angle of incidence. This improvement, easy to implement, considerably improves the results compared to persistence. The Scaled persistence model (Voyant, Motte, et al. 2017), using a clear sky solar radiation model for any horizon Δt, is defined by

$$\hat{X}^{SP}(t + \Delta t) = \frac{\hat{X}(t).\, Xcs\ (t\ +\ \Delta t)}{Xcs\ (t)} \qquad (3.\ 2)$$

Where Xcs is the clear sky model irradiation. This model allows making the persistence technique on the clear sky index. This type of predictor is sometimes the only one that can be used in an operational mode because it does not require historical data and can therefore make a prediction even in the case of measurement faults.

➢ *Auto-Regressive AR model*

Auto-Regressive model assumes that the future value of a variable is a linear combination of past time series of this variable. The AR modeling process is a linear model and the AR model for solar forecast can be described by the following equation:

$$\hat{X}^{AR}(t) = \sum_{i=1}^{i=p} \varphi(i) * x(t - i) + \omega(t) \qquad (3.\ 3)$$

where p is a positive integer and φ with (i = 0,1,…p) are coefficients, while ω is a white noise with mean zero. Here the future value obtained by a linear combination of past observations. The model parameters are φ with (i = 0,1,…p) and p the order of AR model.

The estimation of the parameters φ in the equation (3. 3) is obtained using least-squares(performs linear regression analysis using the least-squares method to find a line from the observed values), for more details see Refs (Singh and Pozo 2019). Additionally, due to the geographical differences, each location corresponds to its own unique model. Consequently, there is an interest to make a selection of the optimal order p. The p order univariate AR (p) models, is determined by information criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criteria (BIC), (G.Schwarz 1978); (Tsay 1984). The order of AR model is optimized using the auto-mutual information factor (Rossi et al. 2006). Results in Table 7 show that the correlation between values at time t and the previous value are stronger than the correlation with the second or third previous value.

Table 7: ACF results on data from 2011 to 2015, all stations grouped

CORR	x_t
x_t	1
x_{t-1}	0,783843
x_{t-2}	0,39928028
x_{t-3}	-0,02368415

For the AR model, we chose an order of (1) which has a 78% correlation.

➢ *Autoregressive Moving Average (ARMA) model*

Known as the Box–Jenkins model (1976), used for forecasting stationary time series. It can be used in fields that deal with a large amount of observed data from the past. It is a linear model used in many fields and more particularly in the field of solar radiation forecasting.

The ARMA model is developed using the combinations of two models, Auto-Regressive and Moving Average (MA). The first one assumes that the future values are individually correlated with previous values up to a certain parameter p, see (3. 4). The second assumes that a time series variable may be related to itself by a moving average process with a certain window size q, where the letter is given by:

$$\hat{X}^{MA}(t) = \sum_{i=1}^{q} \theta(i) * \varepsilon(t-i) \qquad (3.\ 5)$$

Then, ARMA(p , q) model is developed using equations (3. 3) and (3. 5) as:

$$\hat{X}^{ARMA}(t) = \omega(t) + \sum_{i=1}^{p} \varphi(i) * x(t-i) + \sum_{i=1}^{q} \theta(i) * \varepsilon(t-i) \qquad (3.6)$$

Where $\varphi\ (i)$ and $\theta\ (i)$ are the model parameters, p and q are the orders and ε the residue, which is the white noise that produces random uncorrelated variables with zero mean and constant variance. While ω is white noise with mean zero. This is if the model is well parameterized, and the time series is stationary. The ARMA model is therefore a model based on a combination of previous measurements and errors to characterize the current data. In the optimization phase of this model, the orders of the model are determined, where the model parameters (φ and θ) are determined during the learning phase (Stoica

and Selen 2004). We limit p, q ≤ 10 to simplify the process. Typically, this method requires a large amount of historical data to obtain the ARMA model.

This model is widely used in the analysis of time series, and it is quick to implement. There are several developments in this type of model, such as ARMAX, SARMA, SARIMA ... The interested reader can refer to the work by Box and Jenkins on the analysis of time series (Box 1970).

MATLAB simulations were conducted to obtain the ARMA model. Table 8 presents the values of the orders and coefficients for the ARMA model.

Table 8: The realized ARMA model.

p	q	φ	θ
1	2	$\varphi_1 = 1.142$	$\theta_1 = -0.2$ $\theta_2 = 0.18$

The advantages of using the previous models are the simplicity of their implementation, fast calculation time, cost-effectiveness, and accuracy of forecasting in time.

III.3.A.ii. Nonlinear models: Machine Learning methods

Machine learning methods can be used for classification or regression tasks where they have proven their efficiency. The use of machine learning is now widespread in many fields. This type of approach allows solving problems, which are impossible to represent by explicit algorithms. Machine learning models can find relationships between inputs and outputs, even if the representation is impossible, making them particularly suitable for forecasting tasks. First, these models are calibrated using a training dataset, then can be used to predict the result of test data. These algorithms can be divided into different categories (Fan et al. 2019):

- Supervised learning establishes connections, where the goal is to learn a general rule that matches inputs to outputs. That will be by training the system in the context of successive calculations with different inputs and outputs
- Unsupervised learning, which is mainly used for clustering, and is applied to input data without predefined target values. In other words, the model can find in its inputs a hidden structure without knowing the corresponding results.
- Group learning, in this model, it is necessary to train several so-called "weak" learners as members of a larger whole. Their predictions are then combined into a single result to achieve better performance.

Several approximately complex models were used to make the forecast. Deep learning, which is a branch of machine learning based on a set of algorithms modeling high-level abstractions in data using particularly complex model architectures, composed of multiple non-linear transformations, is not taken into account in this study. As GHI forecasting time-series can be framed as a supervised learning problem, supervised learning models are chosen in this study. This section concerns models for which a learning phase is necessary before they can be used. It is necessary to optimally choose

the data that will be provided to the model so that it is properly configured. Three basic steps must be followed in the process of creating a machine learning model:

- Training phase: allows the model to know its parameters.

- Validation phase: model selection is conducted using this measure that monitor the performance.

- Testing phase: allows to test the performance of a trained model on data that have never been involved in learning or testing.

Therefore, it is imperative that the test base includes sufficient different items. Inputs are matched with their target value, which corresponds to the value we are trying to predict (GHI in this study). We collected a set of meteorological variables to be used next to GHI in machine learning multiinputs models, considering the date, hour, and forecast horizon. To determine the correlation between variables we used Pearson's linear correlation coefficients, Table 9, which is known as the best method for measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \qquad (3.\ 7)$$

r= correlation coefficient, x_i values of the x-variable in a sample, $\bar{x}$ the mean of the values of the x-variable.

Table 9: Pearson's linear correlation coefficients between pairs of GHI and each variable of measured data (G), satellite data (S)

VARIABLE		GHI(t-1)	Kc	Kt	Gc	GTOA	ASA	SZA	elv	cos(SZA)
GHI	G	0,732	0,522	0,449	0,552	0,669	-0,075	-0,709	0,709	0,708
	S	0,748	0,513	0,499	0,552	0,669	-0,075	-0,709	0,709	0,708

Table 9 shows that the most correlation is between GHI and GHI(t-1), Solar Zenith Angle, Elevation angle, $G_{TOA,}$ and G_C clear sky model irradiation for ground and satellite data over the five years from 2011 to 2015 all stations grouped. Feature Selection and Principal Component Analysis (PCA) are performed to identify and select only the relevant features that lead to accurate forecasting. PCA is a technique transforms a set of observations of probably correlated variables into a set of values of linearly uncorrelated variables which are called principal components. Four components from seventeen different values that explained 99% of the variance, and two categorical variables: month, and hour, are kept from twelve different inputs and are considered to make the predictions.

After Exploring the optimal input set configuration, the next step is to test the accuracy of the proposed machine learning method for solar forecasting, different tests and validation methods are exploited. K-fold cross-validation is the most prevalent test method used in recent studies (Rohani, Taki, and Abdollahpour 2018); (P. Jiang and Chen 2016). Cross-validation consists of splitting the data repetitively in pairs of train and test sets, called 'folds' to protect against overfitting. As depicted in Figure 22, in k-fold cross-validation the whole data set is split into k folds: at each iteration from 1 to k, one-fold is used as the

testing set and k-1 folds as the training set, until all folds used to build the forecast model, k range between 3 and 10 are the typical values, 5-fold cross-validation was selected as a validation option in this study.

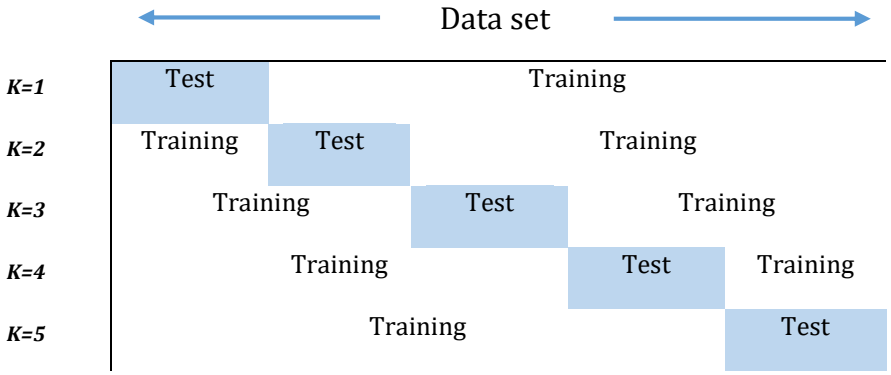


Figure 22: Diagram of k-fold cross-validation with k=5; the blue boxes refer to testing data and the white to training one.

The hyperparameters of each model have been estimated from 12 hourly ground and satellite training data each during 5 years. Models were trained using ground and satellite data separately. The best hyper-parameter set is selected as the following three steps:

- Building several models with different hyper-parameter combinations using a subset of data from the training set.
- Evaluating the models using the unused subset of data from the training set (validation set).
- Choosing the model with the lowest error metric on the validation set.

MATLAB Machine Learning Toolbox was used to train and validate machine learning forecasting models. For this study, we selected four models, namely, Support Vector Machines (SVM), Gaussian Process Regression (GPR) and Regression Trees (TRE, ENsTRE) models. The input data are taken from the database and loaded into MATLAB according to the case and time horizon. Before feeding the models with the inputs, datasets are standardized; in other words, rescaled to have a mean zero and a standard deviation of one; so that the different ranges of the features do not affect the contribution of each one.

➢ *Simple Regression Tree (TRE)*

It can be used for classification and regression. A decision-tree machine-learning algorithm is a type of supervised machine learning, where we give the input and what the corresponding output is in the training data. This algorithm automatically grows a decision tree by splitting each node on an optimal input variable. It stops splitting the nodes when no significant gain is obtained.

Regression trees are an evolution of decision trees, usable in many fields of data processing. A decision tree is a series of choices that lead to a final decision. Graphically, the process is represented by a binary tree, each choice being between several branches and the decisions will be the leaves. This type of method, started in the 1960s, is part of recursive partitioning methods. Formalism as we know it today was carried out by

(Breiman et. al 1984) under the abbreviation CART: "Classification and Regression Tree". The two forms of this modeling qualitative (classification) and quantitative (regression) is compiled. This type of model can be used for forecasting in different fields (De'ath 2007); (Troncoso et al. 2015); (Tso and Yau 2007). There are several variations of regression trees. The first one used to deal with regression problems is classical regression tree (Hastie, T., Tibshirani, R 1986) proposed the first formalism of this regressive model, the mathematical formula is

$$\hat{X}^{RT}\ (t\ +\ \Delta t)\ =\ \sum_{i=1}^{\boldsymbol{t-1}} k_i * H\ (x\ (t\ -\ i)) \tag{3. 8}$$

Where k_i is a factor, H is a function that returns 1 if the data are used otherwise it returns 0. Once the tree is constructed, a regression model is applied to each node. When learning phase, an iterative process, we try to minimize the error between the value at time t and the same predicted value.

➢ *Ensembles of Trees 'Bagging' (ENsTRE)*

The term bagging is an abbreviation for "bootstrap aggregating". Bagging is another improvement level in forecasting models and especially regression trees. This method consists of generating regression trees on samples from the dataset, but unlike the boosting method, successive trees do not depend on responses from previous trees. where each tree is built using a bootstrap sample of the dataset. Bootstrap means created a new sample from the original dataset. Each sample is constructed by drawing with replacement, this is called resampling. The Gaussian assumption on the distribution of samples is not necessary. When the trees have been generated and the corresponding answers are available, a simple majority vote is used to make the forecast. This overall method gives significant improvements in the processing of data concerning complex phenomena whose input-output relationships are difficult to understand. We determined that the optimal number of leaf size is 30 from the training phase. Naturally, this step will have to be conducted as soon as we change the input data.

➢ *Gaussian Process Regression (GPR)*

Gaussian processes are models whose development is quiet (Rasmussen 2006). This type of model is, in fact, a generalization of a multivariate Gaussian (or normal) distribution. These are nonlinear models. The interested reader will refer (Lauret, David, and Calogine 2012). For a univariate forecast, the mathematical formulation of the models is as follows:

$$\hat{X}^{GPR}\ (t\ +\ \Delta t)\ =\ \sum_{i\ =\ 1}^{n} \alpha_i . k_f\ (x_i, x_{test}) \tag{3. 9}$$

With n the number of training data, $\boldsymbol{x_i}$ is the i th input vector for training and $\boldsymbol{x}_{test}$ the test input vector. With the covariance function defined by:

$$k_f\ (x_p, x_q) =\ \sigma_f^2\ .exp\left(\frac{-\ (x_p\ -\ x_q)^2}{2l^2}\right) \tag{3. 10}$$

Where σ_f^2 and l are the hyperparameters of the covariance function, they define the complexity of the model and are also determined during the learning phase. In this study,

the coefficient α_i is determined during the learning phase (resulting from the application of the covariance function on the training data) by connecting the input data with the vector of n target values of the learning.

Table 10: TEST GPR kernel

GPR	**rRMSE**	**rMAE**	**rMBE**
Rational Quadratic	23,96	17,45	-0,20
Squared Exponential	24,00	17,54	-0,24
Exponential	24,61	17,83	-0,10
Matern 5/2	**23,95**	**17,46**	**-0,18**

In GPR model, the realization that is most coherent with the dataset is selected using training data, then the validation set is used to make predictions. The hyperparameters of kernel have been estimated from GHI training data via the minimization of the log marginal likelihood. Matern 5/2 was selected as the function kernel, Table 10. More details on GPR model may be found in (Rasmussen 2006), (Schulz, Speekenbrink, and Krause 2018).

➢ *Support Vector Machine (SVM)*

Support vector machines are a set of supervised learning techniques designed to deal with discrimination or regression problems. they can be used for both classification and regression. (Vapnik 2000), Support vector regression (SVR) is an adaptation data regression method of support vector machines. It has been successfully applied to time series forecasting. The formalism of SVRs is similar to that of Gaussian processes. The mathematical formula of SVRs applied to time series forecasting is (Lauret et al. 2015):

$$\hat{X}^{SVM}\ (t\ +\ \Delta t)\ =\sum_{i\ =\ 1}^{n} \alpha_i\ .k_{rbf}\ (x_i, x_{test})\ +\ b \qquad (3.\ 11)$$

Where, $\boldsymbol{x_i}$ is the i th input vector for training and $\boldsymbol{x}_{test}$ the test vector. The radial basis function k_{rbf} is given using

$$k_{rbf}(x_p, x_q)\ =\ exp\ [\frac{-\ (x_p\ -\ x_q)^2}{2\sigma^2}\ ] \qquad (3.\ 12)$$

The parameter b (bias parameter) is obtained from the previous equation, the hyperparameter σ, defines in particular the complexity of the model. Regarding SVRs, the coefficients α_i are related to the difference of two Lagrange multipliers, a quadratic programming problem is solved to deduce α_i. Unlike artificial neural networks, which are subject to local minima problems, for SVRs the problem is strictly convex and therefore has only one solution. Additionally, it should be noted (unlike Gaussian processes) that not all training models participate in the previous relation. Selecting the right kernel is important in the Support Vector Machine SVM model. We used an optimizable support vector machine that optimizes hyperparameters, a medium Gaussian Kernel was selected.

III.3.B. Satellite-based models

Techniques based on satellite data use either images from sky imaging devices (cameras pointed at the sky) or images from satellites. From these "images", these

techniques can predict the movement of clouds, and from this information, coupled with specific algorithms solar radiance can be predicted.

Satellite solar radiation estimates are also an alternative to a network of ground-based irradiation measurements and acquisitions, nationally or globally, Figure 23. Depending on the treatment of the interaction of solar radiation and the atmosphere, imaging satellites can be divided into two main families (albarelo 2017):

- Geostationary satellites, whose orbital plane is that of the equator and altitude of revolution is approximately 36,000 km. They are suitable for monitoring a particular region. For example, GOES (Geostationary Operational Environmental Satellite, NASA, United States) and Météosat (Europe). This family is characterized by low spatial resolution (1km pixel side for the visible GOES channel, 4 km for the GOES infrared channel) and high temporal resolution (for example, images every 15 minutes for Meteosat-7 and every 30 minutes for GOES).
- Polar-orbiting meteorological satellites, whose orbit is characterized by an inclination close to 90 ° with respect to the equator. They are characterized by high to very high spatial resolution (30 m pixel for Landsat 7, between 5 and 20 m for SPOT 5) and by low temporal resolution (between 2 and 6 images per day). Among the main polar-orbiting satellites are the NOAA AVHRR (National Oceanic and Atmospheric Administration Advanced Very High-Resolution Radiometer) meteorological satellite, as well as Earth observation satellites such as Ikonos or SPOT (Satellite for Earth observation).

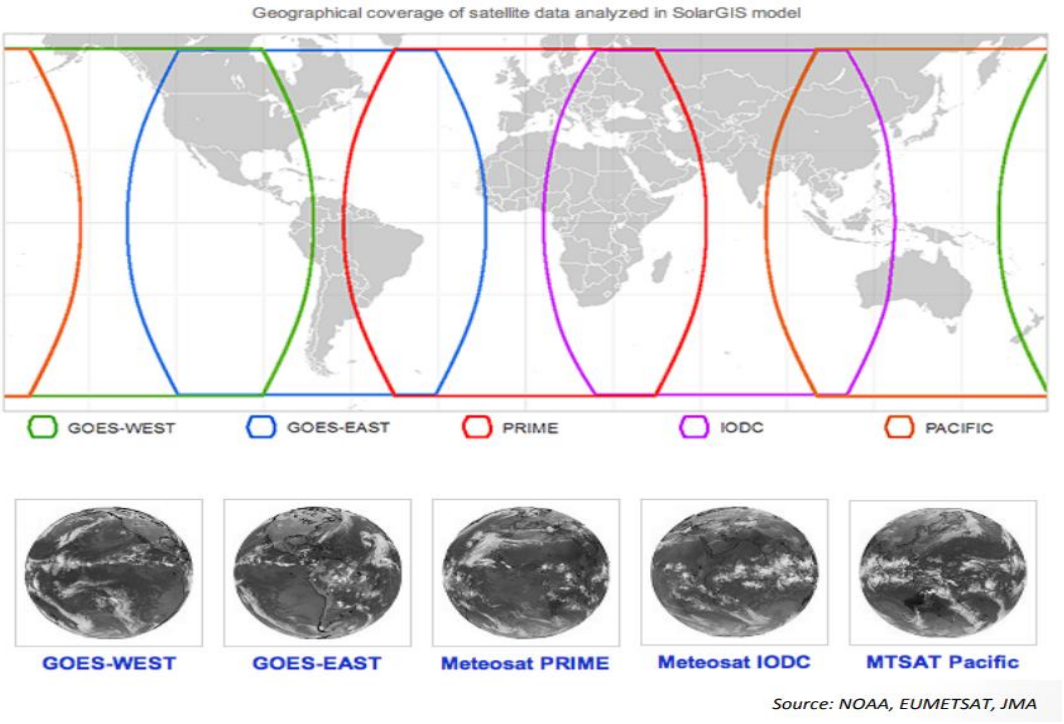


Figure 23: satellite broadcast coverage map worldwide.

Satellite models can be classified into two categories physical and statistical models (Noia, Ratto, and Festa 1993a) and (Noia, Ratto, and Festa 1993b). The advantages of physical models are their generalization due to the use of radiative transfer (RDT) models and the

fact that measurements of ground data are not necessary. However, radiative transfer models require precise and complete measurement of atmospheric morphology as well as careful calibration of satellite measuring devices. Statistical satellite models are based on simple statistical regressions between satellite and ground-based irradiation measurements. As a result, these statistical models are much simpler due to their independence from accurately measuring the composition of the atmosphere, but they suffer from their loss of universality and the need for data measured on the ground.

Heliosat algorithms for estimating solar resources were developed in the 1980s (Cano et al. 1986), The original Heliosat method is based on the principle that a change in cloud cover above ground affects the overall irradiance reaching this surface. This dependency relation was translated by a linear regression between the cloudiness index, (n), and the clarity index, (K_t). Beyer (1996) replaced (K_t) with the clear sky index (K_c), which simplifies the approach with identical results. Then, (Rigollier, Bauer, and Wald 2000) proposed a new evolution of the method now called "Heliosat-2", although this method uses images from the Meteosat satellite (centered on Africa). (Albarelo et al. 2015b) developed Heliosat-2 method to be used with GOES images as input instead of METEOSAT images over French Guiana, the modifications include a change in the calculation of the cloud albedo and in the Linke turbidity factor values. For this study, we choose the optimized method, which offers the best spatial resolution and whose algorithm is accessible. For more details, readers are directed to (Albarelo et al. 2015b) and (Fillol et al. 2017).

III.3.C. Numerical weather prediction method

The NWP model is not a single tool but is a numeric system to forecast weather, based on the exploitation of a chain of complementary numerical models. NWP models generate a probability of cloud occurrence to be used as an input to dynamic models of the atmosphere to then determine the value of solar radiation at ground level. Approaches of these forecasting methods, with their limitations and precision, can be found in the literature ((Diagne et al. 2014); (Elliston and MacGill, 2010); (Espinar et al., 2010); (Heinemann 2006); (Paulescu et al., 2013)). Some of the numerical weather models are as follows:

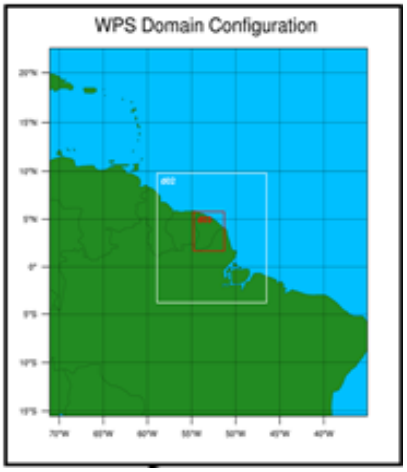
- ✓ The ECMWF model (Europe) "**E**uropean **C**enter for **M**edium-Range **W**eather **F**orecasts".
- ✓ The **GFS** model (American) "**G**lobal **F**orecast **S**ystem model ".
- ✓ The ARPEGE model (World) "**A**ction de **R**echerche **P**etite **E**chelle **G**rande **E**chelle".
- ✓ The AROME model (French) for "**A**pplication of **R**esearch to **O**perations at **ME**soscale".
- ✓ The WRF model (American) "Weather Research and Forecasting model".

WRF offers great flexibility to users with various spatial resolutions (from tens of meters to thousands of kilometers) and a large choice of different physical models. Moreover,

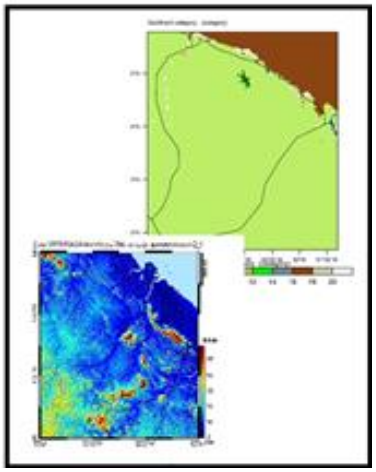
WRF has a large and active worldwide community with a cumulative total over 48 000 of researchers in over 160 countries according to the University Corporation for Atmospheric Research (UCAR).

III.3.C.i. Weather Research and Forecasting Model (WRF)

A numerical weather prediction system developed in the latter 1990's in cooperation between the National Center for Atmospheric Research (NCAR), the National Oceanic and Atmospheric Administration (represented by the National Centers for Environmental Prediction (NCEP) and the Earth System Research Laboratory), the U.S. Air Force, the Naval Research Laboratory, University of Oklahoma, and the Federal Aviation Administration (FAA) for both weather research and operational forecasting. It is interesting to test our proposal on a numerical model like WRF has a physical description of clouds. The parameterization of WRF for French Guiana used in this study was carried out by Mouhamed Diallo (Diallo 2019). GFS forecast were selected for initial and boundary conditions. To choose the initialization, the time needed for WRF to produce a balanced state from GFS coarser initial and boundary conditions must be considered Figure 24. According to (Aryaputera 2015); (Diagne et al. 2014) the spin up time in the ITZ ranges between 6 and 12H. The GFS model is initialized every 6H starting at 00h; consequently, we chose the initialization at 00 UTC to allow for a spin up of 12H before the sunrise in French Guiana.



Initial and boundary conditions



Physiographic data

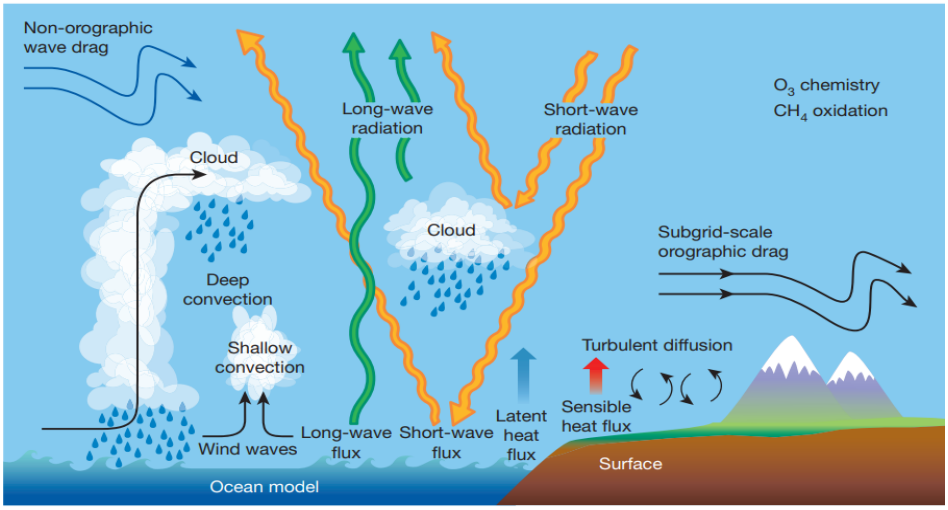


Figure 24: Physical process accounted for NWP models extracted from (Bauer, Thorpe, and Brunet 2015).

First, initialize the grid (preprocessing) as Figure 25, involves the following steps:

1) define the area of interest and the spatial resolution needed;

2) interpolate static data (albedo, topography, …) on the grid (geogrid);

3) interpolate dynamical data (wind speed, temperature, …) on the grid (metgrid).

The static data are provided by UCAR website. The dynamical or grid data are provided by a global NWP, which allows large-scale phenomena. In this study, we choose the Global Forecast System (GFS) data (Environmental Modeling Center, 2003) because archives dating back 1997 are available, free of charge and no pre-processing steps are needed to ingest GFS forecasts in WRF. The GFS model is initialized every 6H starting at 00h.

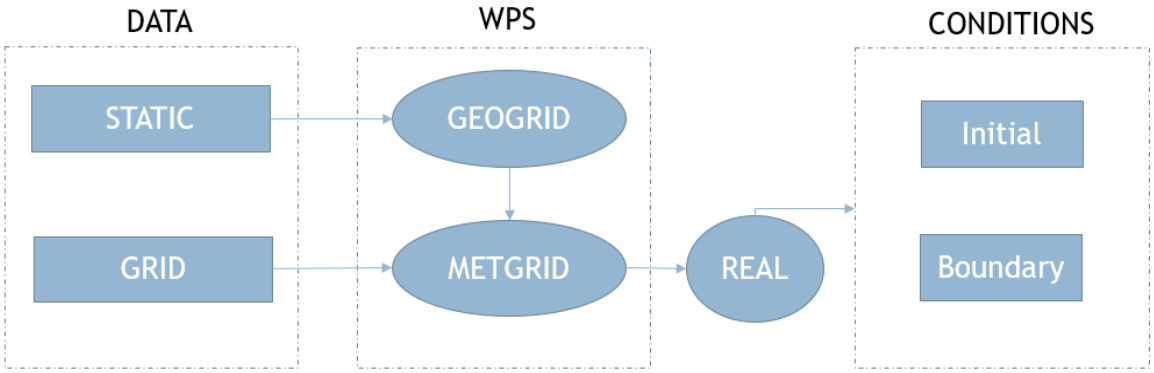


Figure 25: Schematic diagram of WRF Pre-processing System (Rim et al. 2018).

Input data:

- Initial and boundary conditions (ECMWF).
- Physiographic data (soil moisture, orography, albedo, land use, etc.).
- The temporal resolution is 3H.
- The spatial resolution reaches about 3 km at the center of the grid.

Once the initial grid is set, WRF builds successively the next grids (the forecasted grids) by resolving the atmospheric equations. (Diallo 2018) found the best schemes in his theses work for each parameterization for solar forecasting in French Guiana. These parameterizations will be used in this study. Predictions of solar irradiation made using numerical models for the one-day ahead are known to be highly biased, due to the difficulties associated with insufficient performance of the radiative transfer models during partly cloudy situations. The clouds being reproduced in an averaged manner on each pixel, these models cannot resolve a partly cloudy sky. Therefore, it is necessary for us to refine them with a post-processing method.

III.3.C.ii. Post-processing of NWP models

Once forecasts with NWP have been generated, these can be improved by comparing them to data measured during an evaluation period in which improved models are developed and fitted. This MOS (Model Output Statistic) approach works fully if the forecast corrections are updated over time (recursive) and developed separately under different conditions or systems. That is because forecast errors often depend on the hour and year, sky conditions, etc. These models are frequently modified and updated, which requires forecast corrections to adapt them, therefore we will speak of adaptive forecast and correction models. (E. Lorenz 2011) applies a method to forecast solar irradiation in Germany by coupling the MOS correction with the clear sky index and the solar zenith angle to correct the bias related to the forecast. A similar procedure was applied by (Mathiesen and Kleissl 2011) for the United States and different bias structures were found for different NWP models. The elimination of bias has also been studied by (Pelland, Galanis, and Kallos 2013) for North American global radiation forecasts from NWP models, using a linear Kalman filter with 30-60 days sliding training window. Other approaches to forecast from the outputs of NWP models and historical data have been performed. Authors proposed and validated several methods for GHI forecasts in various territories located in the ITZ. (Diagne et al. 2014) proposed Model Output Statistics (MOS) method which is an ANN model for improving the day ahead solar GHI forecasts from Weather Research and Forecasting (WRF) model in Reunion Island. (Soubdhan et al. 2016) combined a Kalman filter and an Auto-Regressive (AR) model in Guadeloupe to produce GHI estimates for forecasting horizons ranging from 1 second to 60 min. In another study using WRF as well as two statistical methods to obtain a day ahead GHI forecast. Yet, results found for territories located in the ITZ were consistently higher than those found for similar climate and models in extra-tropical areas.

The Kalman model predicts the GHI bias that may occur with the numerical model prediction which is WRF in our case. The investigation of the optimal input set in the Pelland's approach (Pelland, Galanis, and Kallos 2013) and replacing the measurements with satellite estimates in the application of the Kalman filter to correct the prediction of WRF forecasts. This latter combined with the Galanis initialization technique (Galanis et al. 2006), Kalman was used for the wind forecast, this combination resulted in a new prediction method.

This family of models is interesting since having significant data histories is not necessary. However, they exhibit relatively restricted performance with difficult apprehension due

to sudden variations in radiation. They are very often used as a reference model to better estimate the contribution of more complex predictive models as WRF model.

In our study, we used a Kalman filter as a tool for improving the forecast of WRF model. which is a recursive estimator (Cheng, 2016), it is a compilation of the estimate at this moment only with the estimate at the previous moment and the measure of this moment (Chaabene et Ben Ammar, 2008). Furthermore, the recursive nature of the Kalman algorithm, unlike other conventional forecasting methods, such as the ARMA model makes it a solution suited to this step. The Kalman filter is more often defined by two phases, which are complementary, Figure 26:

- The prediction phase.

- The correction 'update' phase.

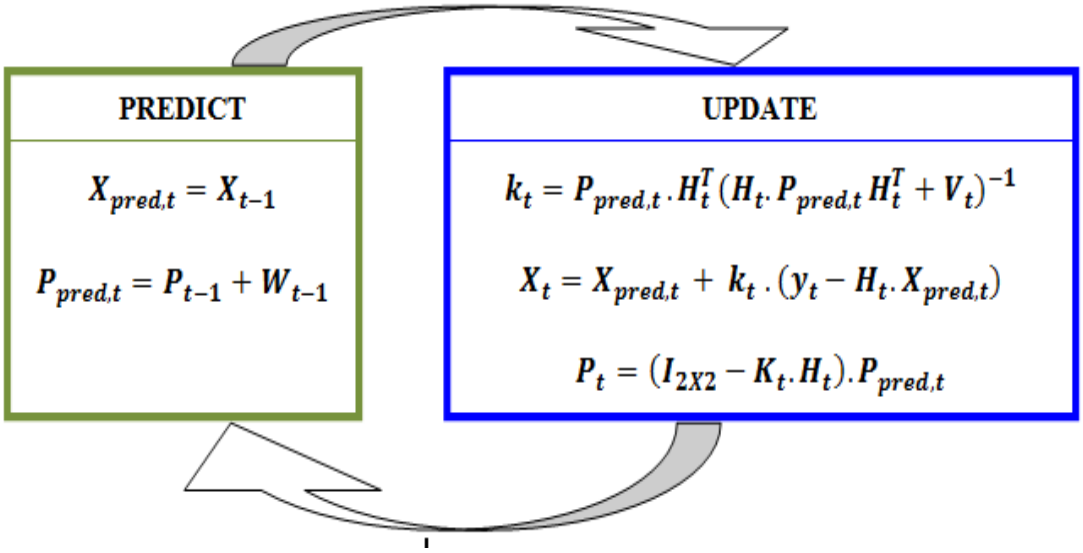


Figure 26: Kalman filtering algorithm (Extracted from (Pelland, Galanis, and Kallos 2013)).

The forecast phase is based on the state at the previous instant to produce a forecast for the present instant. This predicted state is called "a priori", at the moment when it is generated, it is not possible to have an observation of this moment.

In the next phase so-called "update" phase, the "priori" forecast is compiled with the observation of this moment to refine the forecast. The result of the compilation is then called a "posterior" forecast. For the general case, the sequence of the prediction and correction phases, as well as the applied equations are written as

1. **State Prediction**

Let $\hat{x}_{t+h\backslash t}$ be the state to be predicted at time step h based on information available at time t:

$$\hat{x}_{t+h\backslash t} = M_t . x_t + \omega_t . \quad (3.13)$$

Here ω_t is white noise represents the state random change from t to t+h, which corresponds to a covariance W (= N (0,W)). M_t is the matrix, which connects the real state to the observation, in this study we started from the assumption that the observation is the real state; therefore, M is an identity matrix.

At time step h, an observation of the actual state y_t , takes the following form:

$$y_{t+h} = H_{t+h}.x_{t+h\backslash t} + \varepsilon_t \quad (3.\ 14)$$

$$H_{t+h} = [1, \quad \widehat{WRF}_{t+h}, \quad Cos(\Theta_z(t+h))] \quad (3.\ 15)$$

H_t is a matrix of 3 columns; given as follow:

In which ε_t is the observation noise, which is assumed to be an average Gaussian white noise with a covariance V (= N (0, V)). The initial state of the model and noise at each time step is assumed to be mutually independent. This is in fact a "posterior" predicted state at time t, which comes from observations up to time t.

2. **Covariance Prediction**: It is then necessary to judge the accuracy of the model. We calculate the covariance matrix of the "posterior" error, which is called P (an estimate of the predicted state). The definition of the form of P is:

$$\hat{P}_{t+h\backslash t} = P_t + W_t. \quad (3.\ 16)$$

During the compilation of this matrix P, we defined a “sliding window” for the computation of the covariance matrix allowing us to consider the seasonality due to the nature of the data constituting the signal.

3. **Gain Correction**: deduce the gain of the Kalman filter K, which is defined by:

$$K_{t+h} = \hat{P}_{t+h\backslash t} * H^T{}_{t+h} / (H_{t+h} * \hat{P}_{t+h\backslash t} * H^T{}_{t+h} + V) \quad (3.\ 17)$$

4. **State Correction**: Then we introduce the correction factor described and defined by A.C. Harvey (1990):

$$x_{t+h} = \hat{x}_{t+h\backslash t} + K_{t+h} * (y_{t+h} - H_{t+h} * \hat{x}_{t+h\backslash t}) \quad (3.\ 18)$$

5. **Covariance Correction:** updating the error covariance term

$$P_t = (1 - K_t * H_t) * \hat{P}_{t+h\backslash t}. \quad (3.\ 19)$$

Then, sequences of the next prediction and correction phases are repeated.

From these formulations, we define the forecast model for a horizon h = 1:

$$\hat{x}_{t+1\backslash t} = M_{t+1} * \hat{x}_{t\backslash t} \quad (3.\ 20)$$

For the generalized formula for any horizon h (1H to 6H):

$$\hat{x}_{t+\Delta t\backslash t} = M_{t+\Delta t} * \hat{x}_{t\backslash t} \quad (3.\ 21)$$

In simple words, the Kalman filter is a model which:

- Compare the forecast to the observation;

- Attempt at every moment t to make the best forecast match with the observation.

First forecasts of the Global Forecast System (GFS) model are downloaded. Second, in order to downscale the temporal and spatial resolution of this forecast, data from GFS are used as input to the WRF model. WRF builds the forecasted grids of solar irradiation. As we said the spatial resolution of this grid can reach 3 km. Third and last, a post-processing method as Kalman filter bias correction is applied to obtain one to six hours ahead of forecasting using WRF output and satellite data, Figure 27.

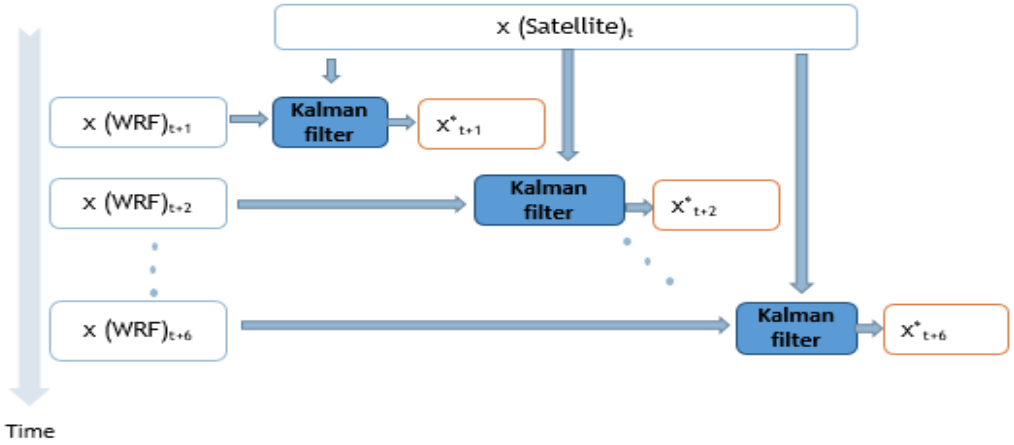


Figure 27: A Kalman filter bias correction method with satellite data for 1 to 6H ahead.

Using measurement data in post-processing step reduces the spatial resolution. In other words, the post-processing of WRF is applying only on the point where we have the measurement. Our contribution in this part will be testing and validated the use of satellite data to correct the WRF forecast's bias. GHI forecasts from WRF were compared with ground-based irradiation measurements to analyze forecast accuracy. This way allows us to correct the output of WRF model in any geographical point in French Guiana and for several forecast horizons.

III.3.D. Hybrid method

Ensemble forecasts are classical in meteorology for uncertainty quantification. However, Ensembles combine multiple members to create a single output that should have better performance than any other ensemble member. Most of the articles in the literature related to comparisons between different complex models, the models are compared with each other, and often hybridized to observe the possible improvements in their performance due to these hybridizations. Hybridization is putting together different models to use them together and observe the effects on performance when forecasting. This method often generates improvement but also the complexity of the models. To reduce model uncertainties, using hybrid systems have been suggested. A hybrid system or a multi-model ensemble is characterized by a combination of any forecasting methods. This approach captures the advantages of each ensemble member and forecasting

method. It combines ensemble members to produce a single forecast more accurately than any individual model of the ensemble. (Mallet, Stoltz, and Mauricette 2009) summarized and tested this method on forecasts of respectively electricity consumption and ozone concentrations. Still, (Yokohata et al. 2013) studied climatological ensembles of top atmosphere radiation and radiation in cloud-free conditions. (Cesa-Bianchi and Lugosi 2006) detail the solid mathematical background of hyprid methods. Recently (J. Thorey et al. 2015) compare the ensemble forecasts of solar radiation from TIGGE (THORPEX Interactive Grand Global Ensemble) and used a sequential aggregation method to combine linearly irradiation forecasts from these several weather forecast centers. (J. Thorey et al. 2015) found that the aggregated forecast is more accurately than any individual forecasts. Nevertheless, this approach only used the weather prediction model as ensemble members. To the best of the author's knowledge, no previous studies have attempted to combine weather forecasts model and statistical methods to improve hour ahead of solar irradiation forecasts.

The Ridge Regression method was used to linearly combine all forecasts with weights varying by models. Sequential aggregation updates the weights before any forecast using satellite-derived data, we validated our method over three months of GHI ground-based irradiation measurements of 2016 with hourly data of GHI from six stations of the French national weather service.

We apply a ridge regression method (J. Thorey et al. 2015) to aggregate each ensemble member's GHI forecasts. Data from one site used to obtain every model weight via the equation (3.22):

$$w = argmin(\sum_{i=1}^{n}(X_i - X'^{T}_i w)^2 + \sum_{j=1}^{p} w_j^2) \tag{3.22}$$

Data from other sites are used to validate these weights to derive an improved GHI forecast equation (3.23).

$$\hat{X}^{ME}(t) = \sum_{T=1}^{m} w^T \hat{X}^T(t) \tag{3.23}$$

Where, m is the number of models; w is The model's weight and $\hat{X}^T(t)$ is the GHI forecasted by the model at time t. Consequently, the aggregated ensemble has a forecast horizon of one to six hours. The different steps of the ensemble method used to improve the GHI forecasts in the ITZ are shown in Figure 28.

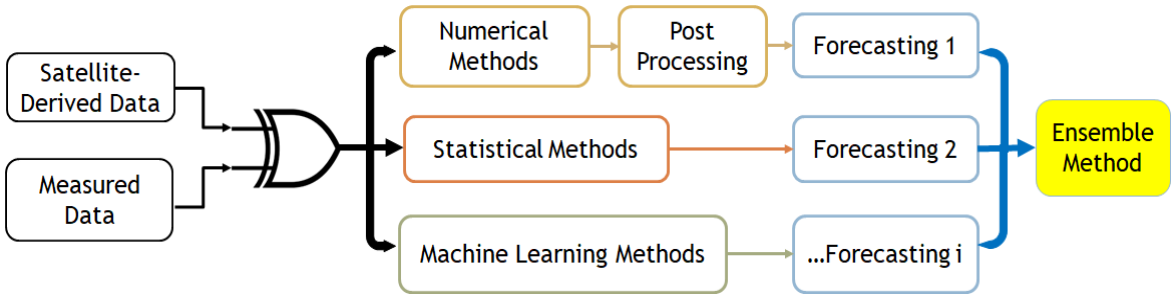


Figure 28: A basic flow diagram of the process.

The first model of the ensemble is a hybrid physical statistical model; it is obtained by post-processing physical model: Weather Research and Forecasting model (WRF) output with a Kalman filter bias correction method, the second, third and fourth models are statistical: the smart persistence, Auto Regression moving average and machine learning models.

III.4. Summary of the models

To forecast the GHI, we considered a coupling with an alternative resource that accounts for cloud properties. Satellite-derived data were chosen as the alternative resource because they can provide information on the optical properties of the cloud. The cloud fraction observed by satellite is used as a proxy of cloud transmittance. Every chosen model of these models constitutes an interesting panel of techniques to tackle forecasting problems. In the family of models without learning and in general, it is possible to assert that these models are easier to implement (persistence and smart persistence). In the family of models with training step (much more complex) it is possible to "classify" the models by the complexity of implementation:

- Classic methods, are the least complex (AR, ARMA).
- Methods based on Kernel functions (Gaussian process, Support vector machine).

The methods of strengthening, bagging and random forests are based on simple regression trees. However, the nuances in their operation, whether it is the reinforcement algorithm or the sampling of training data from several methods, give these models greater complexity of implementation and less readability. For physical methods, we chose to use the WRF model to make the GHI forecasts. WRF numerical model is free, has a dynamic scientific community and the initialization data are easy to acquire. Indeed, a bibliographic study has shown that the Kalman filter is a powerful post-processing technique in the field of forecasting meteorological variables. In fact, the treatments are done on a desktop computer and the results are available quickly. The program requires low computing power and allows results to be obtained in real time. Concerning Hybrid Methods, which are the most complicated methods because they require the simultaneous use of more than one of the models. We choose a Sequential Aggregation method, one of the benefits of aggregation is that large improvements are achieved at the most difficult time steps, even with a single member of the ensemble (Jean Thorey 2018).

For the 6 ground-based irradiation measurements sites, forecasts of GHI at hourly time steps and for time horizons of 1H to 6H were produced. The purpose of this research is not to compete with various solar forecasting tools that are academically or commercially available today, but to generate our own solar forecasting results using satellite-derived data with simple, inexpensive and effective methods.

In the next chapter, 1H forecasts for each model will be first presented using one year of measured and then satellite-derived data. The next 1 to 6H forecasts will be discussed. Three months' data of 2016 are used to reduce the bias of a numerical model. Finally,

satellite-based method forecasts will be gathered in an ensemble method to get one and more accurate forecasts.

Chapter .IV. Simulation and forecasting results

IV.1. Introduction

We have described, in the previous chapters, the main tools that we have used to achieve the hourly GHI multi-horizon forecast using satellite-derived irradiance as observations. After discussing the need for forecasting, we have reviewed all of the predictive methodologies. which ranges from collecting and preprocessing data in Chapter .II to describe in detail the methods tested in Chapter .III. The third chapter focused on the development of an operational methodology by detailing the models that we have implemented and by proposing a hybrid forecasting approach. The goal is to show that satellite-derived data can replace ground measured data. Then study the accuracy improvement of the models forecasts due to hybrid ensemble method.

This chapter contains the results and discussions of empirical tests of the methods chosen. All the experiments are conducted using the MATLAB environment; The tests are focused on comparing the predictive performance for several ground measured and satellite-derived irradiance. This chapter is divided into four main sections relating to four experimental conditions. The most objective conclusions will then be drawn up to highlight the strengths and weaknesses of the different models according to observations.

The first section of Chapter .IV relates to the models for intraday GHI forecasting per hourly time step in French Guiana. These models were previously described in (III.3.A.i), Our contribution will be the use of satellite-derived irradiance data as input to statistical methods to forecast solar irradiation in an ITZ. These models will be validated and compared over six meteorological sites (Rochambeau, Saint Laurent, Saint Georges, Maripasoula, île Royale, and Kourou) in French Guiana. Then we will try to relate the results obtained in a hybrid method to improve the forecast.

As we saw in Chapter .I, forecasting demand from power grid operators is increasing. In addition to the expected value of the resource, these users should have a forecasting framework. Thus, in the second section of this chapter, we will briefly present the development of the different forecast models used in this study. Then, in the third section, we will present a preliminary assessment of statistical forecasting techniques using hourly satellite-derived irradiance as input to statistical methods. In the fourth section of this chapter, four machine learning models will be used to forecast GHI using satellite-derived irradiance data as input. The fifth section of this chapter will deal with improving the prediction of the numerical weather prediction model using satellite-derived irradiance data as input in kalman filter. Our proposal with satellite-derived irradiance data will allow us to obtain the following framework; using satellite-derived irradiance data as observation outperform the post-processing technique predictions. The sixth section will present the work and experiments performed within the framework of the ensemble method. The operating mode will differ a little from what can be found in the literature, and thus lay the foundations for a new forecasting model on the horizon and the geographic zone. Finally, a general summary of our simulations and results will be presented and discussed. The new approach is the use of hourly satellite-derived irradiance as inputs to build and train different models to forecast the GHI in the ITZ. Using satellite-derived irradiance as input to statistical methods has the advantage of

removing ground-based irradiation measurement stations and then makes it possible to make the forecast at any point in the satellite image.

IV.2. Forecast of horizontal global irradiation (6h/1h)

The first part of the simulations concerns the development of the different forecast models. Additionally, this step allowed us to test the protocol for preparing and using data for conducting the forecast. In this section, we will detail the forecasts made using all the models we have produced. The goal of most of this work was to evaluate models for intraday forecasting GHI at hourly time step according to two sets of observations.

IV.2.A. Forecast models Construction

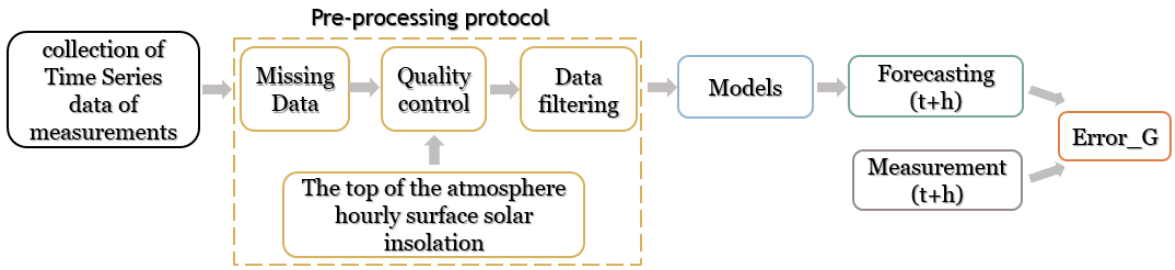
Hourly satellite-derived irradiance is used to realize the prediction with ten models. The developed and tested models fall into three main groups:

1. **Statistical Models:**
   a. **Linear Models**
      i. Persistence model (P).
      ii. Scaled Persistence model (SP).
      iii. Auto Regression model (AR).
      iv. Auto Regression Moving Average (ARMA).
   b. **Non-linear Models**
      i. Gaussian Processes Regression (GPR).
      ii. Support Vector Machines applied to regression (SVM).
      iii. Simple Regression Trees (TRE).
      iv. Ensembles of Trees 'Bagging' (ENsTRE).

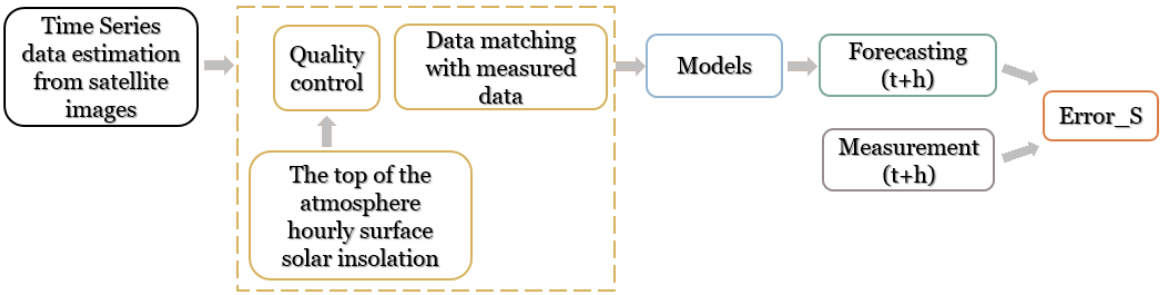
2. **Numerical models:**
   Weather Research and Forecasting model (WRF), numerical weather forecasting model, improved by the use of the Kalman filter.

3. **Hybrid models:**
   Combine all the previous predictions into a single improved one with a sequential aggregation method (J. Thorey et al. 2015).

The protocol followed for constructing the forecast models is presented in Figure 29 for measured data and then satellite one.



a. Measured Data



b. Satellite-Derived Data

Figure 29: Representation of the methodology to construct the forecast models

The first steps up to pre-processing have been detailed in section II.4 of this document. Once; the data are pre-processed, they are ready to be used by the different models. The different models are presented in detail in Figure 29.

IV.2.B. Models performance comparison

We classified in Table 11 each forecast model used in this study. This classification is based on two criteria: their structure complexity and the number of parameters required for their implementation. This table aim to ease the comparison between each model and highlight both their advantages and drawback.

Table 11: Models Classification according to their complexity. A mathematical description of these model was given in Chapter .III. Complexity degree increase from + to ++++.

	MODELS	*PROS & CONS*	*TIME*	*CONFIGURATION PARAMETERS*	*COMPLEXITY*
1.	***P***	Reference model. No historical data requirement. A high correlation between past and future.	+	-	-
2.	***SP***	No historical data requirement. A high correlation between past and future. Clear sky irradiation.	+	Clear sky model data.	+
3.	***AR***	Forecast using a linear combination of predictors. Less information is required.	++	The number of historical data.	++
4.	***ARMA***	Takes into account trends, cycles, seasonality, errors, and other non-static types of data when making forecasts.	+++	The number of historical data.	++
5.	***GPR***	Directly captures the model uncertainty. Add prior knowledge by selecting different kernel functions.	+++++	The number of historical data. Covariance function. Hyperparameters.	+++
6.	***SVM***	Add prior knowledge by selecting different kernel functions.	+++	The number of historical data. Covariance function. Hyperparameters.	+++
7.	***TRE***	Simple to understand, interpret and visualize. Over-fitting issues.	+++	The number of historical. Data Reinforcement Algorithm. The number of trees.	+++
8.	***ENsTRE***	Reduce the variance of a decision tree. Not bias, and solves over-fitting issues in a model.	+++	The number of historical data. The number of trees. The number of samples.	++++
9.	***WRF***	Open source code. Easy to acquire data initialization.	++++++	Download GFS data. WRF Parameterization.	++++
10.	***Kf***	Recursive nature of the algorithm. Efficient technique in meteorological variables forecasting. Reduce the bias in the GHI forecasts without needing a long historical data archive	++	Solar Zenith Angle. Hyperparameters. Update window.	++
11.	***Ens***	Coupling different methods. Improve the prediction performance over any contributing member of the ensemble.	++++	Different forecasting methods.	+++

The simplest model to implement is the Persistence model, this model consists of considering that the predicted data are equal to the previous data while keeping in mind that it is only efficient in clear sky condition so it is only useful as a reference model. Next is the Scaled Persistence model, which despite its simplicity performs better. Slightly more complex models are recursive, Auto-Regressive processes and with Moving Average. These models have approximately the same number of parameters. Models based on machine learning as Regression Trees, Gaussian Processes, and Support Vector Machines are even more complex.

Finally, the more complex models are numerical ones such as WRF. WRF output are improved using the Kalman filter. The complexity lies in determining the update window. Further details on these update windows were given Chapter .III.

The last one is the one based on ensemble techniques and data sampling. They require more parameters to implement and data sampling can be complex, making these models the most computationally intensive. This construction phase was important and allowed us to familiarize the operation. Subsequently, we will compare these different models on both ground and satellite-derived data series separately. After studying the

characteristics of the stations and sky conditions in each station in II.5, we decided to use the data from all stations for the simulations. This choice is intended to facilitate the discussion and results presentation for the ten different models, the two datasets used as input, and horizons from 1 to 6H ahead. For the 6 ground-based irradiation measurement sites, we made forecasts of GHI at hourly time steps and for horizons from 1H to 6H ahead. We will first present the one-hour forecast results with the ground and satellite-derived data separately by identifying the best model according to the uncertainties obtained. Second, we will compare the results for the remaining forecast horizons. Details on prediction errors by datasets, model, station, and time horizon can be found in the ANNEX.

We will present graphically the various parameters for estimating the relative forecast error RMSE, MAE, and MBE. MBE will not only allow us to judge the accuracy of the models, but also to determine whether these predictive models with satellite-derived data overestimate or underestimate the reel value. The MAE, RMSE is more sensitive to large forecast errors (and weighs larger deviations) and is recognized to be more representative of the quality of a model (COST action ES1002 Weather Intelligence for Renewable Energies (WIRE), 2012). It is used especially when small errors are tolerable and large errors have negative consequences. Models will also be compared with each other using the efficiency score of the model "skill score" or "forecast score". Using this parameter, the improvement (or not) brought can be estimated using a prediction model (more complex) compared to a reference model (naive and simple to implement); this reference model in our study is the Persistence.

The next sections are devoted to the use of models on the available datasets, identifying model selection parameters, and showing the forecasting performance from one to six hours ahead obtained for models of each category.

IV.3. Preliminary assessment of statistical forecasting techniques solar irradiation forecast

The objective of this section is to intraday forecast hourly GHI with good precision over French Guiana located in the ITZ. A time series of images are retrieved from the meteorological geostationary satellite GOES-13 to estimate GHI (as seen in III.3.B). We made a hypothesis in an ITZ, and we studied it for simple and primary cases. Models based on Persistence and Auto Regression were first tested. The mathematical methods to build the models are introduced in Chapter .III of this document.

In this section, the two-phase realization of the AR and ARMA models is implemented in the MATLAB System Identification Toolbox. Once the ARMA and AR parameters are computed, one can forecast hourly GHI for the validation period. To produce GHI forecasts at time (t + forecast horizon), these models need observations at time t or (t, t-dt). In this section we start by presenting the results using data from each station.

Figure 29 shows the results of the 1H ahead forecast error for each station. For regression models, the differences in forecast error estimates between ground measurement and satellite-derived data are less than 2%. Moreover, the differences in the performance of persistence model between the ground and satellite-derived data range from 0.2 to 1.5% depending on the station. For the 1H time horizon, models perform the same with ground and satellite-derived data, noting that the least performing is the persistence model (P). The KR station has the lowest forecast error estimates among the six stations with nRMSE and nMAE values around 37.3%, 30.8% for ground data and 37.4%, 30.4% for satellite-derived data, respectively. Whereas the IR station has the highest forecast error estimates among the six stations with nRMSE, nMAE around 42.4% 34.7% for ground data and 41.5% 33.8% for satellite-derived data, respectively. Persistence model is a naïve estimator known to perform poorly and is usually used as a reference model for comparison with other predictors, as is the case in this study. The Scaled Persistence (SP) model with ground data is more accurate than with satellite derived data. For the KR station, the nRMSE and nMAE are about 25.3%, 19.3% for ground data and 32.5%, 21.4% for satellite-derived data, respectively.

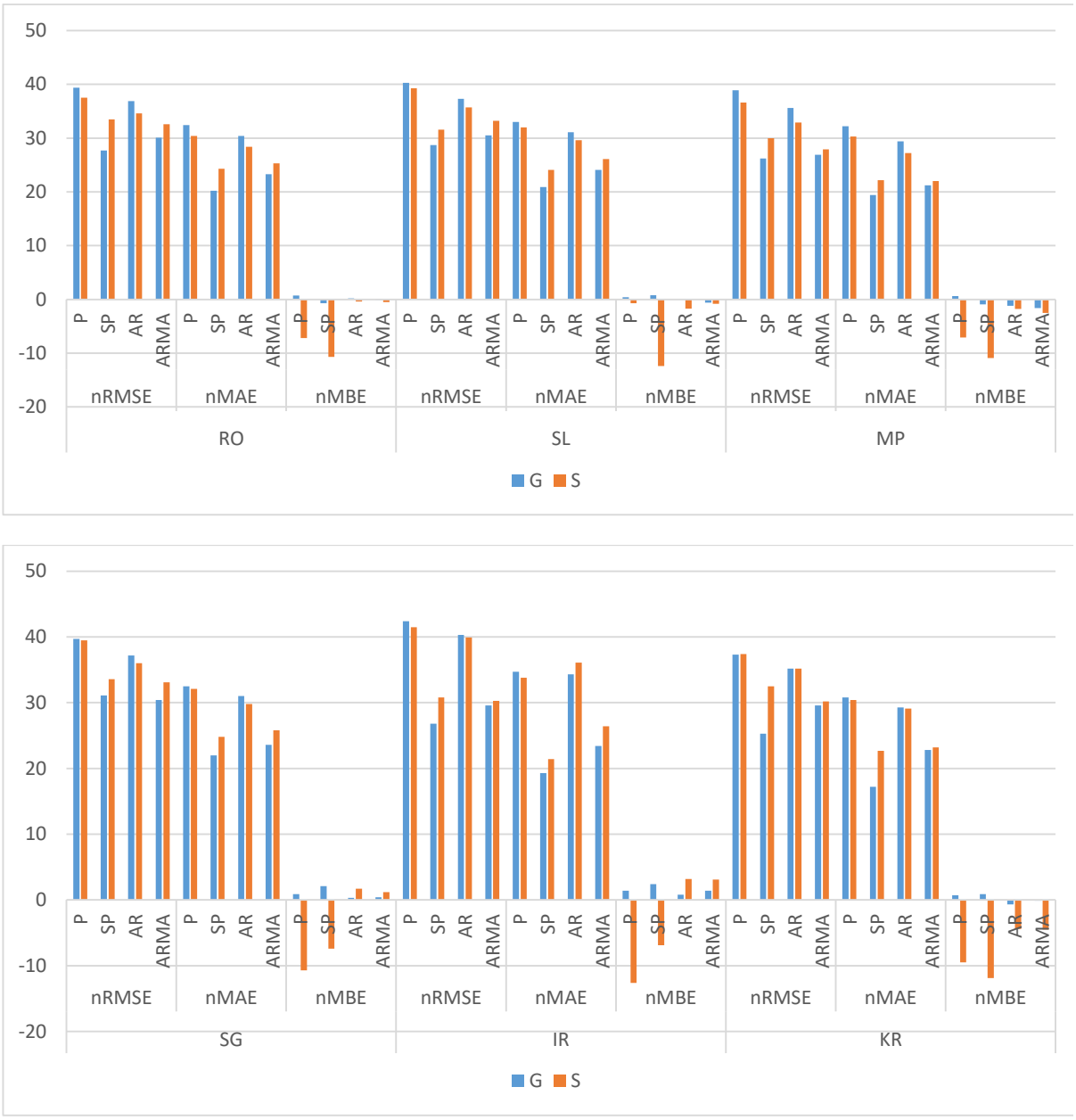


Figure 30: 1H forecasting errors according to station for Ground measurement (G), Satellite-derived data (S) with four statistical models in 2016.

Representing the MBE in the form of histograms allows us to judge if the models over- or underestimate the observation. If the model is well built, its MBE should be close to zero. The nMBE results presented in Figure 30 show that using satellite-derived data in the persistence-based method underestimate the observations according to the station by about 10.7% to 3.2% for P model and by about 12.4 to 6.9% for SP model. While the use of satellite-derived data with the regression-based method tends to vary between overestimation or underestimation of observation according to the station. For 1H time horizon, results reveal that, the best performance models are ARMA and then SP.

The next step is to plot the forecasting errors against the horizons. It seems interesting to represent several forecasting horizons accuracy for ground and satellite-derived data graphically. This representation highlights the performance of these models for multistep-ahead in French Guiana, as in Figure 31 and Figure 32.

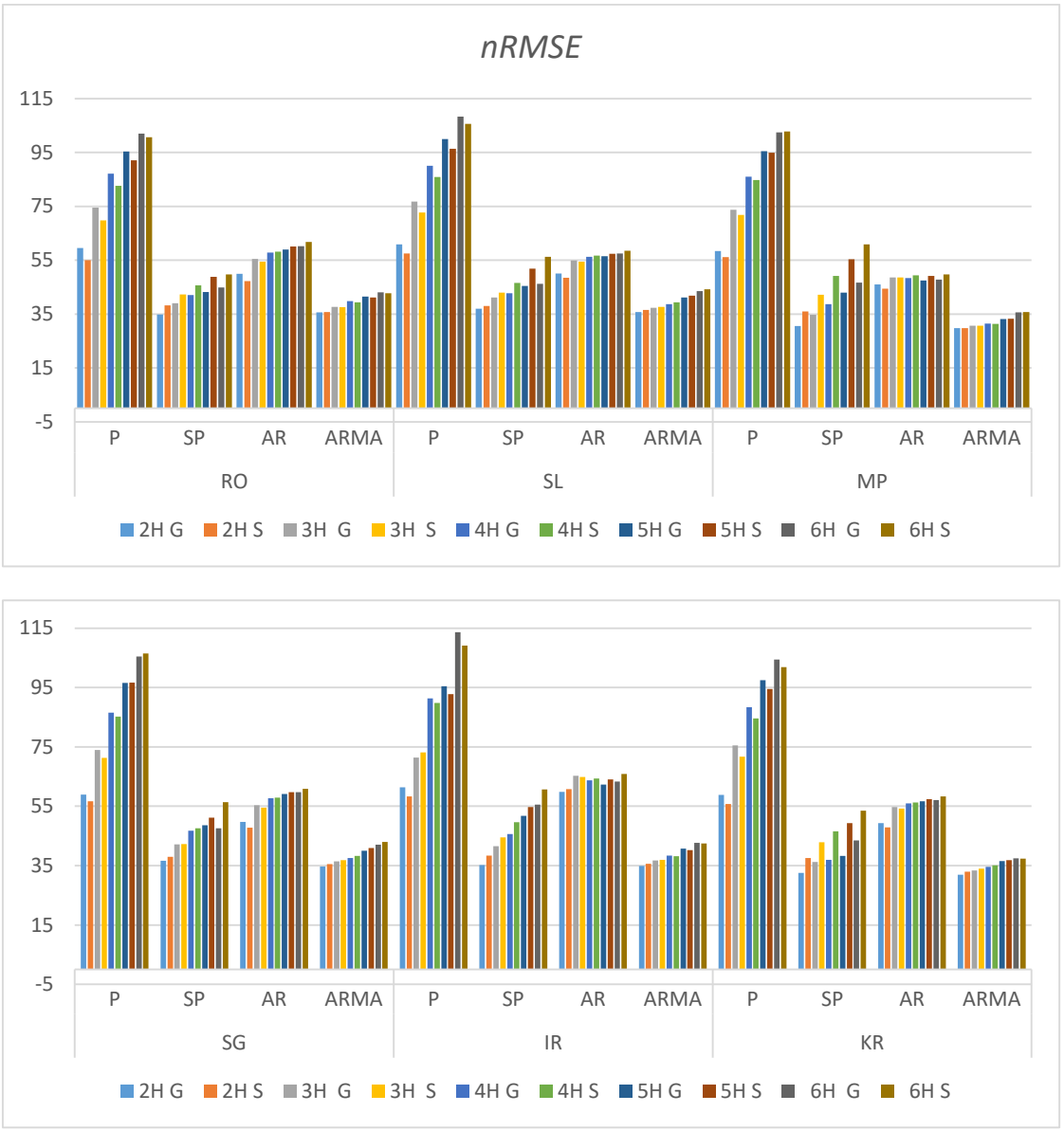


Figure 31: nRMSE forecasting errors vs. time Horizon for Ground measurement (G), Satellite-derived data (S) with four statistical models in 2016 according to station.

Figure 31, satellite-derived data with the ARMA model improves the P model depending on the station from 4.9% to 8.7% for the 2H horizon and 57.9% to 67.1% for the 6H horizon. This difference can be taken into consideration-to justify the use of a complex model such as ARMA. Forecasting-errors show that models perform-better with ground data for a one-hour time horizon.

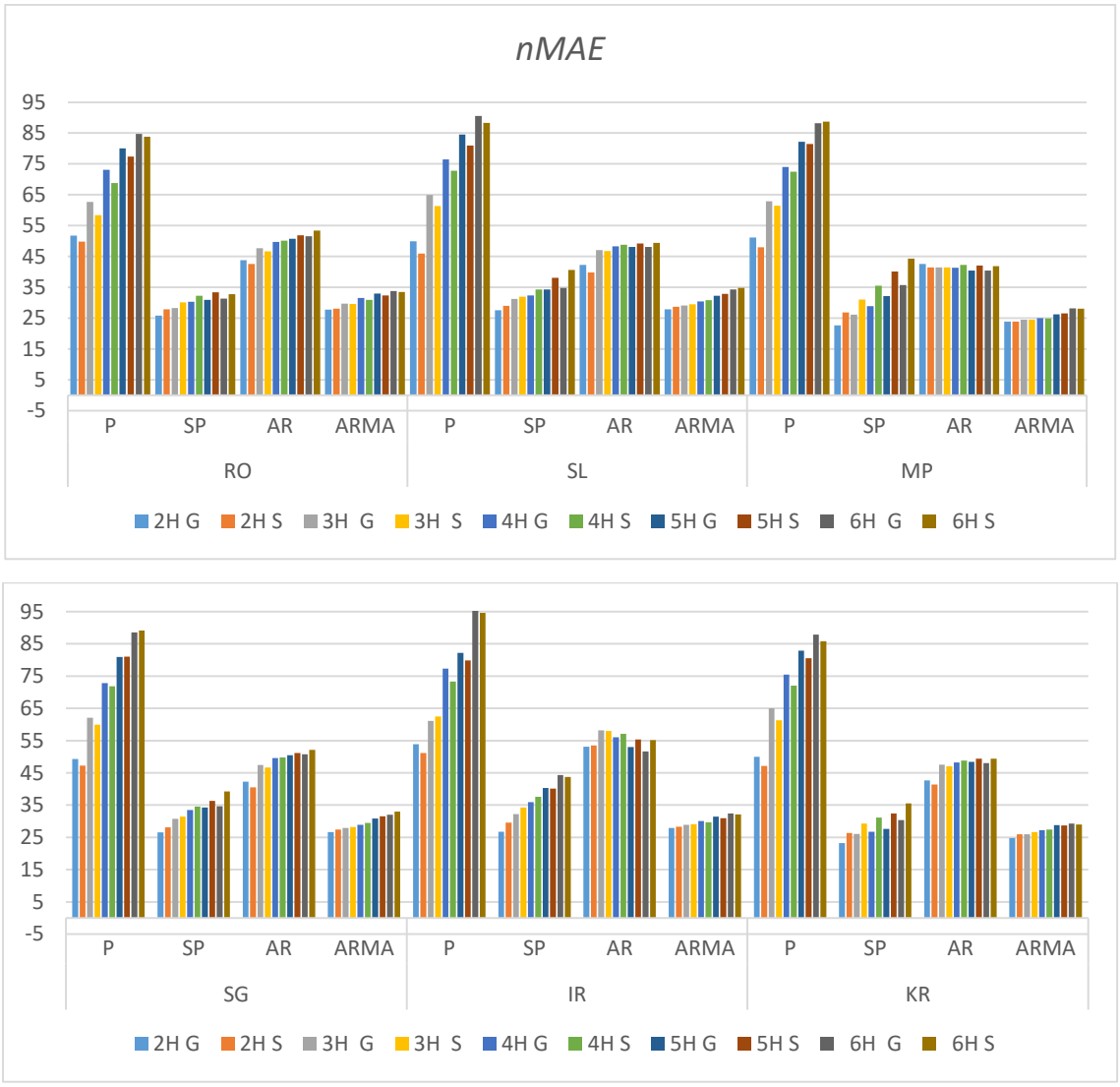


Figure 32: nMAE forecasting errors vs time. Horizon for Ground measurement (G), Satellite-derived data (S) with four statistical models in 2016 according to station.

However, for horizons over 2H, satellite-derived data are more suitable to make the forecast. Since satellite-derived data give information at the top of cloud processes, they have more information about the next value and are more suitable to make the forecast than the Ground one.

The nMBE forecasting errors of four statistical models depending on the station, time horizon are shown in Figure 36.

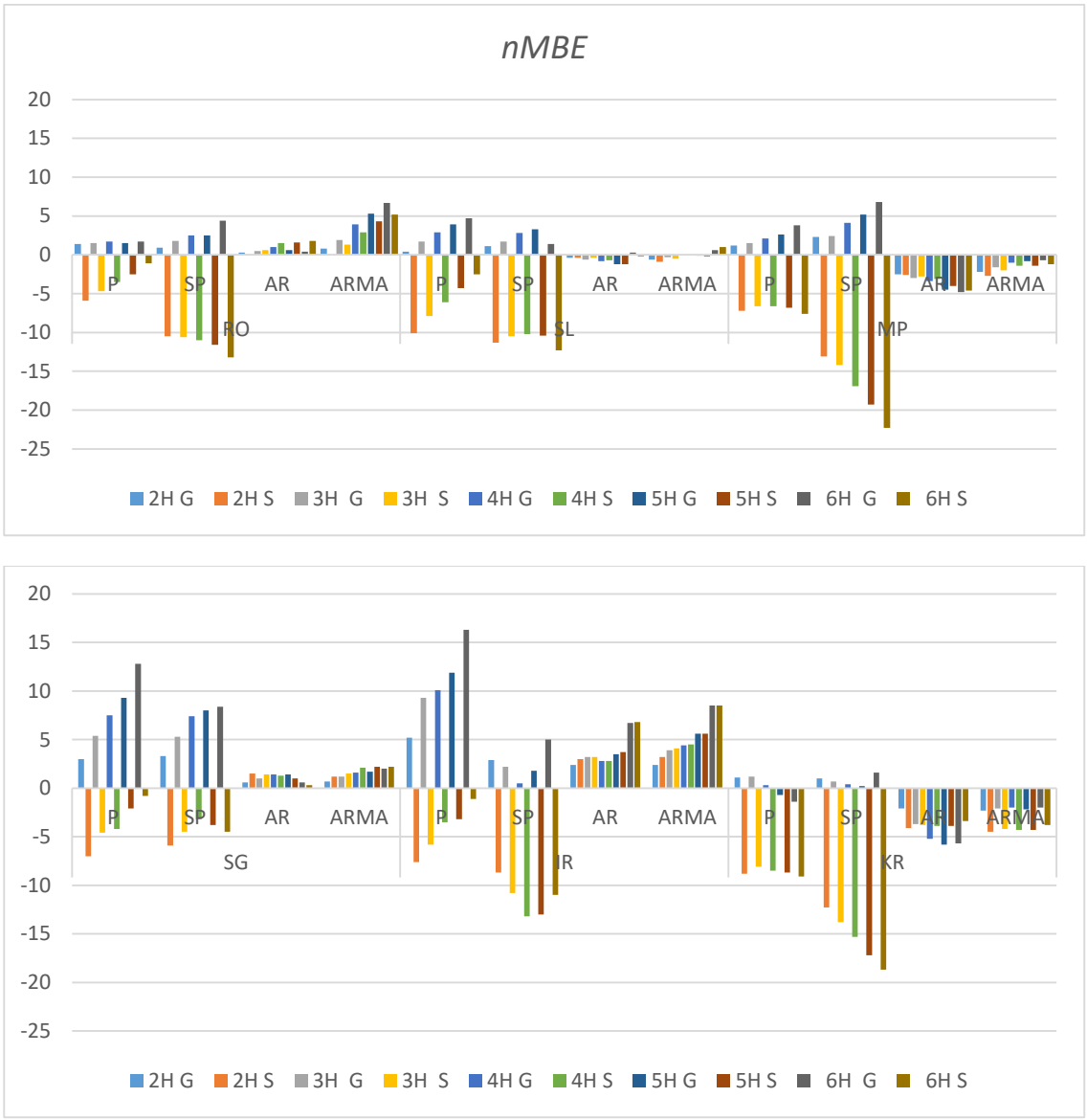
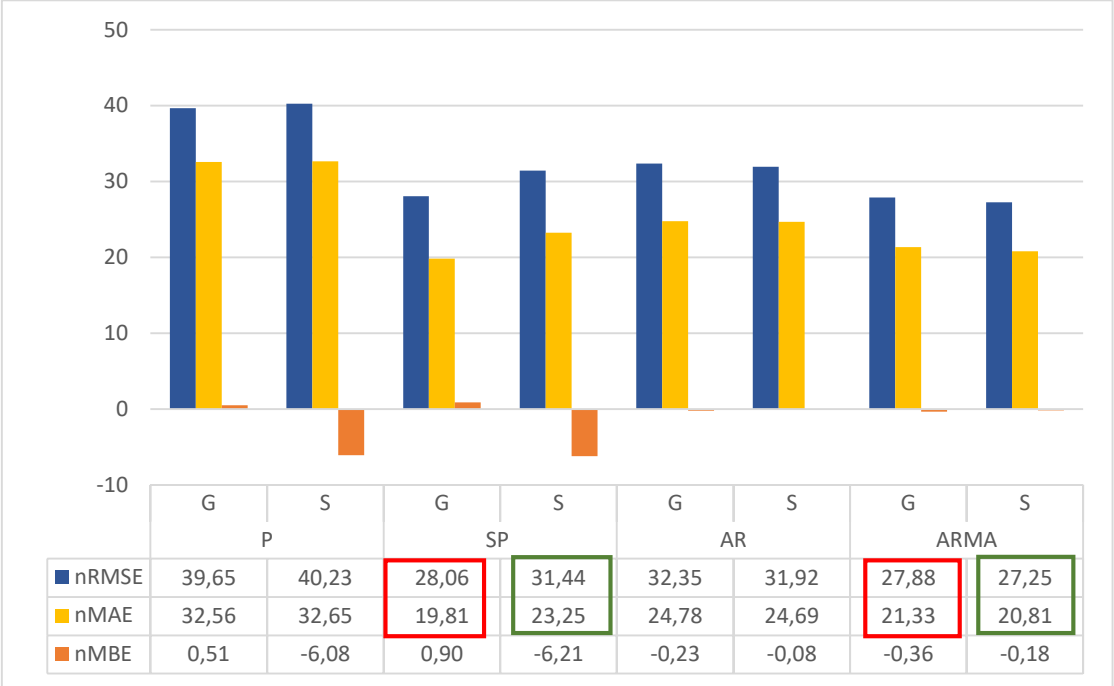


Figure 33: nMBE forecasting errors vs. time Horizon for Ground measurement (G), Satellite-derived data (S) with four statistical models in 2016 according to station.

We note that using ground data as inputs to the persistence-based method leads to an overestimation of the observation. For P model, the nMBE values range from 0.4% to 1.4% and 1.7% to 16.3% depending on the station for 1H to 6H horizon, respectively. While using satellite-derived data as input to P model leads to underestimation of the observation. Regarding the regression-based model, we note that the two models perform the same way in overestimating or underestimating the observations, depending on the station.

Next we present the performance of the models using the data of the six stations together. Forecast results of six stations together for a one-hour horizon are recorded in Figure 34. Figure 34 shows that when the ground or satellite-derived data are used as inputs for the regression models, the differences between the forecast error estimate are negligible. Moreover, the difference in nRMSE and nMAE between the forecast error estimate using the persistence model when the ground or satellite-derived data are used as inputs are small. For 1H time horizon, the less performing model is persistence (P) with values of

nRMSE and nMAE about 39.65%, 32.55% for ground data, respectively, and 40.23%, 32.65% for satellite-derived data, respectively. For 1H, the Scaled Persistence (SP) model with ground data is more accurate than satellite-derived data, with nRMSE and nMAE of 28.06%,19.80%, and 31.44%, 23.24%, respectively.



	G	S	G	S	G	S	G	S
	P		SP		AR		ARMA	
nRMSE	39,65	40,23	28,06	31,44	32,35	31,92	27,88	27,25
nMAE	32,56	32,65	19,81	23,25	24,78	24,69	21,33	20,81
nMBE	0,51	-6,08	0,90	-6,21	-0,23	-0,08	-0,36	-0,18

Figure 34: 1H forecasting errors for Ground measurement (G), Satellite-derived data (S) with four statistical models in 2016.

Figure 34 shows that using satellite-derived data with the four statistical models underestimates the observation with varying values between the persistence-based method and the regression-based method. Where P and SP models using satellite-derived data underestimate the data by about 6.07%, and 6.2% respectively. AR and ARMA models using satellite-derived data underestimate the data by about 0.07% and 0.18%, respectively.

Results reveal that for satellite-derived data, the model that shows the best overall performance is ARMA. For the second and third models, the choice is more delicate between AR and SP with a small difference in the forecast error estimate. We may explain the poor performance of AR due to its prediction method, as we obtain forecast by a linear combination of the past hour observation.

The performance of several forecasting models using ground and satellite-derived data for multistep-ahead in French Guiana is shown in Figure 35.

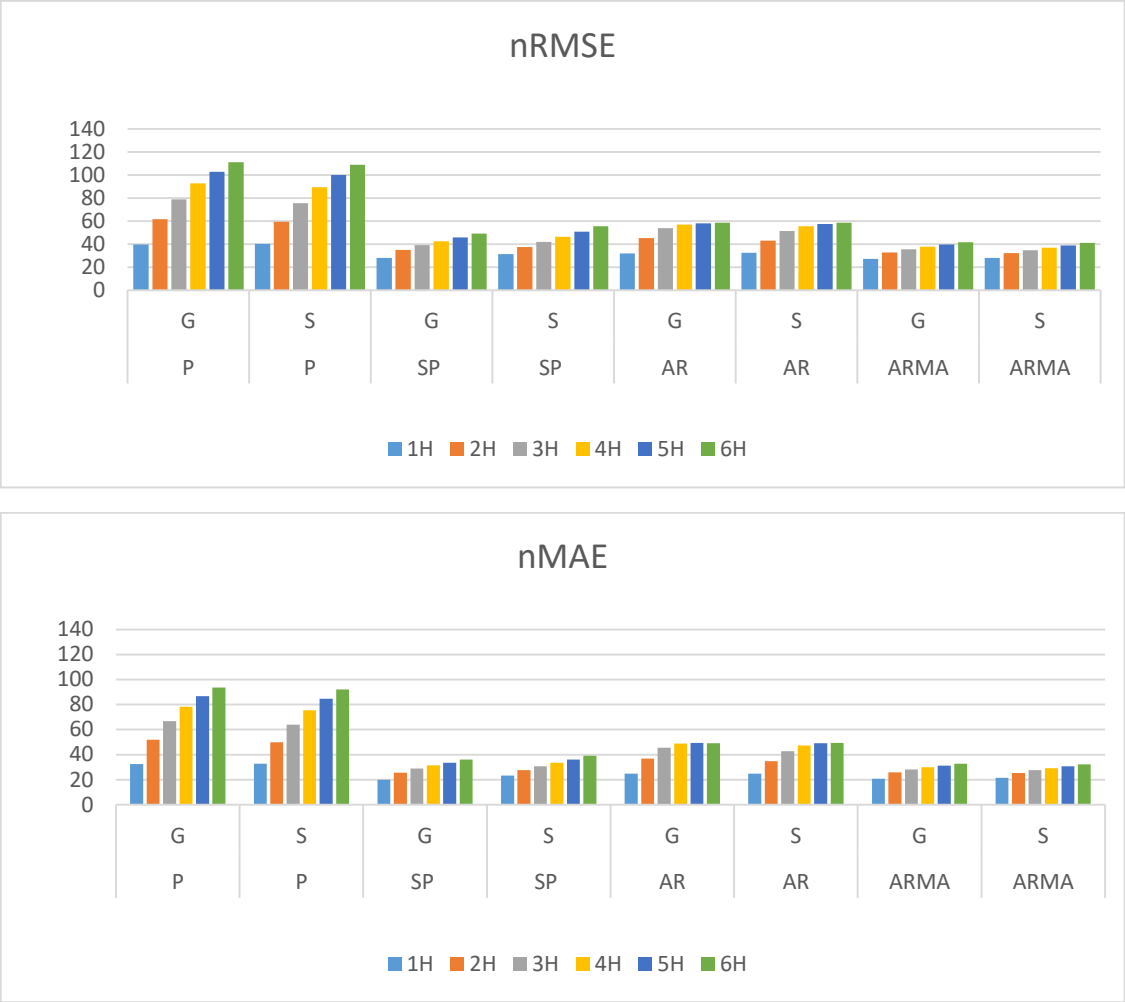


Figure 35: nRMSE, nMAE forecasting errors vs. Horizon for Ground measurement (G), Satellite-derived data (S) with four statistical models in 2016 for 6 time horizons.

Using satellite-derived data, the difference between the best-performing model ARMA model, and the reference P varies from 27.25% to 40.23% for 1H time horizon, and from 41.72% to 109.05% for 6H time horizon, respectively. As we mentioned before, this difference can be taken into consideration to justify the use of a complex model as ARMA. Forecasting errors show that models perform better with ground data for one-hour time horizon. However, for horizons over 2H, satellite-derived data combined with regression-based models are more suitable to make the forecast than with ground one. According to the time horizon, models tend to vary between over- or underestimate the observation, Figure 36.

	G P	S P	G SP	S SP	G AR	S AR	G ARMA	S ARMA
1H	0,51	-6,08	0,90	-6,21	-0,08	-0,23	-0,18	-0,36
2H	1,61	-4,97	1,47	-6,26	-0,13	-0,23	0,00	-0,28
3H	2,87	-3,54	1,88	-6,34	-0,39	-0,32	0,59	0,19
4H	3,26	-2,60	1,87	-6,87	-1,03	-0,62	0,94	0,60
5H	3,64	-1,55	1,97	-7,38	-1,71	-0,97	1,16	0,90
6H	4,57	-0,02	2,51	-7,67	-1,22	-0,35	1,88	1,69

Figure 36: nMBE forecasting errors vs. Horizon for Ground measurement (G), Satellite-derived data (S) with four statistical models in 2016.

nMBE of P with ground data overestimate the observation from 0.50% to 4.5% for horizons 1H to 6H, respectively. While P models with satellite-derived data underestimate the observation from -6.07% to -0.02% for horizons 1H to 6H, respectively.

Based on the foregoing, we observed similar results using data from stations separately and together. The decision was made to discuss the results of prediction errors using data from six stations together. A decision that does not conflict with our purpose and spares us complexity and distraction. For the following, we present only the prediction error results using the data from the six stations together and not separately. Results for each of the stations are presented in the ANNEX.

Figure 37 represents the skill score as a function of time horizon considering the (P) as a reference. Negative results mean that the model is worse than the benchmark, and positive results mean it is better. Positive results from this figure show that all Ground-based and Satellite-based models perform better than the reference for all time horizons.

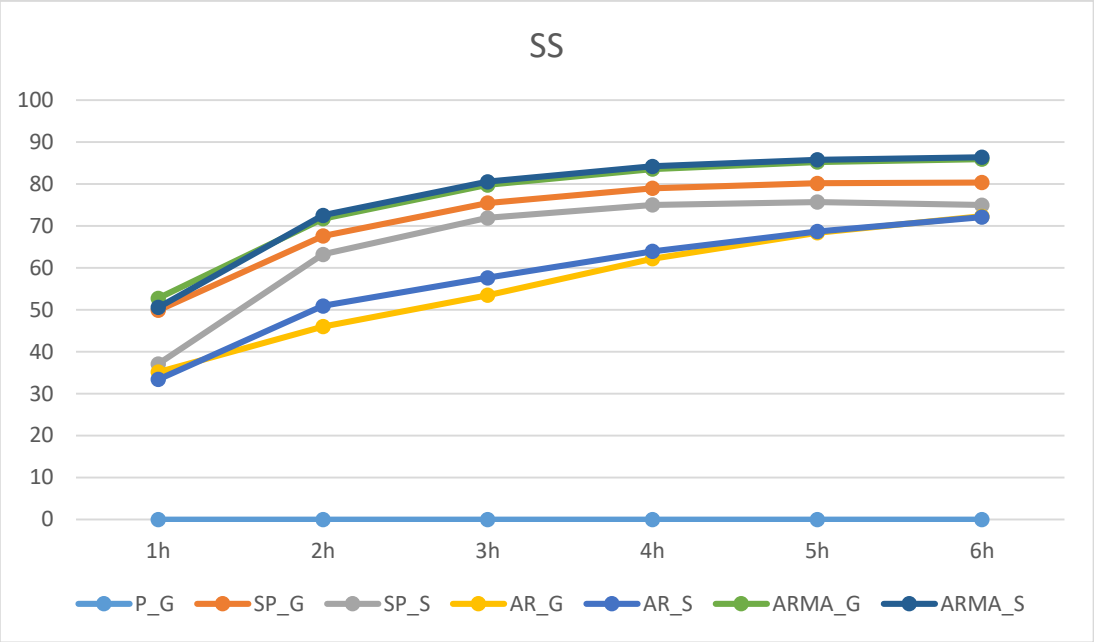
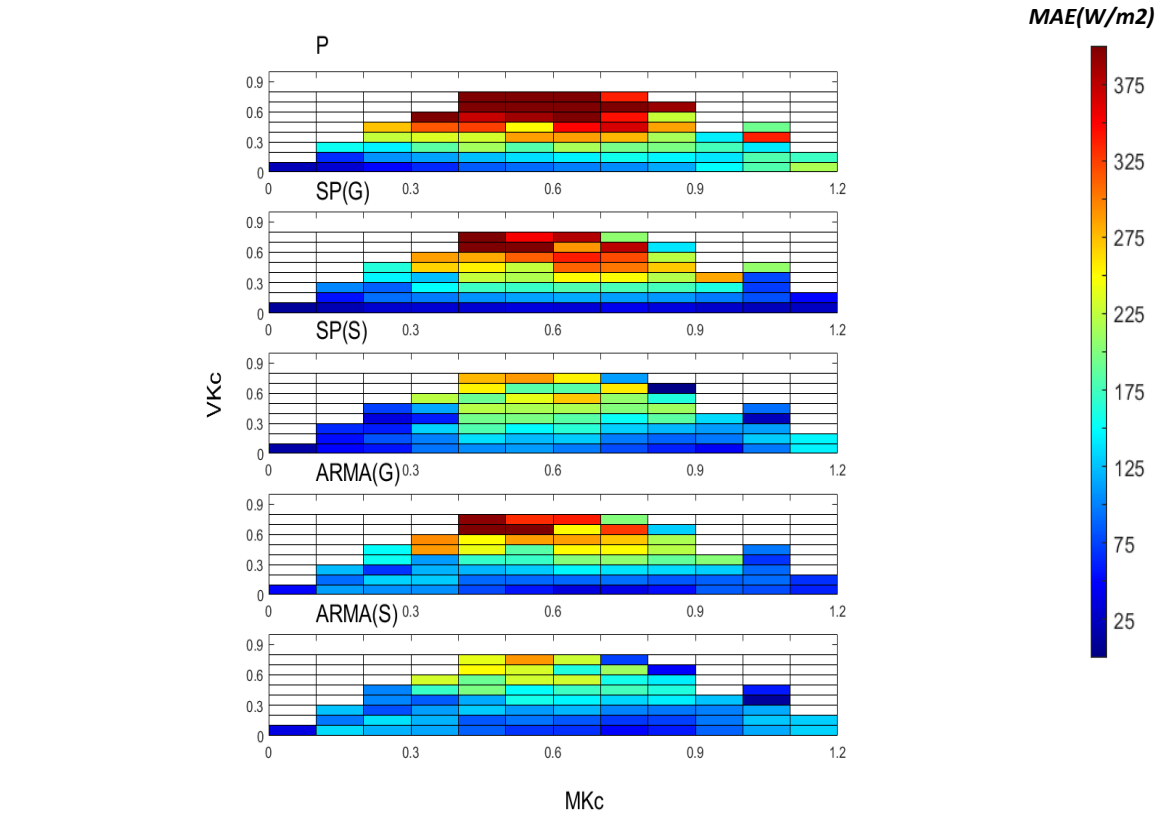


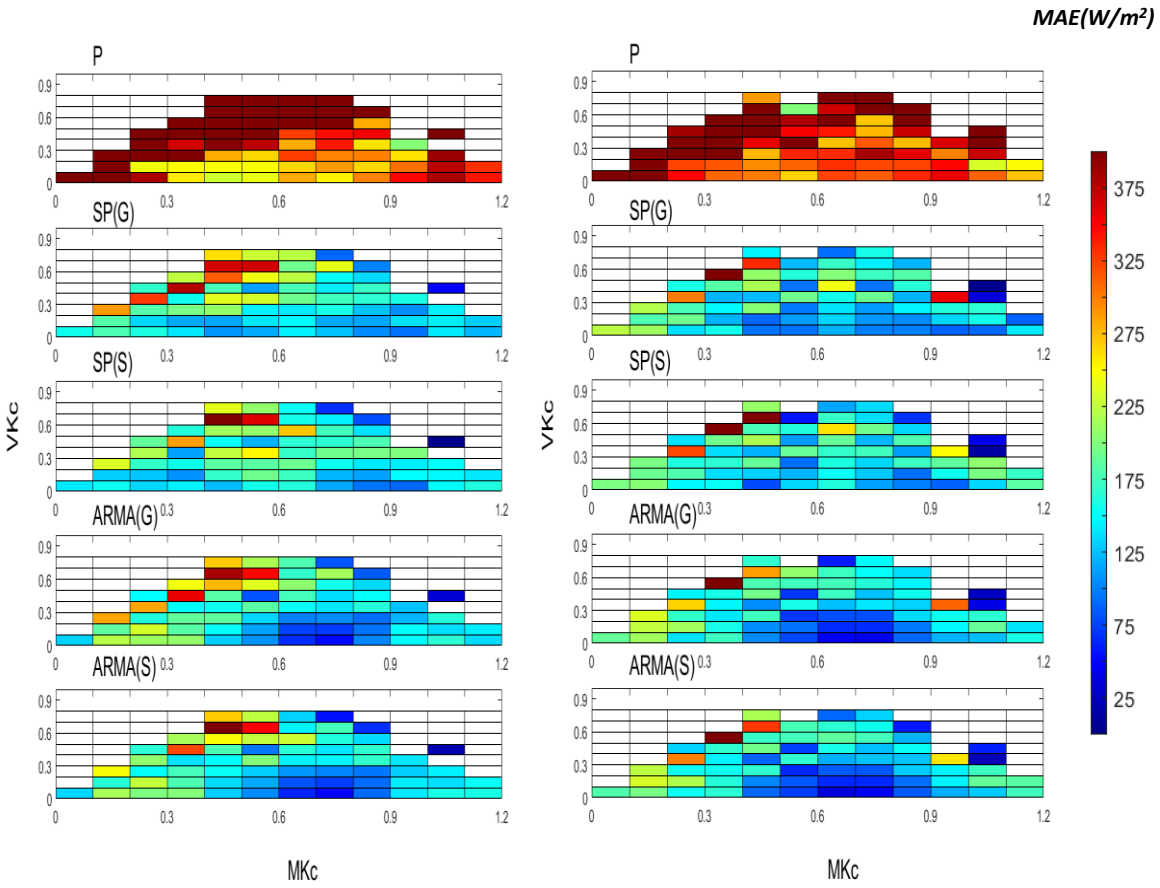
Figure 37: Statistical models' Skill score compared to P according to forecast horizons.

To further analyze the effect of the use of satellite-derived data, it seems interesting to present the MAE value for the reference model and the best two statistical models as a function of Mean clear sky index (MKC) and mean clear sky index variability (VKC) using data from MP station and for three horizons h + 1, h + 3 and h + 6 in Figure 38.

Figure 38, a) shows that the Ground-based models perform better than Satellite-based models for clear sky conditions with low variability i.e. VKc<0.2. Higher MAE values for each model are found under cloudy and highly variable sky conditions i.e. 0.4<MKc<0.65 and VKc>0.2. Note that, under cloudy and highly variable sky conditions the Satellite-based models perform better than Ground-based models. Figure 38, a) also shows that the Ground-based models perform better than Satellite-based models for VKc values lower than 0.2 and independent of the MKc value. By contrast, for VKc values higher than 0.2, the Ground-based model data perform poorer than Satellite-based models.



a) Time horizon= 1H



b) Time horizon= 3H *c) Time horizon= 6H*

Figure 38: MAE for the reference model (Persistence) and the best two models (Scaled Persistence, ARMA) using ground data (G) then satellite-derived data (S) for MP station data in 2016.

Figure 38, b) c) shows that Satellite-based models perform better than Ground-based models for 3 and 6 time horizons, especially under cloudy and highly variable sky conditions. Figure 38, b) c) also shows that the Satellite-based models perform better than Ground-based models for VKc values higher than 0.2 and independent of the MKc value. For clear sky conditions with low variability; the Satellite-based models perform similar to Ground-based models.

The scatter diagrams of forecast irradiation with satellite-derived versus observed data for the ARMA model are illustrated in Figure 39, for six-time horizons.

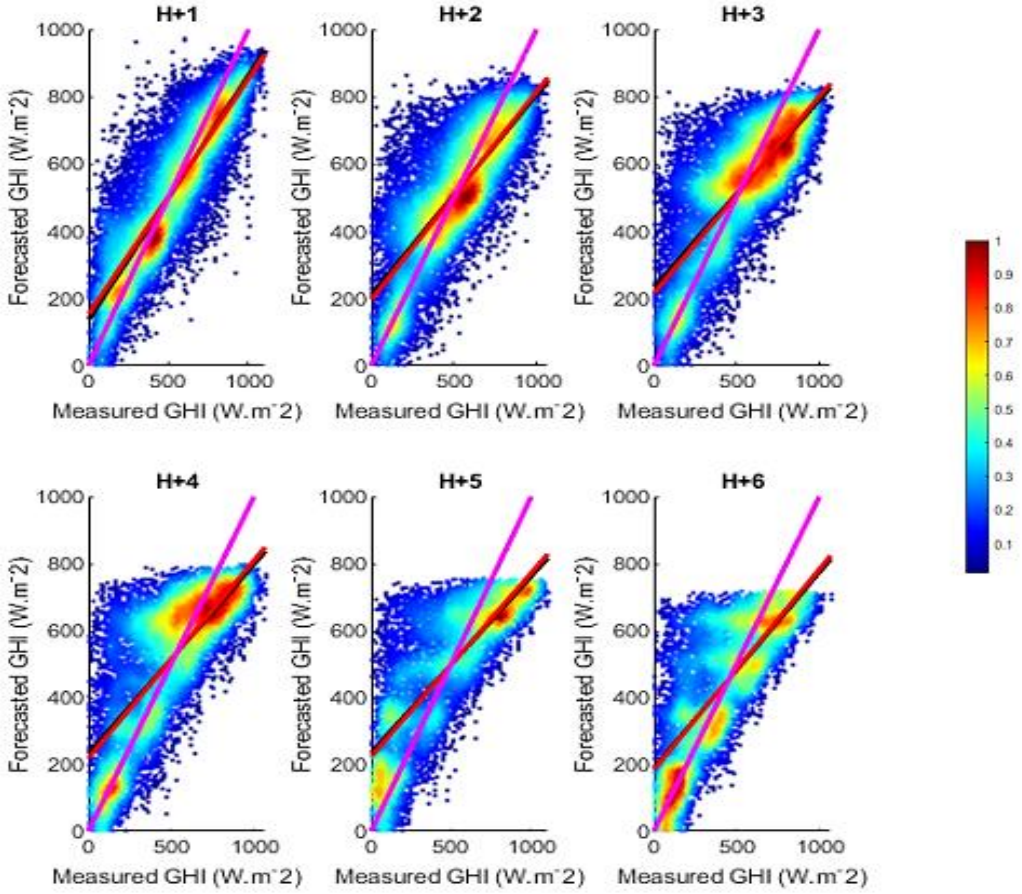


Figure 39: Scatter plot of forecast irradiation with satellite data versus observed irradiation for ARMA model for six-time horizons. The black line represents the GHI forecasted by ARMA using satellite data, the red line is the GHI forecasted by ARMA using ground data and the magenta line depicts the line corresponding to x = y (i.e. perfect correlation).

We note the following:

- The more the time horizon increases, the more the dots are scattered, which seems logical.
- ARMA_G line interposes the ARMA_S line for all horizons, this can be taken as an indication of similar performance.
- Increasing the angle between the ideal predictor and the two ARMA models with the horizon.

- For time horizons over 3H, we can notice that ARMA gives a high value to predict the lowest GHI values. For example, for a time horizon of 6H, to predict the value of 17H (GHI should have a low value where the next hour 18H is the sunset hour) the ARMA model will use the high value of 11H as the past value.
- Even if there are dots whose predicted value is very far from the measured one, however, we can see that most of the values are around the ideal predictor.

Summary

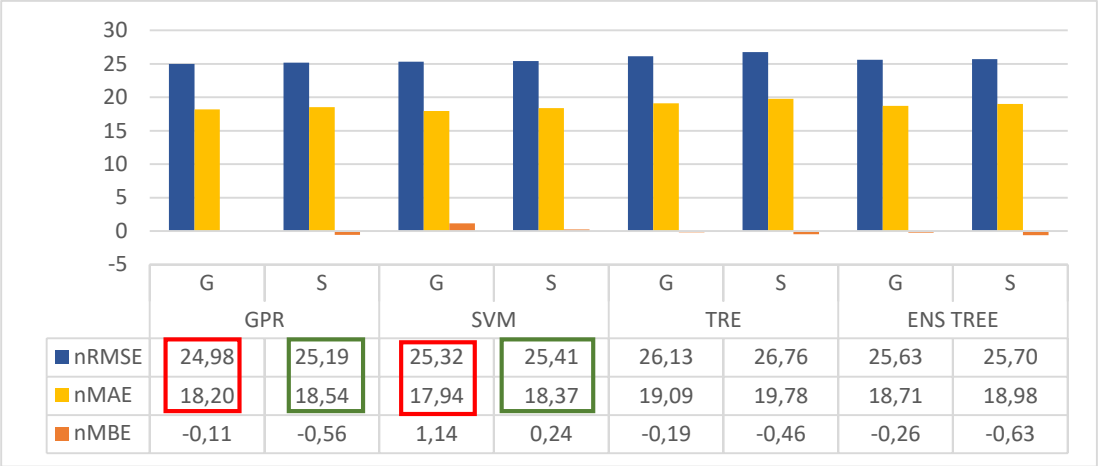
As a conclusion of this paragraph on combining satellite-derived data with four Well-used statistical models, we note the following:

- Persistence (as we expected), does not seem to be suitable.
- AR and ARMA models perform comparably using the two datasets.
- Although the models' performances with satellite-derived data are close. The ARMA model is the most suitable. ARMA model improves the persistence model with nRMSE from 13% to 70% for 1H to 6H ahead, respectively.
- When predictive models give very similar results, the choice of the predictor can be made based on the complexity of the implementation and the computation time.
- The SP model presents interesting performances with satellite-derived data and has the advantage of being easy to implement.
- The models with ground data have better performance than satellite-derived data for 1H time horizon with a difference in nRMSE of 0.5%, whereas regression-based models with satellite-derived data are more suitable to make the forecast than with ground one for horizons higher than 2H.
- For 1H time horizon, Ground-based models perform better than Satellite-based models for low variable sky conditions. For the other sky conditions, the Satellite-based models perform better than ground one.
- Horizons higher than 1H, Satellite-based models perform better than ground one for all sky conditions.

IV.4. Machine learning models for intraday solar irradiation forecast

Geostationary satellite-derived data in French Guiana are used to evaluate the performance of several machine learning techniques in solar irradiation forecast. This section proposes the use of GHI estimations from satellite images with machine learning techniques to forecast hourly GHI in French Guiana. These estimations are combined with exogenous variables related to sun position to be used as inputs to learn and validate machine learning models. Five years (2011 to 2015) of hourly GHI data from six stations are used for the training phase of four machine learning techniques: Gaussian Process Regression (GPR), Support Vector Machine (SVM), Regression Trees (TRE), Ensembles of Trees (ENsTRE). Generalizability and stability were evaluated using different sizes of training data with 5-fold analysis. These models have been selected as regression techniques to be evaluated, due to their good performance in similar problems in the past. Hourly GHI data of 2016 are used for the validation phase. We study the intraday-forecasting performance for the validation process. One to six hours ahead forecasts were obtained using satellite-derived then ground measured data of 2016. Model's hyperparameters were obtained by optimization, for more details see Chapter .III. Forecasting errors results according to time horizon and station using ground-measured data and satellite-derived data with the four machine learning models are presented in Annex *A*. In this section only results using data from the six stations will be discussed.

Figure 40, shows one hour ahead of forecasting performance. We can observe that nRMSE and nMAE values between ground measured and satellite-derived data for GPR, SVM, and ENsTRE models are close. Combining satellite-derived data with simple TRE model gives a slightly less accurate performance than with ground measured data. The nRMSE results show that combining satellite-derived data with GPR gives a really close forecasting performance to the model combined with ground one. The nRMSE and nMAE between the best model GPR and the reference model P using satellite-derived data range from 25.1% to 40.23%, and 18.54% to 32.56%, respectively (see persistence model results in 0). Moreover, we cannot ignore the very identical results for SVM and ENsTRE between ground measured and satellite-derived data. According to forecast performance, SVM is the second best model. Whereas, simple TRE model based on simple regression trees gives poor results due to its high probability of overfitting.



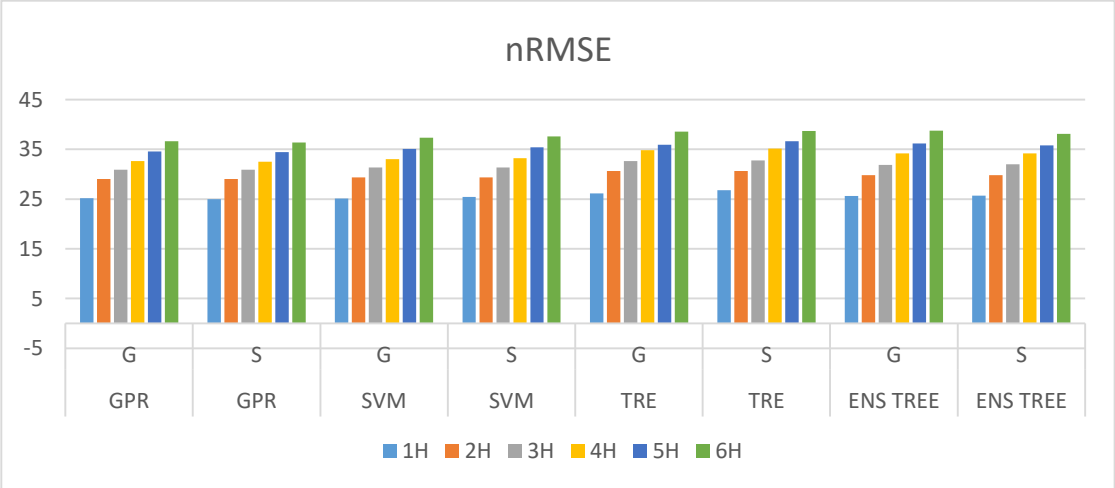
	G	S	G	S	G	S	G	S
	GPR		SVM		TRE		ENS TREE	
nRMSE	24,98	25,19	25,32	25,41	26,13	26,76	25,63	25,70
nMAE	18,20	18,54	17,94	18,37	19,09	19,78	18,71	18,98
nMBE	-0,11	-0,56	1,14	0,24	-0,19	-0,46	-0,26	-0,63

Figure 40: 1H Horizon forecasting errors for ground measured and satellite-derived data of 2016 with several machine learning models, Ground measurement (G), Satellite-derived data (S).

Observing nMBE in Figure 40, we find that SVM with satellite and ground data trend to overestimate the observation. Whereas the other models underestimate the observation. From nMBE, we can say that satellite and ground combining with machine learning models behave in the same way as concerning underestimate or overestimate the data.

Results reveal that for satellite-derived data, GPR is the model that shows the best overall performance from both nRMSE and nMAE. For the second and third models, the choice is more delicate between SVM and ENsTRE with really small differences in the result.

The next step is to plot the forecasting errors against the horizon to illustrate the performance of these four machine learning models for multistep-ahead in French Guiana, as in Figure 41. The graph shows the accuracy metrics of the validation process using ground then satellite-derived data from six meteorological stations of 2016 as input.



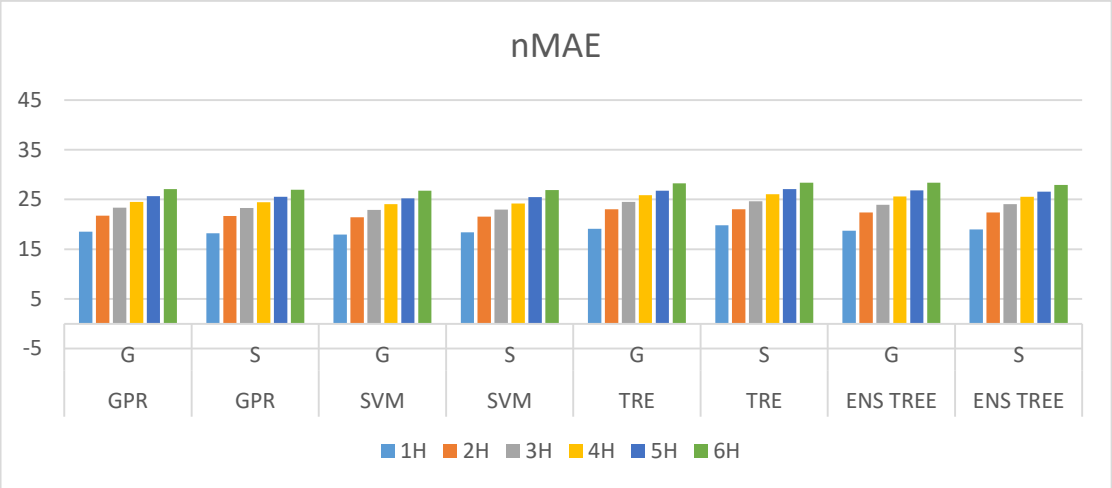
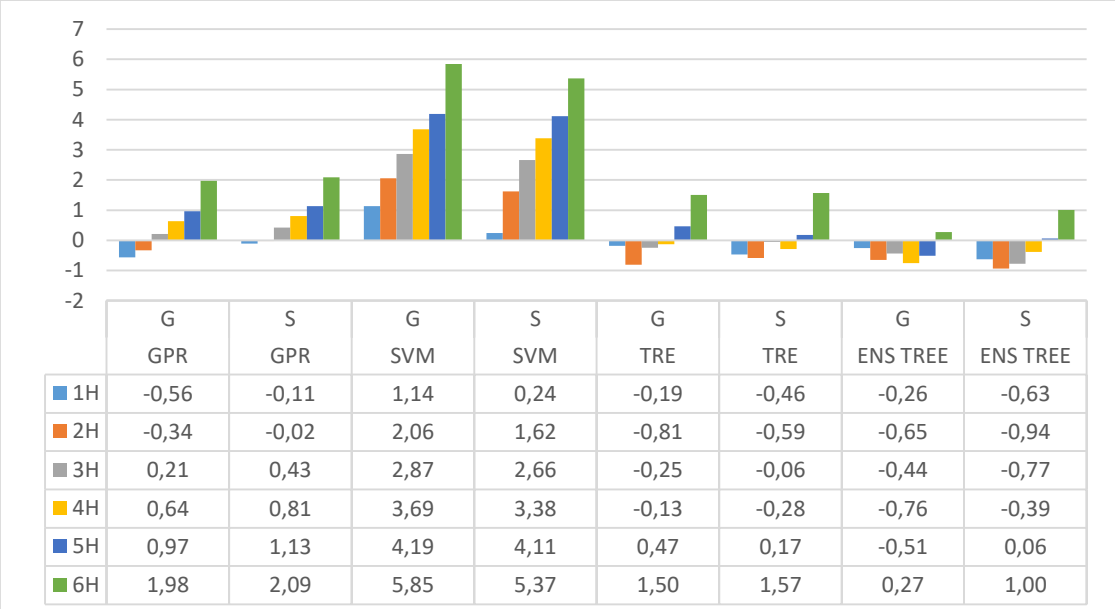


Figure 41: nRMSE, nMAE forecasting errors vs. Horizon for Ground measurement (G), Satellite-derived data (S) with several machine learning models for all stations of 2016.

We can observe a good performance with all machine learning techniques, where the evolution is linear relative to the forecast horizon. GPR model has the best nRMSE for all time horizons. Moreover, the GPR model has very identical performance combining models with satellite-derived data and ground one from 1 to 6H.



	G GPR	S GPR	G SVM	S SVM	G TRE	S TRE	G ENS TREE	S ENS TREE
1H	-0,56	-0,11	1,14	0,24	-0,19	-0,46	-0,26	-0,63
2H	-0,34	-0,02	2,06	1,62	-0,81	-0,59	-0,65	-0,94
3H	0,21	0,43	2,87	2,66	-0,25	-0,06	-0,44	-0,77
4H	0,64	0,81	3,69	3,38	-0,13	-0,28	-0,76	-0,39
5H	0,97	1,13	4,19	4,11	0,47	0,17	-0,51	0,06
6H	1,98	2,09	5,85	5,37	1,50	1,57	0,27	1,00

Figure 42: nMBE errors vs. Horizon for Ground measurement (G), Satellite-derived data (S) with four machine learning models for all stations of 2016.

Figure 42 shows that models tend also to vary between over- or underestimate the observation according to the time horizon. Yet, satellite-based or ground-based models behave the same according to over- or underestimate the observation. For best model GPR, nMBE with ground data increases from -0.50% to 1.9% from 1 to 6H, respectively, with satellite-derived data nMBE decreases from -0.11% to -0.024 for 1H to 2H, respectively. Then, nMBE increases from 0.42 to 2.08% for 3 to 6H.

Figure 43 shows the skill score results of the different machine learning models compared to the P model and according to the forecast horizon. Positive results from this figure show

that all models perform better than the reference model for all time horizons. Even though the four models perform similarly with the ground and satellite-derived data, GPR and SVM perform slightly over the others compared to the P model.

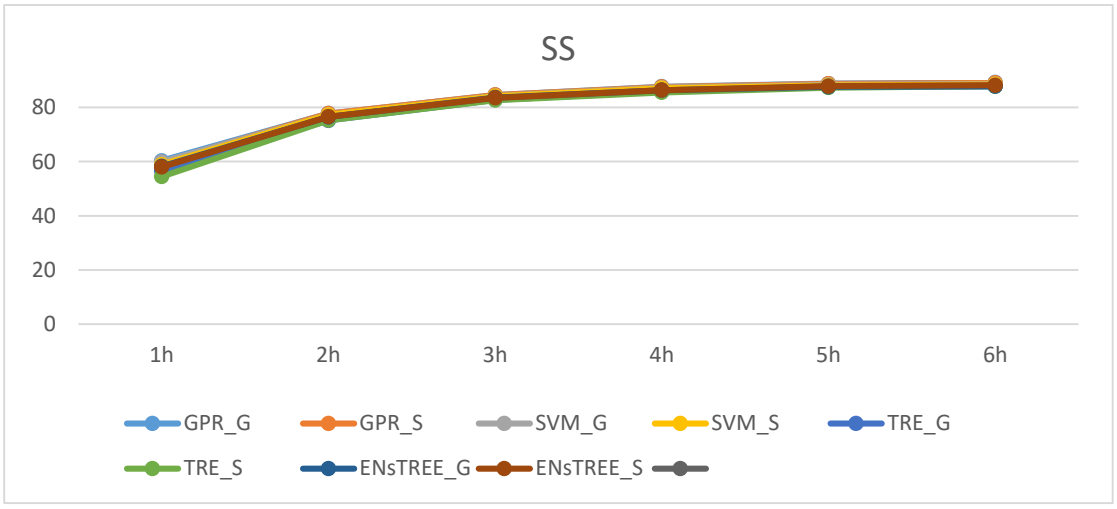
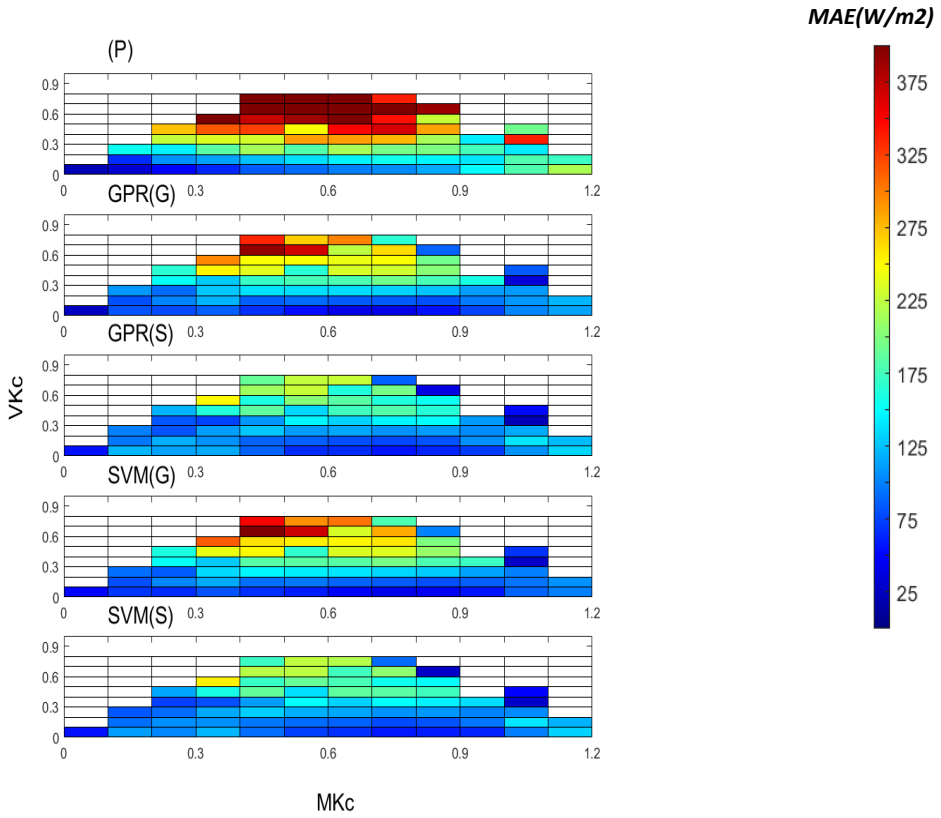
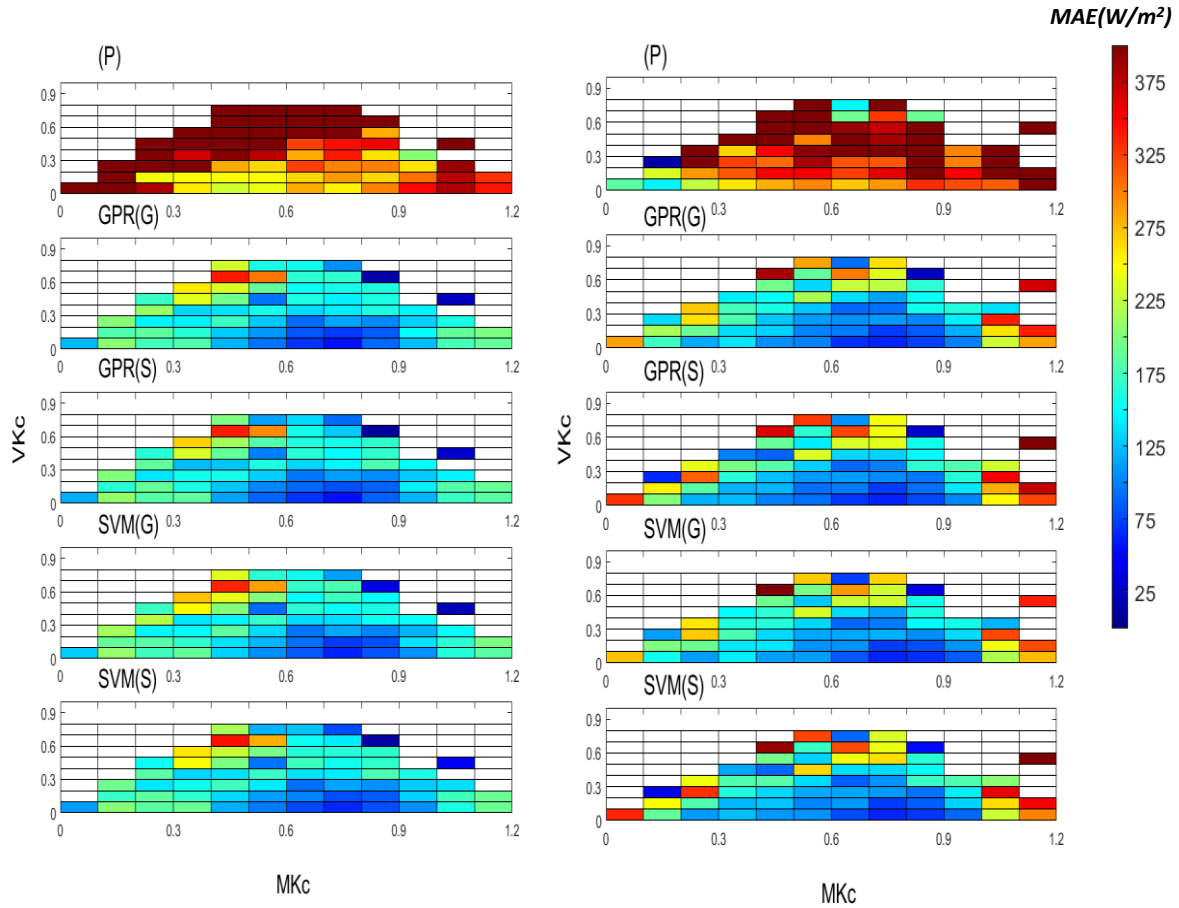


Figure 43: Machine learning models' Skill score compared to P according to forecast horizons.

To further analyze the effect of the use of satellite-derived data, it seems interesting to present the MAE value for the reference model and the best two machine learning models as a function of MKC and VKC using data from MP station and for three horizons h + 1, h + 3 and h + 6 in Figure 44.



a) Time horizon= 1H



b) Time horizon= 3H *c) Time horizon= 6H*

Figure 44: MAE for reference model Persistence and best two models Gaussian Process Regression (GPR), Support Vector Machine (SVM) using ground data (G) then satellite-derived one (S). For MP station data in 2016.

Figure 44, a) shows that for clear sky conditions with low variability the Ground-based machine learning models perform better than Satellite one i.e. MKc>0.65 and VKc<0.2. Under cloudy and highly variable sky conditions, we find higher MAE values for each model. Note that, Satellite-based machine learning models perform better than Ground one under cloudy and highly variable sky conditions. Figure 44, b) c) shows that models behave in a convergent manner with slight differences. Satellite-based models perform better than Ground-based models for 3 and 6 time horizons, especially under cloudy and highly variable sky conditions. For 3H forecast Satellite-based machine learning models perform better than Ground one for highly variable sky conditions VKc values higher than 0.2. Figure 44, a) b) and c) also show that forecast accuracy decreases with time horizon but remains much better than P model.

We have presented in Figure 45 the scatter diagrams of forecast irradiation with satellite versus observed irradiation for GPR model for six-time horizons;

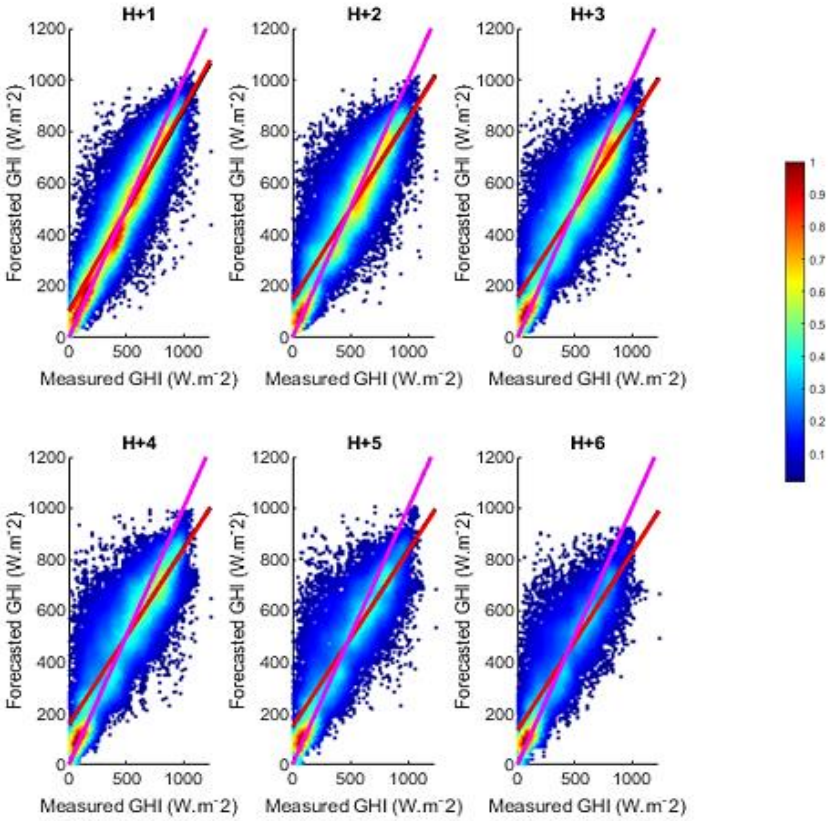


Figure 45: Scatter plot of forecast irradiation with satellite versus observed irradiation for the best model (GPR) for six-time horizons of 2016. The black line represents the GHI forecasted by GPR using satellite data, the red line is the GHI forecasted by GPR using ground data and the magenta line depicts the line corresponding to x = y (i.e. perfect correlation).

From Figure 45, we note the following:

- The more the time horizon increases, the more the dots are scattered; the same behavior as in the previous section.
- GPR model tends to overestimate low values of GHI and underestimate high values for all horizons.
- Horizons more than 2H, the GPR_G line interposes the GPR_S line for all horizons, which can be taken as an indication of similar performance.
- Increasing the angle between the ideal predictor and the GPR models for 1H and 2H ahead, for horizons more than 2H this angle starts to stabilize.
- Even if there are dots whose predicted values are very far from the measured one, however, we can see that most of the values are around the ideal predictor.

Summary

As a conclusion of this paragraph on combining satellite data with machine learning models, we note the following:

- Although the performances are close, the GPR model with a nRMSE of 24.9% to 36.9% is the most suitable to be combined with satellite data, followed by SVM with a nRMSE of 25.4% to 37.6% for horizons 1H and 6H, respectively;
- The simple TRE regression model is not suitable;
- The uncertainties are much smaller than for the previous case;
- The forecast errors between the satellite and ground data are similar regardless of the horizon considered;
- The models have similar nMBE performance to overestimate or underestimate the observation with ground measured and satellite-derived data.
- The Gaussian Process Regression (GPR) and support vector machine with K-fold cross validation model is elected for modeling solar radiation for 1 to 6H.
- For 1H time horizon, Ground-based models perform better than Satellite-based models low variable sky conditions. Horizons higher than 1H, Satellite-based models perform better than ground one for all sky conditions.

IV.5. Combining satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting

This section studies a globally applicable approach to improve hourly GHI forecasts of numerical weather prediction (NWP) in French Guiana. This area has a tropical climate regulated by the ITZ and it has not only a high solar potential but also a high variability, this is due to rapid cloud transformations caused by the ITZ. This variability causes a technical challenge to obtain accurate GHI forecasts. For a large area geographic data collection, there is no better observational tool than an NWP. A NWP provides more information up to several days ahead, however, there are significant biases and random errors in the GHI forecasts of NWP. Weather Research and Forecasting model (WRF) can compute a forecast up to 48 H with high spatial resolution (this model is detailed in III.3.C.i). However, it is well-known that this model's output is highly biased. To reduce this bias, we propose a post-processing technique via a specific Kalman filter. We chose the Kalman filter as a post-processing algorithm because it reduces the bias in the GHI forecasts without the need for a long historical data archive, as it can adapt to changes in weather quickly. We propose using data derived from satellite images as observation for the bias correction technique because satellite images have broader coverage than ground-based irradiation measurements. Our method was validated over a period of three months of 2016 of hourly GHI data from the WRF model. Hourly measured data are collected at six ground stations of the French national weather service.

A global view of the proposed intra-day forecasting method is illustrated in Figure 46. Indeed, it combines in a statistical model, namely the linear Kalman filter, the outputs of a numerical forecasting model, the satellite estimation data, and the cosine of the zenith solar angle made at the experimental sites.

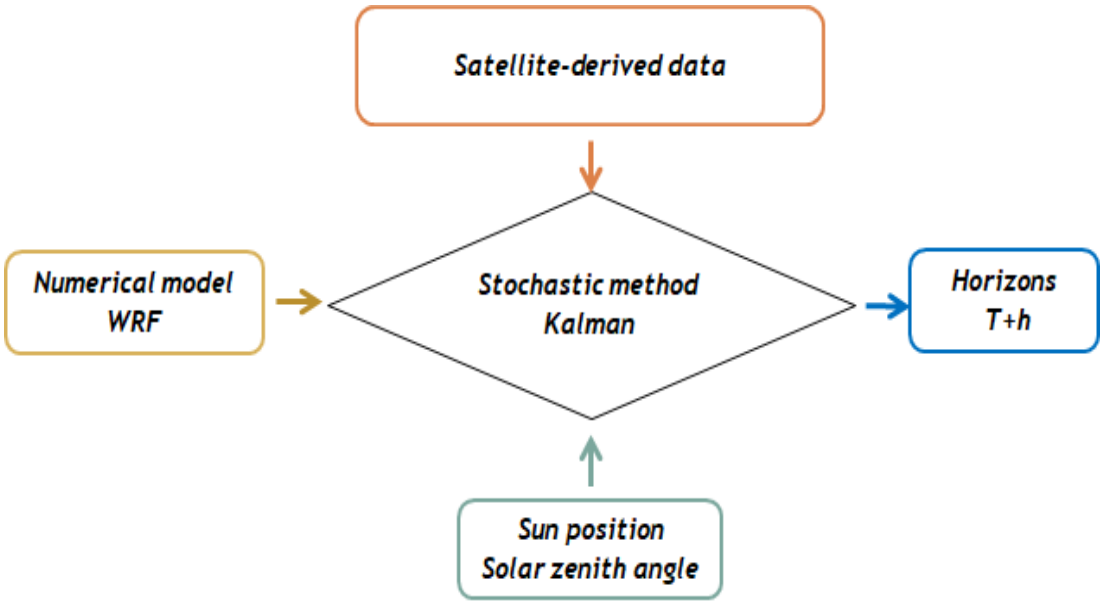


Figure 46: Global view of the forecasting approach for horizon 1H to 6H.

As we mentioned before, three months of 2016 were selected in French Guiana. September, May, represent the dry and rainy seasons, respectively. The third month March describes the latency period between the ITCZ southward to the northward motion over French Guiana. Depending on the selected year March either behaves as a rainy or a dry season month. In this section, WRF prediction errors before and after processing using ground-measured data and satellite-derived data from all stations are discussed. More details on WRF prediction errors by station and time horizon can be found in Annex B.

For all stations, hourly GHI forecasts were tested against ground-based irradiation measurements in Figure 44. WRF post-processing method via kalman filter improves WRF forecasts. For all-sky conditions, the post-processing method using satellite-derived

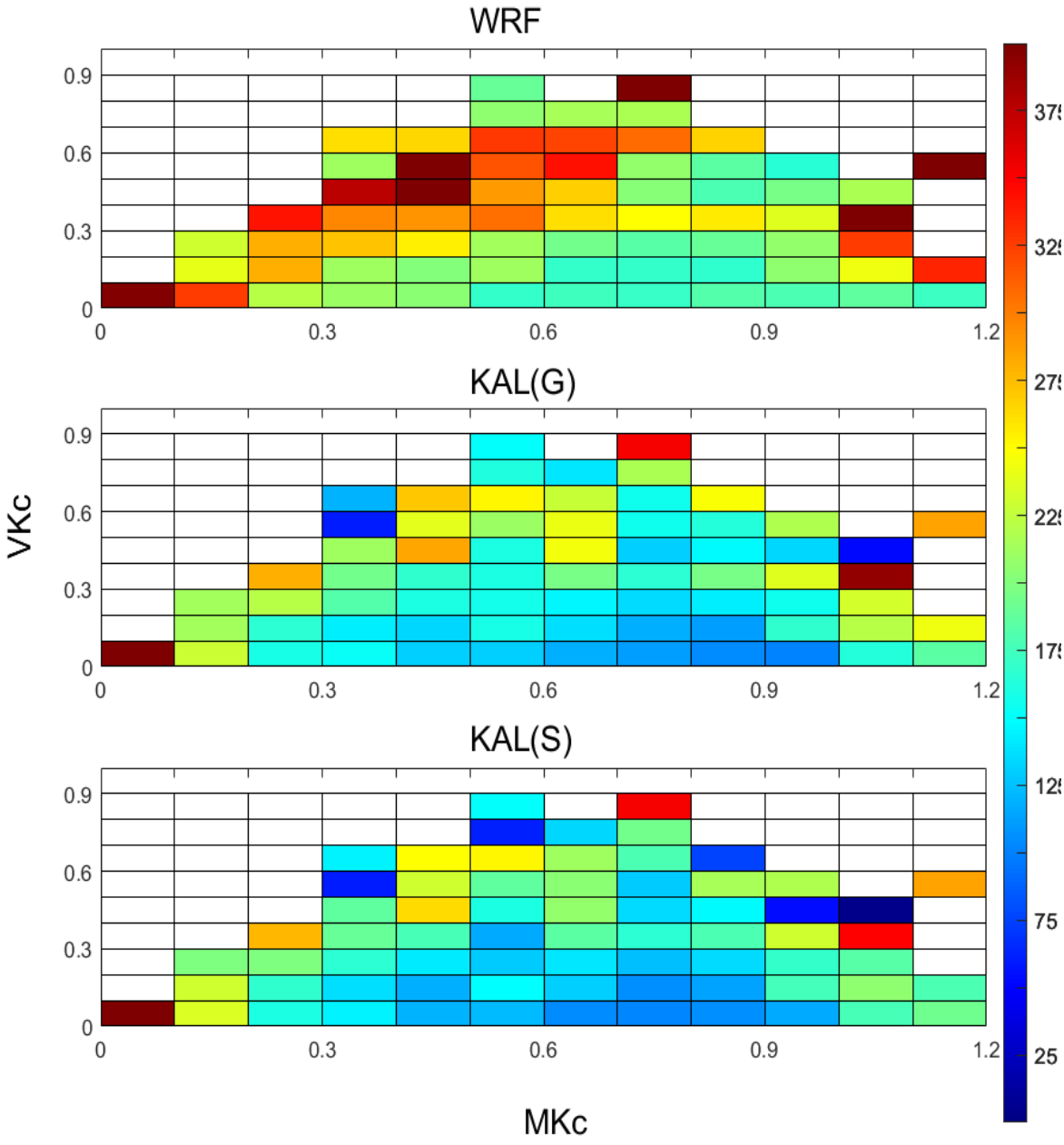


Figure 47:1H MAE for WRF, WRF Kalman with ground data, WRF Kalman with satellite-derived data. Three months of 2016.

data performs close to the post-processing method using ground-based irradiation

measurements. Satellite-derived data perform better than ground one for clear sky conditions with low variability. Moreover, we note a better performance with satellite-derived data for high variability and independent of sky condition.

Multi-horizon forecast errors for post-processing WRF with ground and satellite-derived data in Figure 48. nRMSE nMAE and nMBE of WRF and the post-processing methods show that the post-processing method using satellite-derived data performs similarly to that using ground-based irradiation measurements for all time horizons. We also found from Figure 48, Hourly post-processing of GHI forecasted with Kalman filtering using satellite gave better results than direct WRF output. Using satellite-derived data as observation decreases the nMBE of WRF forecasts to more than the half, from 9.93% 15.4% to 3.7% 8,5% for 1H to 6H ahead, respectively. For all stations, we found that using satellite-derived data as observation decreases nRMSE, nMAE and nMBE of WRF from 1 to 6H ahead.

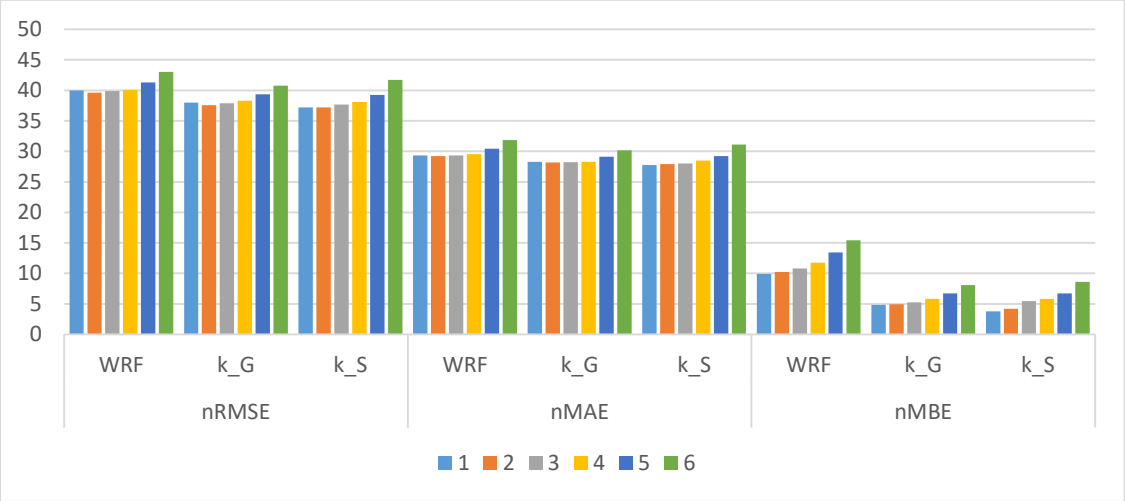
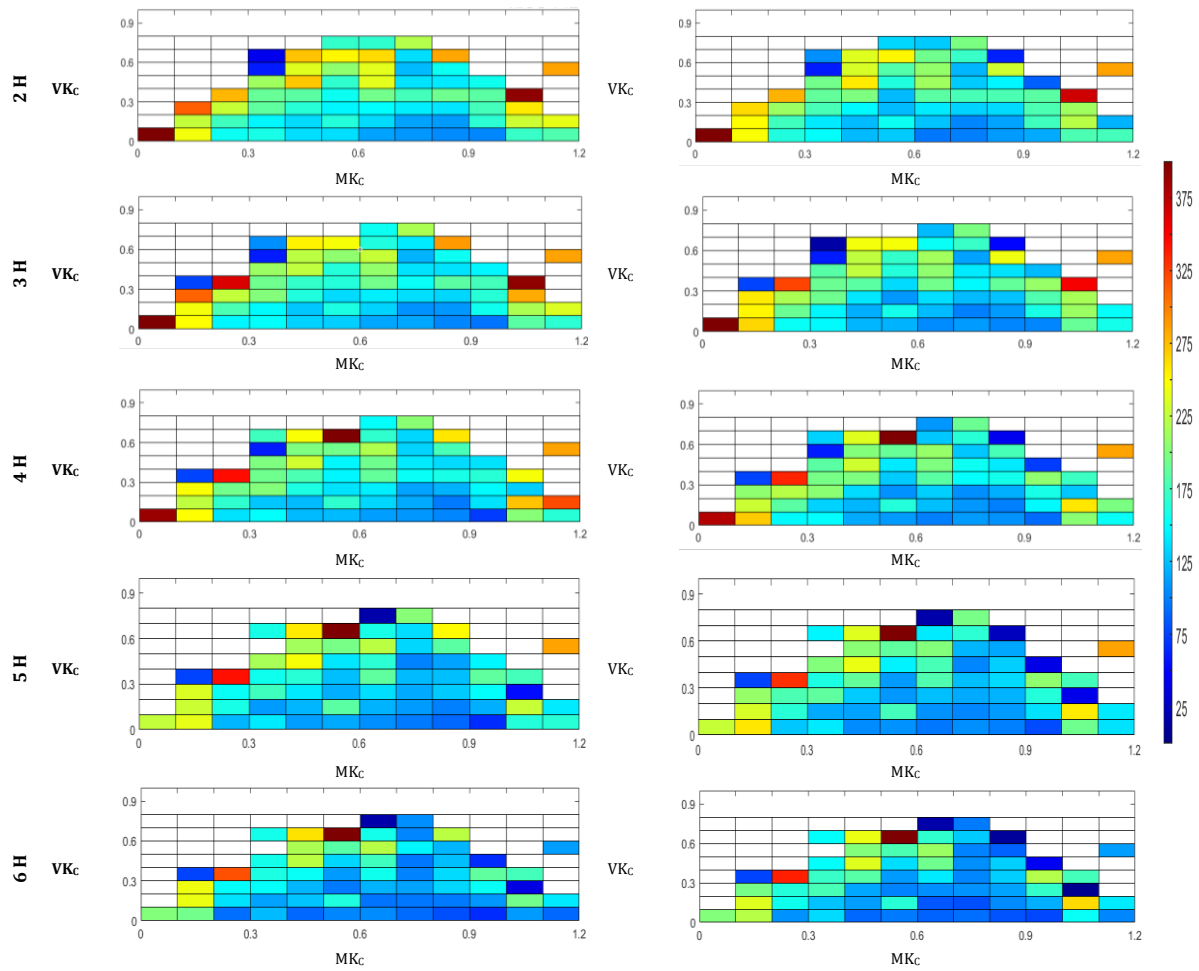


Figure 48: nRMSE nMAE and nMBE for post-processing WRF with ground data k_G and satellite-derived data k_S for horizons of 1, 2, 3, 4, 5, 6 for three months of 2016.

The objective of the next step is to study the effect of post-processing of the WRF model using the satellite. The graphical representation of MAE depending on the MKC and VKC according to the time horizon for each model is shown in Figure 49. Results from post-processing WRF with ground data, and with satellite data are shown in Figure 49.

Using ground-based irradiation measurements with the Kalman filter improves WRF forecasts for all time horizons. The first and second columns in Figure 49 are the application of the Kalman hybrid method using ground and satellite-derived data, respectively. We can notice that the post-processing of the WRF model with satellite-derived data has lower MAE than that with ground one for clear sky conditions with high variability. We note a close performance of ground data with satellite-derived data for low variability and independent of sky condition.



b) Kalman filter using ground data *c) Kalman filter using satellite-derived data*

Figure 49: 2 to 6H MAE for WRF Kalman with ground data, WRF Kalman with satellite-derived data. Three months of 2016.

The scatter plot of hourly forecasted GHI with satellite-derived data versus hourly measured GHI in Figure 50 shows a comparison between post-processing WRF with ground data and post-processing WRF with satellite-derived data results.

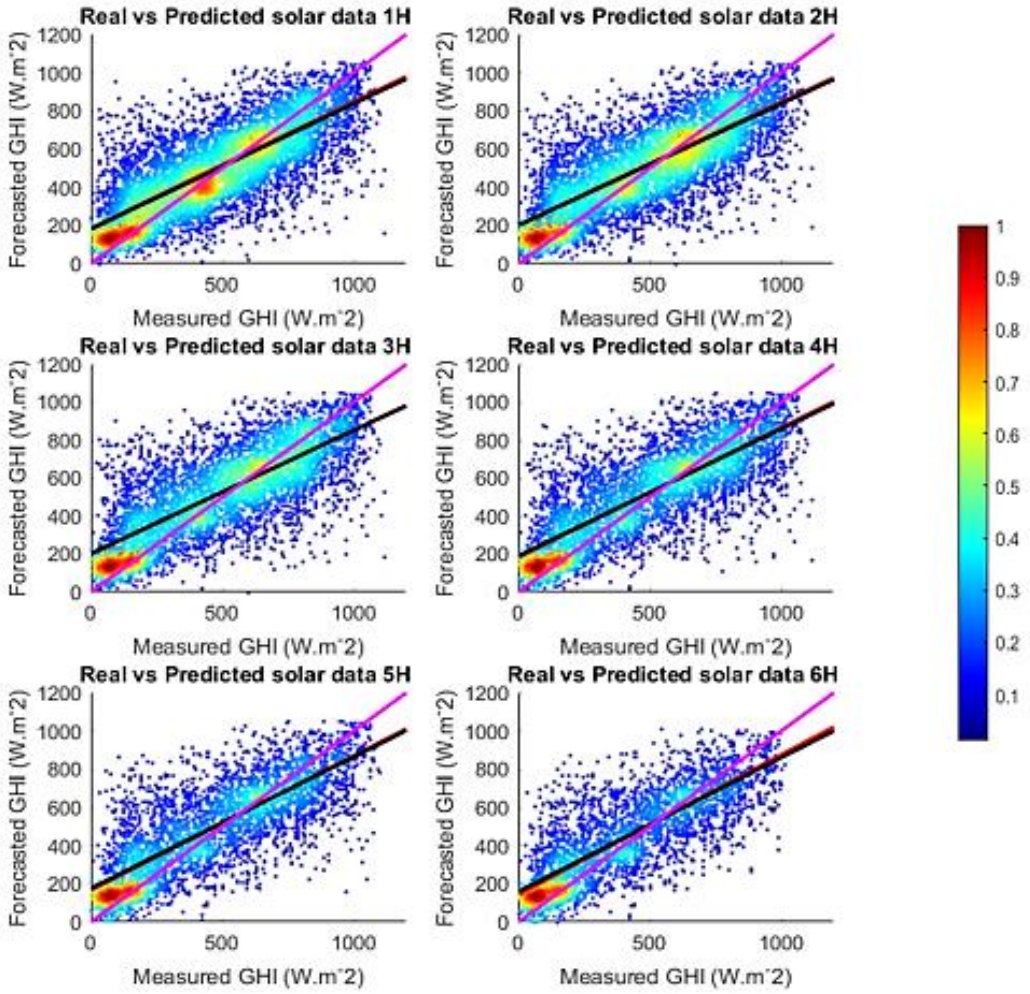


Figure 50: Scatter plot of measured and post-processing WRF forecasting according to satellite-derived data for horizons of 1 to 6H, for three months of 2016. The black line represents the GHI forecasted by Kalman filtering using satellite data, the red line is GHI forecasted by Kalman filtering using ground data and the magenta line depicts the line corresponding to x = y (i.e. perfect correlation).

We notice that for both post-processing techniques, high GHI values are underestimated. Moreover, low values of GHI are overestimated and forecasts are very scattered. It is due to the variability of the zone of study. Even if there are dots whose predicted value is far from the measured one, however, we can see that most of the values are around the ideal prediction line.

Summary

In conclusion of this study on post-processing of GHI forecasted with WRF using satellite for intra-day forecast, we note:

- Kalman filter improves the 24H forecast of WRF at hourly time steps, for horizons of 1H to 6H using ground or satellite datasets in ITZ.
- Forecast results obtained using satellite-derived data were compared to those obtained from measured data. We found that the post-processing method using satellite-derived data performs similarly as when measured data are used.

- Kalman filter performance by time horizon decreases slightly compared to the previous time horizon case.
- For all-sky conditions and six stations data, we found that using satellite-derived data as observation decrease nMBE by approximately 7 - 3% according to the time horizon.
- Using satellite-derived data as an observation in the WRF post-processing technique decrease MBE by approximately 50% according to the time horizon. This is an interesting case because it will allow to reduce WRF bias on any geographical location in French Guiana and to maintain the advantage of WRF's high spatial resolution.
- A global post-processing technique of WRF by satellite-derived data is successfully applied.

IV.6. Hybrid method using satellite-derived data to improve intra-day solar irradiation forecast

The objective of this section is to improve the intra-day GHI forecasts in the ITZ using satellite-derived data and ensembles of forecasts from several models in French Guiana. First, we produce GHI forecasts from four statistical and machine learning methods using satellite estimations. The four models are SP, ARMA, GPR, SVM. We used 5 years of hourly historical GHI to train models and compute the model's parameters. Second, we acquired WRF forecasts in French Guiana (Mouhamet 2018), we post-processed these WRF outputs using a Kalman Filter bias correction method (Pelland, Galanis, and Kallos 2013) to refine the forecast of WRF. These WRF corrected outputs are used as an ensemble new member. Hourly GHI Satellite estimations for three months of 2016 are used to correct forecasted values from WRF. Each model was performed at different horizons (1H, 2H, 3H, 4H, 5H, and 6H). In the last step, we apply a ridge regression method (Thorey et al. 2015) to aggregate each ensemble member's GHI forecasts. Data from SL site used to obtain every model weight. The accuracy of each model is validated by comparison with hourly data of GHI from five stations of the French national weather services as a reference forecast. We used the Adjusted R-square aR^2, to compare the performance of satellite-based models (detailed in section II.3.B). In this section, results are discussed using satellite-derived data from all stations to improve solar forecasts. Results depending on the station and time horizon are shown in Annex C.

Figure 51 presents the aR^2 according to the horizon. We have selected five models to be members of the ensemble method. The five ensemble members are SP, ARMA, GPR, SVM, WRF-Kalman. WRF prediction contains information on the state of the atmosphere in the future because the WRF model numerically solves a system of differential equations that describe the physical and chemical processes of the atmosphere. For this reason, it is interesting to use WRF-Kalman as a member of the ensemble method.

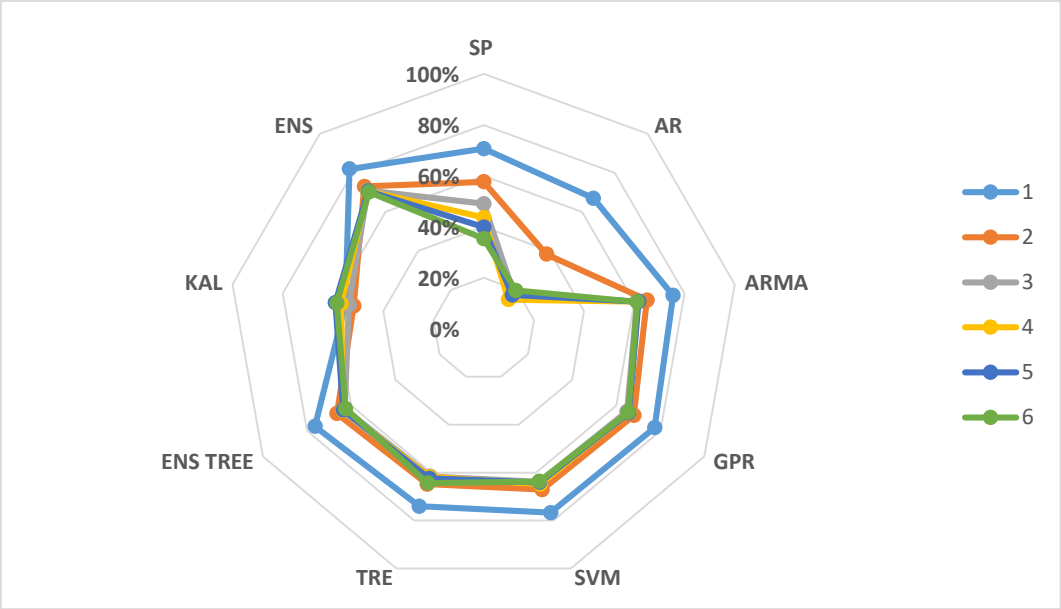
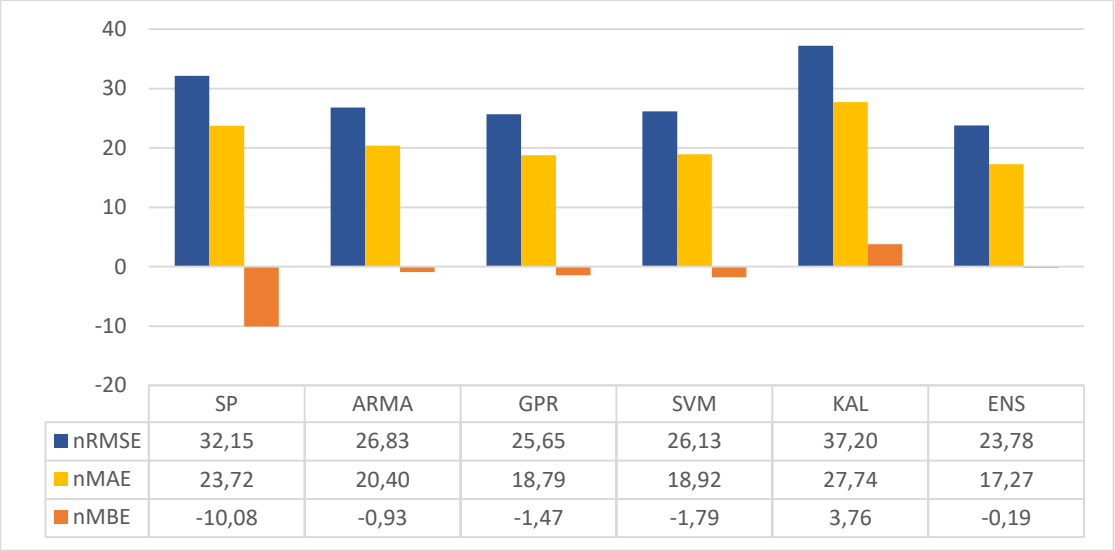


Figure 51: aR^2 according to the horizon, all models, and all stations together.

Figure 52, presents the nRMSE, nMAE, and nMBE of the ensemble aggregation and each ensemble member. This figure shows that the ensemble aggregate obtained using the ridge regression method has lower nRMSE, nMAE, and nMBE values.



	SP	ARMA	GPR	SVM	KAL	ENS
nRMSE	32,15	26,83	25,65	26,13	37,20	23,78
nMAE	23,72	20,40	18,79	18,92	27,74	17,27
nMBE	-10,08	-0,93	-1,47	-1,79	3,76	-0,19

Figure 52: nRMSE, nMAE, and nMBE for each ensemble member and ensemble aggregation. 5 stations were merged; SL is the training station for computing the weights so it was excluded.

nRMSE for ENS shows an improvement of almost 16% compared with the persistence reference model. We found that using a Ridge regression method decreases the normalized forecast bias of the reference model by 10% approximately. Figure 52 also shows that the aggregate tends to underestimate the GHI with approximately equal to 0.2%. Yet, we may conclude that using the ridge regression method to aggregate the ensemble member improves the GHI forecasts. The effect of the aggregation is analyzed in Figure 53, where we present the MAE value as a function of MK_C and VK_C using data from five stations merged. The ensemble member's properties are transferred to the aggregated model. For all sky conditions and low variability, the MAE of the ensemble aggregate is lower than the MAE of each member. For a clear sky with highly variable sky conditions, the MAE of the ensemble aggregate is lower than the MAE of statistical models and similar to the MAE for both machine learning models and the post-processing method. Yet, the MAE of the post-processing method for cloudy sky with highly variable sky conditions has lower MAE than the ensemble aggregate.

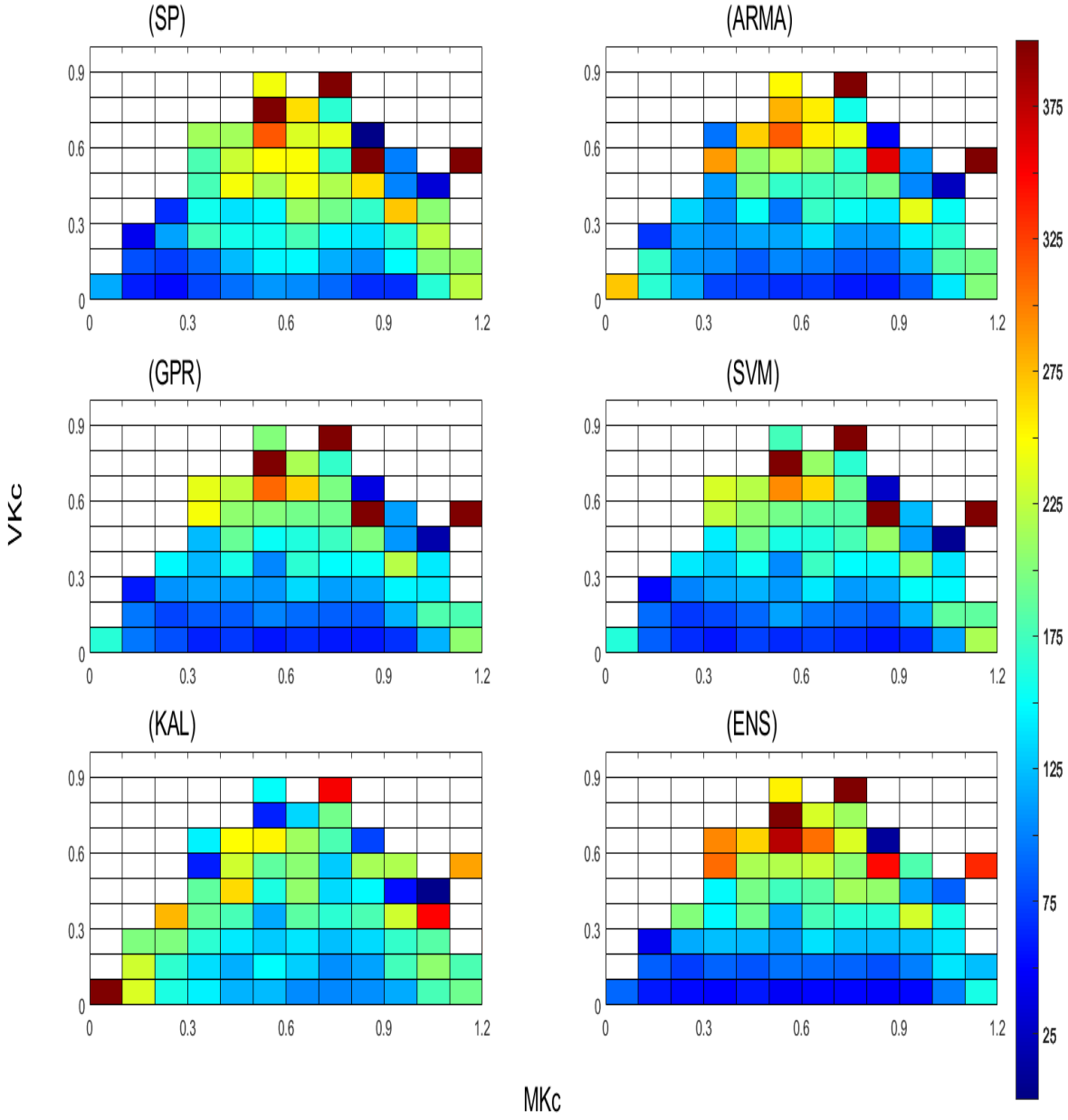


Figure 53: MAE for SP, ARMA, GPR, SVM, WRF-Kalman with satellite-derived data and ensemble aggregation. 5 stations were merged, SL is the training station for computing the weights and was excluded.

The next step is to plot the forecasting errors against the horizon. It is interesting to represent multistep-ahead forecasting accuracy for satellite-derived data graphically, to illustrate the performance of the improvement of the ensemble method compared to its members for multistep-ahead in French Guiana, as in Figure 54.

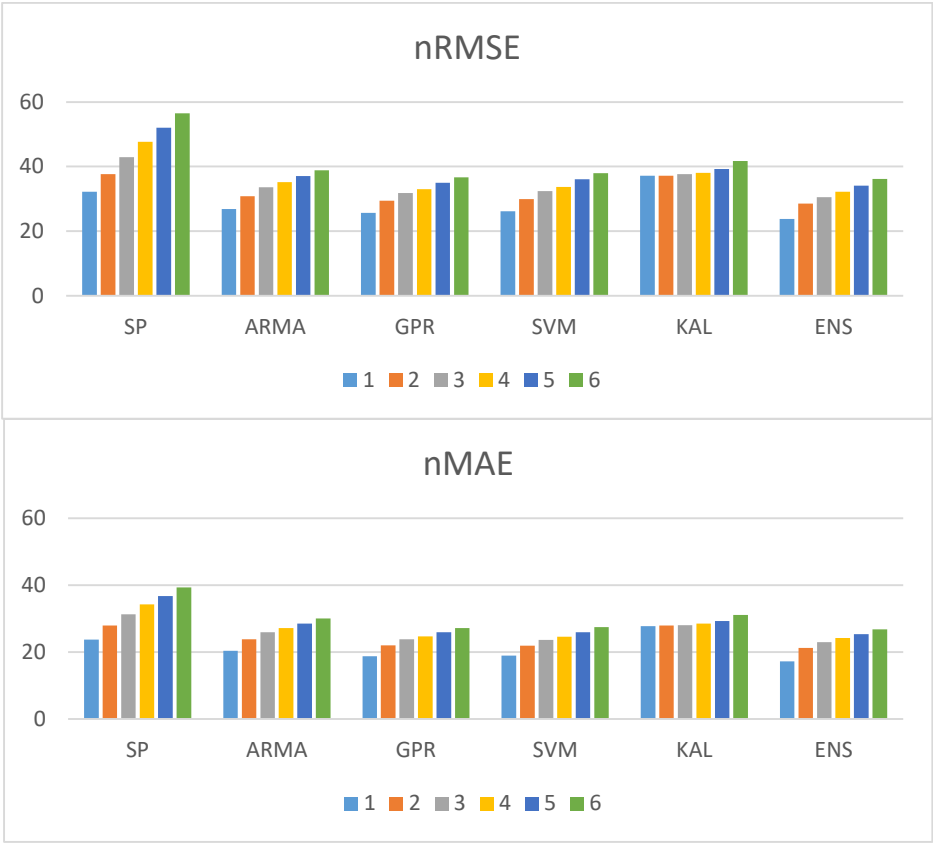


Figure 54: nRMSE, nMAE forecasting errors vs. Horizon for Satellite-derived data combined with each ensemble, Scaled persistence SP, Auto Regression Moving Average ARMA, Gaussian Process Regression GPR, Support Vector Machine SVM, WRF-Kalman KAL, and with the ensemble aggregation ENS. Three months' data from five stations were merged, SL is the training station for computing the weights and was excluded.

The graph shows the accuracy metrics of the ensemble method and its members using satellite-derived data from five meteorological stations of 2016 as input. The nRMSE and nMAE of the ensemble aggregate are lower than the nRMSE and nMAE of the members for all time horizons. For time horizons higher than 4H, the ensemble method has forecast errors close to the best ensemble member GPR. We explain this phenomenon by the higher weight given to GPR by the Ridge regression aggregation method. The computed weights are 0.34, 0.31, -0.12, 0.46 and 0.28, respectively, for ARMA, SVM, SP, GPR and KAL.

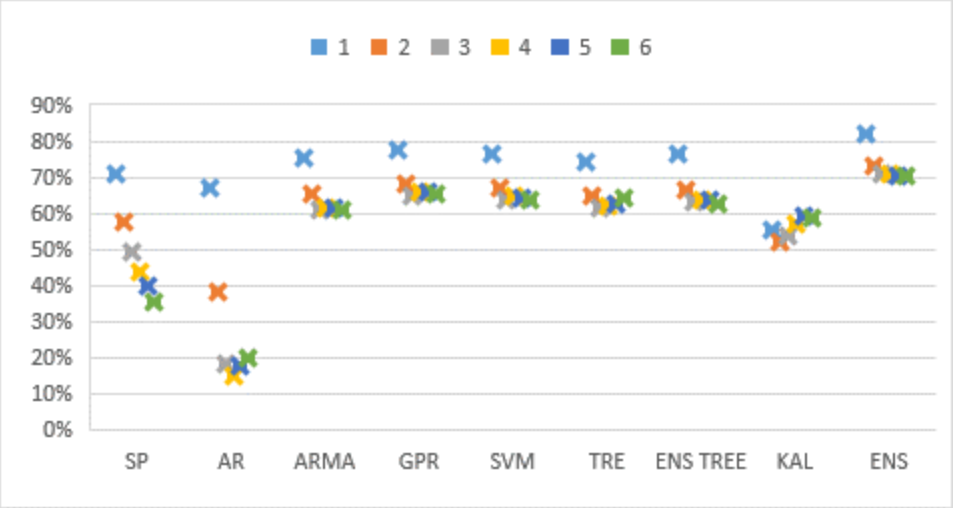


Figure 55: aR^2 according to the horizon, of the ensemble method and its members combined with satellite-derived data from five meteorological stations of 2016 as input.

The aR^2 according to the horizon in Figure 55 shows the ensemble method and its five members, note that the ensemble method has the best aR^2 for all time horizons.

Summary

In conclusion of this section to improve GHI forecasts in the ITZ using aggregation method combined with satellite-derived data from GOES, we note the following:

- The ability of each forecasting model is successfully combined.
- The nRMSE and nMAE of the ensemble aggregate are lower than the best member model.
- For all time horizons, the ensemble aggregate has a lower forecast error than the SP, ARMA, GPR, SVM, and Kalman. Yet, it has close forecast errors than GPR. We have explained this phenomenon by the different weights of each model.
- For all-sky conditions merged, we found that the ensemble aggregate decreases the 1H forecast error of the most accurate single model by 2%.
- The ensemble aggregate decreases the 1H forecast error of almost 15% compared with the persistence reference model.

IV.7. Summary

We provide a brief final summary here as we have provided four summaries relating to each subsection. The main purpose of the current study is to forecast GHI in an ITZ using satellite-derived data from GOES-13. The novelty of our approach is the use of satellite-derived data to forecast GHI with different forecast horizons in French Guiana. For intraday forecast using satellite-derived data, we achieved the following studies:

Firstly, we started to forecast GHI using simple and primary cases, therefore four well-known statistical forecasting models; Persistence (P), Scaled Persistence (SP), Auto Regressive (AR) and Auto Regressive Moving Average (ARMA) models are studied.

Secondly, four machine learning models are used to forecast GHI; Gaussian Process Regression (GPR), Support Vector Machine (SVM), Simple Regression Tree (TRE) and Ensembles of Trees 'Bagging' (ENsTRE).

Thirdly, we used satellite-derived data with a post-processing technique to improve 24-h WRF predictions. Then investigate the effectiveness of this proposal from 1 to 6 horizons with 2.8% and 1.3 % of improvement.

Finally, we tried to go further and improve GHI forecasts multi-hours ahead in French Guiana by using the ensemble of five GHI forecasts. The ensemble includes one hybrid physical statistical model (WRF- Kalman), the two best statistical forecasting models (SP, ARMA), and two best machine learning models (GPR, SVM), we combine these ensemble members using a Ridge regression method.

In this chapter, we have conducted four comparative studies of using satellite-derived data instead of ground measured one in French Guiana:

- Forecast of hourly GHI for hourly horizons of 1 to 6H with four statistical models;
- Forecast of hourly GHI for hourly horizons of 1 to 6H with four Machine learning models;
- Improving WRF forecasts using the Kalman filter with satellite-derived data for hourly horizons of 1 to 6H;
- Finally, using the best five predictor models for horizons of 1 to 6H as members in a hybrid ensemble method to improve hourly GHI forecasts;

As mentioned earlier, satellite data has been increasingly used to predict Global Horizontal Irradiance GHI. Several recent studies have shown an interesting model performance obtained by adding satellite data to the ground measurement. (Jang et al. 2016) used satellite images and support vector machine to predict the solar power. For 1H horizon, the nRMSE improvement from the reference model is 13% compared to 15% in our study. Another study used satellite data is (Aguiar et al. 2016), that combine ground measurements with exogenous inputs provided by satellite and NWP data in order to improve intra-day solar forecasting. With nRMSE improvement of 1% to 2%, for 1H to 6H ahead respectively. (Yagli 2020) used five ML models to generate 1-h-ahead forecasts of GHI based on ground measurements and satellite-derived data. The

forecasts are combined using the trimmed average technique to form ensemble forecasts. The improvement in nRMSE using the ensemble vary between 0.1 and 1%, depending on the station.

All results in our study showed that satellite-based model performance was close to models built with ground data for an intraday forecast.

Chapter .V. Conclusion and outlook

V.1. Conclusion

Solar energy is a renewable and clean source of energy increasingly used worldwide. During the day, photovoltaic solar panels convert solar energy into electrical one. Solar energy has a behavior that can cause strong fluctuations in the electrical power injected into the electrical network. The limitation of producing electricity from solar energy can be overcome, by storing electricity during the periods of high production to recover it during periods of low production. Storage would stabilize the photovoltaic power plants production. It requires a large amount of storage by a power reserve to smooth the fluctuations generated, which is expensive. Another solution is to anticipate the availability of production sources by forecasting the production of intermittent solar energy and therefore facilitate network management. This solution enables easy network management. Additionally, accurate forecasting of solar irradiation can improve the efficiency of the renewable energy conversion chain, reduce the risk caused by extreme weather conditions, and can optimize power grid management. In other words, accurate forecasting of the resource allows power grid managers to adapt their actions and maintain the quality and safety of the power grid at all times. The interest of our study is to test the robustness of ten models using satellite-derived data in French Guiana. The interest of this research is to study the effectiveness of using satellite data rather than measured data. To investigate this study, we tested the robustness of ten models using data from satellites in French Guiana.

Models from three main categories were chosen and compared:

- Statistical models:
    - ✓ Models based on the persistence of weather conditions: Persistence (P), the Scaled Persistence (SP), and regression: Auto Regression (AR) and Auto Regression Moving Average (ARMA).
    - ✓ Models based on machine learning: Gaussian Process Regression (GPR), Support Vector Machine (SVM), Regression Trees (TRE), Ensembles of Trees (ENsTRE). These models require a learning phase to determine various parameters of the model. Thus, the time series is partitioned into two groups, one for the learning phase (80%) and one for the test phase (20%). Cross-validation was used by the k-fold method to prevent the learning phase from being subject to seasonal trends and to strengthen the validation of our models.
- The numerical model (weather researching and forecasting model WRF) is based on a mathematical representation of atmosphere behavior. Therefore, it gives a more detailed description of the atmospheric processes. kalman filter is used to post processed WRF forecasts.
- Hybrid model with estimated satellite data are used to improve solar predictions in tropical environments. Aggregation method decreases the forecast error and produces a more realistic spatial pattern of predicted irradiation using a

combination of five methods to improve the solar irradiation forecast in the French Guiana tropical climate.

The implementation of these models requires the availability of satellite-derived data, in addition to the availability of ground measurement to compare models' performance. Hourly data from six meteorological stations in French Guiana over 6 years (2011 to 2016) are used. Before using ground measured and satellite-derived data to feed the forecast models. This data must go through a preprocessing protocol. which consists of collecting the data, controlling their quality by detecting outliers and missing data periods. After these preprocessing protocols, the data are ready to be used as inputs to models. A comparison of these models was conducted on the different datasets. Model performance was estimated by several error metrics widely used in solar forecasting. The main findings are as follows:

- From the analysis of statistics and machine learning models, it appears that the 1H forecast quality of these models using ground data is better than satellite one by 0.1 to 2% of nRMSE, then for 2H to 6H time horizon data satellite-derived data are more efficient by 0.1 to 1% of nRMSE.
- Although the statistical models' performances with satellite-derived data are close, the ARMA model is the most suitable, then the SP. The SP model presents interesting performances with satellite-derived data. Moreover, it has the advantage of being easy to implement.
- It is difficult to compare the machine learning models as predictors because of their similar performances. Machine learning models are suited for short-term forecasting, from a few minutes to a few hours.
- Even though, all of the machine learning models seem interesting to predict the GHI, we choose GPR and SVM as the elected models to be members in the hybrid ensemble model.
- Using satellite data in the post-processing of numerical model gives a result very close to the case with ground data. The advantage of using satellite data is to maintain the resolution spatial of the numerical model.
- The Kalman filter improves WRF forecasts for horizons from 1 to 6H.
- Hybridizing satellite-based models to get the best out of each gives an interesting improvement to try to keep the forecast error as low as possible.
- For all models the model's performance difference between ground and satellite data is insignificant.
- Even if the use of satellite-derived data gives results that are very close to the case of using ground data. The advantage of using satellite data is its broader coverage than ground-based irradiation measurements.

To conclude, forecasts generated using satellite-derived irradiance estimations have comparable quality compared to those generated using ground-based data. Our study indicates the use of satellite-derived data as a viable tool for forecasting GHI in French

Guiana a region in an ITZ. The information they can provide will establish reliable planning for producing electricity from solar energy.

V.2. Outlook

The conclusion of the thesis work revealed interesting prospects for improvement. The use and development of several models, each combined with satellite-derived data, conclude that satellite estimation is a reliable source to intraday GHI forecast. The use of satellite-derived data estimation to predict GHI, is a topic that continues to be the subject of many recent articles. At the end of this work, we will present a set of improvements for further studies:

- ✓ Extend the temporal depth of the solar irradiation estimates by considering data from GOES-8 and GOES-12;
- ✓ Test satellite-derived data to other methods sites, climates, or even to other similar climate regions in the world, at least, accounting for the GOES satellite coverage;
- ✓ Results could be improved by applying more elaborated pre-processing combined with our method to substantially improve forecasting performance;
- ✓ More effective selection of satellite pixels would give us satellite-derived data more accurate and essential information about the solar radiation behavior;
- ✓ The methodology presented here may be extended to longer time horizons, by adding NWP data with model plus complex;
- ✓ Using clustering to gain more information about the data to improve the prediction accuracy.
- ✓ Using more features to improve post-processing techniques;
- ✓ The contribution of other features and the use of a deeper neural network;
- ✓ Using a sequential aggregation method with weights varying with time and by model and including other forecasting methods that perform better under highly variable sky conditions.
- ✓ Study the influence of our method on the production prediction;

References

ADEME. 2017. 'Analyse Des Impacts Prévisibles Des Objectifs Énergétiques En ZNI Sur l'activité et l'emploi'.

Aguiar, L. Mazorra, B. Pereira, P. Lauret, F. Díaz, and M. David. 2016. 'Combining Solar Irradiance Measurements, Satellite-Derived Data and a Numerical Weather Prediction Model to Improve Intra-Day Solar Forecasting'. *Renewable Energy* 97 (November): 599–610. https://doi.org/10.1016/j.renene.2016.06.018.

Ahmed, R., V. Sreeram, Y. Mishra, and M.D. Arif. 2020. 'A Review and Evaluation of the State-of-the-Art in PV Solar Power Forecasting: Techniques and Optimization'. *Renewable and Sustainable Energy Reviews* 124 (May): 109792. https://doi.org/10.1016/j.rser.2020.109792.

Albarelo, Tommy. 2017. 'Estimation de l'Irradiation Solaire sur le Plateau des Guyanes: apport de la Télédétection Satellite', 108.

albarelo, tommy. 2017. 'ESTIMATION DE L'IRRADIATION SOLAIRE SUR LE PLATEAU DES GUYANES : APPORT DE LA TELEDETECTION SATELLITE'.

Albarelo, Tommy, Isabelle Marie-Joseph, Antoine Primerose, Frédérique Seyler, Lucien Wald, and Laurent Linguet. 2015a. 'Optimizing the Heliosat-II Method for Surface Solar Irradiation Estimation with GOES Images'. *Canadian Journal of Remote Sensing* 41 (2): 86–100. https://doi.org/10.1080/07038992.2015.1040876.

Albarelo 2015b. 'Optimizing the Heliosat-II Method for Surface Solar Irradiation Estimation with GOES Images'. *Canadian Journal of Remote Sensing* 41 (2): 86–100. https://doi.org/10.1080/07038992.2015.1040876.

'Albioma_DDR'. 2018.

Alonso-Suárez, R., M. David, V. Branco, and P. Lauret. 2020. 'Intra-Day Solar Probabilistic Forecasts Including Local Short-Term Variability and Satellite Information'. *Renewable Energy* 158 (October): 554–73. https://doi.org/10.1016/j.renene.2020.05.046.

André, Maïna, Richard Perez, Ted Soubdhan, James Schlemmer, Rudy Calif, and Stéphanie Monjoly. 2019. 'Preliminary Assessment of Two Spatio-Temporal Forecasting Technics for Hourly Satellite-Derived Irradiance in a Complex Meteorological Context'. *Solar Energy* 177 (January): 703–12. https://doi.org/10.1016/j.solener.2018.11.010.

Antonanzas-Torres, F., R. Urraca, J. Polo, O. Perpiñán-Lamigueiro, and R. Escobar. 2019. 'Clear Sky Solar Irradiance Models: A Review of Seventy Models'. *Renewable and Sustainable Energy Reviews* 107 (June): 374–87. https://doi.org/10.1016/j.rser.2019.02.032.

Aryaputera, Aloysius W. 2015. 'Very Short-Term Irradiance Forecasting at Unobserved Locations Using Spatio-Temporal Kriging'. *Solar Energy*, 13.

Bauer, Peter, Alan Thorpe, and Gilbert Brunet. 2015. 'The Quiet Revolution of Numerical Weather Prediction'. *Nature* 525 (7567): 47–55. https://doi.org/10.1038/nature14956.

Benamrou, Badr, Mustapha Ouardouz, Imane Allaouzi, and Mohamed Ben Ahmed. 2020. 'A Proposed Model to Forecast Hourly Global Solar Irradiation Based on Satellite Derived Data, Deep Learning and Machine Learning Approaches'. *Journal of Ecological Engineering* 21 (4): 26–38. https://doi.org/10.12911/22998993/119795.

Bourbonnais, Régis, and Michel Terraza. 2010. *Analyse Des Séries Temporelles : Applications à l'économie et à La Gestion*. Vol. 1.

Bovolo, C. Isabella, Ryan Pereira, Geoff Parkin, Chris Kilsby, and Thomas Wagner. 2012. 'Fine-Scale Regional Climate Patterns in the Guianas, Tropical South America, Based on Observations and Reanalysis Data'. *International Journal of Climatology* 32 (11): 1665–89. https://doi.org/10.1002/joc.2387.

Box, George E P. 1970. 'Time Series Analysis', 709.

Breiman et. al. 1984. *Classification and Regression Trees.Pdf*.

Bright, Jamie M. 2019. 'Solcast: Validation of a Satellite-Derived Solar Irradiance Dataset'. *Solar Energy* 189 (September): 435–49. https://doi.org/10.1016/j.solener.2019.07.086.

Cano, D., J.M. Monget, M. Albuisson, H. Guillard, N. Regas, and L. Wald. 1986. 'A Method for the Determination of the Global Solar Radiation from Meteorological Satellite Data'. *Solar Energy* 37 (1): 31–39. https://doi.org/10.1016/0038-092X(86)90104-0.

Cesa-Bianchi, Nicolo, and Gabor Lugosi. 2006. *Prediction, Learning, and Games*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511546921.

Cornejo-Bueno, L., C. Casanova-Mateo, J. Sanz-Justo, and S. Salcedo-Sanz. 2019. 'Machine Learning Regressors for Solar Radiation Estimation from Satellite Data'. *Solar Energy* 183 (May): 768–75. https://doi.org/10.1016/j.solener.2019.03.079.

De'ath, Glenn. 2007. 'BOOSTED TREES FOR ECOLOGICAL MODELING AND PREDICTION'. *Ecology* 88 (1): 243–51. https://doi.org/10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2.

Deo, Ravinesh C. 2017. 'Forecasting Long-Term Global Solar Radiation with an ANN Algorithm Coupled with Satellite-Derived (MODIS) Land Surface Temperature (LST) for Regional Locations in Queensland'. *Renewable and Sustainable Energy Reviews*, 21.

Deo 2019. 'Universally Deployable Extreme Learning Machines Integrated with Remotely Sensed MODIS Satellite Predictors over Australia to Forecast Global Solar Radiation: A New Approach'. *Renewable and Sustainable Energy Reviews*, 27.

Diagne, Maimouna, Mathieu David, John Boland, Nicolas Schmutz, and Philippe Lauret. 2014. 'Post-Processing of Solar Irradiance Forecasts from WRF Model at Reunion Island'. *Solar Energy* 105 (July): 99–108. https://doi.org/10.1016/j.solener.2014.03.016.

Diagne, Maimouna, Mathieu David, Philippe Lauret, John Boland, and Nicolas Schmutz. 2013a. 'Review of Solar Irradiance Forecasting Methods and a Proposition for Small-Scale Insular Grids'. *Renewable and Sustainable Energy Reviews* 27 (November): 65–76. https://doi.org/10.1016/j.rser.2013.06.042.

Diagne 2013b. 'Review of Solar Irradiance Forecasting Methods and a Proposition for Small-Scale Insular Grids'. *Renewable and Sustainable Energy Reviews* 27 (November): 65–76. https://doi.org/10.1016/j.rser.2013.06.042.

Diallo, Mouhamet. 2018. 'Solar irradiance forecast and assesment in the intertropical zone', 161.

Diallo 2019. 'Estimation et Prévision de l'éclairement Solaire Au Sol En Zone Intertropicale'.

Dong, Zibo, Dazhi Yang, Thomas Reindl, and Wilfred M. Walsh. 2014. 'Satellite Image Analysis and a Hybrid ESSS/ANN Model to Forecast Solar Irradiance in the Tropics'. *Energy Conversion and Management* 79 (March): 66–73. https://doi.org/10.1016/j.enconman.2013.11.043.

Du, Jun, Judith Berner, Roberto Buizza, Martin Charron, Pieter Leopold Houtekamer, Dingchen Hou, Isidora Jankov, et al. 2018. 'Ensemble Methods for Meteorological Predictions.' *U.S. Department of Commerce, National Oceanic and Atmospheric Administration (NOAA), National Weather Service, National Centers for Environmental Prediction (NCEP)*. https://doi.org/10.7289/v5/on-ncep-493.

Duffie, John A, and William A Beckman. 2013. 'Solar Engineering of Thermal Processes', 928.

Elliston, Ben, and Iain MacGill. 2010. 'The Potential Role of Forecasting for Integrating Solar Generation into the Australian National Electricity Market', 12.

Espinar, Bella, Philippe Blanc, Lucien Wald, Carsten Hoyer-Klick, Marion Schroedter Homscheidt, and Thomas Wanderer. 2012. 'On Quality Control Procedures for Solar Radiation and Meteorological Measures, from Subhourly to Montly Average Time Periods', 2.

Fan, Junliang, Lifeng Wu, Fucang Zhang, Huanjie Cai, Wenzhi Zeng, Xiukang Wang, and Haiyang Zou. 2019. 'Empirical and Machine Learning Models for Predicting Daily Global Solar Radiation from Sunshine Duration: A Review and Case Study in China'. *Renewable and Sustainable Energy Reviews* 100 (February): 186–212. https://doi.org/10.1016/j.rser.2018.10.018.

Fillol, Erwann, Tommy Albarelo, Antoine Primerose, Lucien Wald, and Laurent Linguet. 2017. 'Spatiotemporal Indicators of Solar Energy Potential in the Guiana Shield Using GOES Images'. *Renewable Energy* 111 (October): 11–25. https://doi.org/10.1016/j.renene.2017.03.081.

Fouilloy. 2019. 'Comparaison de Méthodes d'apprentissage Automatique de Prévision de La Ressource Solaire Pour Une Application à Une Gestion Optimisée Des Réseaux Intelligents;' CORSE-PASCAL PAOLI.

Galanis, George, P Gouka, P Katsafados, G Kallos, and I Pytharoulis. 2006. 'Applications of Kalman Filters Based on Non−linear Functions to Numerical Weather Prediction.Pdf'.

Geiger, M., L. Diabaté, L. Ménard, and L. Wald. 2002. 'A Web Service for Controlling the Quality of Measurements of Global Solar Irradiation'. *Solar Energy* 73 (6): 475–80. https://doi.org/10.1016/S0038-092X(02)00121-4.

G.Schwarz. 1978. 'Estimating the Dimension of a Model,.Pdf' 6 (2): 171–76.

Guijo-Rubio, D. 2020. 'Evolutionary Artificial Neural Networks for Accurate Solar Radiation Prediction', 11.

Gupta, Anuj, Kapil Gupta, and Sumit Saroha. 2021. 'A Review and Evaluation of Solar Forecasting Technologies'. *Materials Today: Proceedings*, May, S2214785321034167. https://doi.org/10.1016/j.matpr.2021.04.491.

Hastie, T., Tibshirani, R. 1986. 'Generalized Additive Models.Pdf'.

He, Chuan, Jiandong Liu, Fang Xu, Teng Zhang, Shang Chen, Zhe Sun, Wenhui Zheng, et al. 2020. 'Combination of Meteorological Factors with Machine Learning Methods'. *Energy Conversion and Management*, 15.

Heinemann, Detlev. 2006. 'Forecasting of Solar Radiation', 10.

Huang, Guanghui, Zhanqing Li, Xin Li, Shunlin Liang, Kun Yang, Dongdong Wang, and Yi Zhang. 2019. 'Estimating Surface Solar Irradiance from Satellites: Past, Present, and Future Perspectives'. *Remote Sensing of Environment* 233 (November): 111371. https://doi.org/10.1016/j.rse.2019.111371.

Huva, Robert, Hadrien Verbois, and Wilfred Walsh. 2020. 'Comparisons of Next-Day Solar Forecasting for Singapore Using 3DVAR and 4DVAR Data Assimilation Approaches with the WRF Model'. *Renewable Energy* 147 (March): 663–71. https://doi.org/10.1016/j.renene.2019.09.011.

Iqpal, Muhammad. 1983. *An Introduction To Solar Radiation - 1st Edition.*

Jang, Han Seung, Kuk Yeol Bae, Hong-Shik Park, and Dan Keun Sung. 2016. 'Solar Power Prediction Based on Satellite Images and Support Vector Machine'. *IEEE Transactions on Sustainable Energy* 7 (3): 1255–63. https://doi.org/10.1109/TSTE.2016.2535466.

Jarvis, A, E Guevara, H.I Reuter, and A.D Nelson. 2008. 'Hole-Filled SRTM for the Globe V4'.

Jiang, Hou, Ning Lu, Guanghui Huang, Ling Yao, Jun Qin, and Hengzi Liu. 2020. 'Spatial Scale Effects on Retrieval Accuracy of Surface Solar Radiation Using Satellite Data'. *Applied Energy* 270 (July): 115178. https://doi.org/10.1016/j.apenergy.2020.115178.

Jiang, Hou, Ning Lu, Jun Qin, Wenjun Tang, and Ling Yao. 2019. 'A Deep Learning Algorithm to Estimate Hourly Global Solar Radiation from Geostationary Satellite Data'. *Renewable and Sustainable Energy Reviews* 114 (October): 109327. https://doi.org/10.1016/j.rser.2019.109327.

Jiang, Ping, and Jiejie Chen. 2016. 'Displacement Prediction of Landslide Based on Generalized Regression Neural Networks with K-Fold Cross-Validation'. *Neurocomputing* 198 (July): 40–47. https://doi.org/10.1016/j.neucom.2015.08.118.

Kallio-Myers, Viivi, Aku Riihelä, Panu Lahtinen, and Anders Lindfors. 2020. 'Global Horizontal Irradiance Forecast for Finland Based on Geostationary Weather Satellite Data'. *Solar Energy* 198 (March): 68–80. https://doi.org/10.1016/j.solener.2020.01.008.

Kurtz, Benjamin, Felipe Mejia, and Jan Kleissl. 2017. 'A Virtual Sky Imager Testbed for Solar Energy Forecasting'. *Solar Energy* 158 (December): 753–59. https://doi.org/10.1016/j.solener.2017.10.036.

Laguarda, A., G. Giacosa, R. Alonso-Suárez, and G. Abal. 2020. 'Performance of the Site-Adapted CAMS Database and Locally Adjusted Cloud Index Models for Estimating Global Solar Horizontal Irradiation over the Pampa Húmeda'. *Solar Energy* 199 (March): 295–307. https://doi.org/10.1016/j.solener.2020.02.005.

Lara-Fanego, V., J.A. Ruiz-Arias, D. Pozo-Vázquez, F.J. Santos-Alamillos, and J. Tovar-Pescador. 2012. 'Evaluation of the WRF Model Solar Irradiance Forecasts in Andalusia (Southern Spain)'. *Solar Energy* 86 (8): 2200–2217. https://doi.org/10.1016/j.solener.2011.02.014.

Lauret, Philippe, Mathieu David, and Didier Calogine. 2012. 'Nonlinear Models for Short-Time Load Forecasting'. *Energy Procedia* 14: 1404–9. https://doi.org/10.1016/j.egypro.2011.12.1109.

Lauret, Philippe, Cyril Voyant, Ted Soubdhan, Mathieu David, and Philippe Poggi. 2015. 'A Benchmarking of Machine Learning Techniques for Solar Radiation Forecasting in an Insular Context'. *Solar Energy* 112 (February): 446–57. https://doi.org/10.1016/j.solener.2014.12.014.

Lesperance, Wilna, Jules Sadefo Kamdem, Laurent Linguet, and Tommy Albarelo. 2018. 'Renewable Energy in French Guiana: Prospects towards a Sustainable Development Scenario'. In *2018 2nd International Conference on Smart Grid and Smart Cities (ICSGSC)*, 133–36. Kuala Lumpur: IEEE. https://doi.org/10.1109/ICSGSC.2018.8541267.

Lorenz, E. 2011. 'Local and Regional Photovoltaic Power Prediction for Large Scale Grid Integration: Assessment of a New Algorithm for Snow Detection'. *Progress in Photovoltaics* 20: 760–69.

Lorenz, Elke, Annette Hammer, and Detlev Heinemann. 2004. 'Short Term Forecasting of Solar Radiation Based on Satellite Data', 8.

Mallet, Vivien, Gilles Stoltz, and Boris Mauricette. 2009. 'Ozone Ensemble Forecast with Machine Learning Algorithms'. *Journal of Geophysical Research* 114 (D5). https://doi.org/10.1029/2008JD009978.

Marchesoni-Acland, Franco, and Rodrigo Alonso-Suárez. 2020. 'Intra-Day Solar Irradiation Forecast Using RLS Filters and Satellite Images'. *Renewable Energy* 161 (December): 1140–54. https://doi.org/10.1016/j.renene.2020.07.101.

Marquez, Ricardo, and Carlos F.M. Coimbra. 2011. 'Forecasting of Global and Direct Solar Irradiance Using Stochastic Learning Methods, Ground Experiments and the NWS Database'. *Solar Energy* 85 (5): 746–56. https://doi.org/10.1016/j.solener.2011.01.007.

Marquez, Ricardo, Hugo T.C. Pedro, and Carlos F.M. Coimbra. 2013. 'Hybrid Solar Forecasting Method Uses Satellite Imaging and Ground Telemetry as Inputs to ANNs'. *Solar Energy* 92 (June): 176–88. https://doi.org/10.1016/j.solener.2013.02.023.

Mathiesen, Patrick, and Jan Kleissl. 2011. 'Evaluation of Numerical Weather Prediction for Intra-Day Solar Forecasting in the Continental United States'. *Solar Energy* 85 (5): 967–77. https://doi.org/10.1016/j.solener.2011.02.013.

Mazorra Aguiar, L., B. Pereira, M. David, F. Díaz, and P. Lauret. 2015. 'Use of Satellite Data to Improve Solar Radiation Forecasting with Bayesian Artificial Neural Networks'. *Solar Energy* 122 (December): 1309–24. https://doi.org/10.1016/j.solener.2015.10.041.

Mazorra Aguiar, L., J. Polo, J.M. Vindel, and A. Oliver. 2019. 'Analysis of Satellite Derived Solar Irradiance in Islands with Site Adaptation Techniques for Improving the Uncertainty'. *Renewable Energy* 135 (May): 98–107. https://doi.org/10.1016/j.renene.2018.11.099.

Messenger, Roger A., and Jerry Ventre. 2004. *Photovoltaic Systems Engineering*. 2nd ed. Boca Raton: CRC Press.

Mihalakakou, G., M. Santamouris, and D. N. Asimakopoulos. 2000. 'The Total Solar Radiation Time Series Simulation in Athens, Using Neural Networks'. *Theoretical and Applied Climatology* 66 (3–4): 185–97. https://doi.org/10.1007/s007040070024.

Miller, Steven D., Matthew A. Rogers, John M. Haynes, Manajit Sengupta, and Andrew K. Heidinger. 2018. 'Short-Term Solar Irradiance Forecasting via Satellite/Model Coupling'. *Solar Energy* 168 (July): 102–17. https://doi.org/10.1016/j.solener.2017.11.049.

Mouhamet, Diallo. 2018. 'Conception d'un dispositif de prédiction et de stockage intelligent de l'énergie photovoltaïque par traitement d'image satellite', 15.

Muneer, T, and F Fairooz. 2002. 'Quality Control of Solar Radiation and Sunshine Measurements – Lessons Learnt from Processing Worldwide Databases'. *Building Services Engineering Research and Technology* 23 (3): 151–66. https://doi.org/10.1191/0143624402bt038oa.

Noia, M., C. F. Ratto, and R. Festa. 1993a. 'Solar Irradiance Estimation from Geostationary Satellite Data: I. Statistical Models'. *Solar Energy* 51 (6): 449–56. https://doi.org/10.1016/0038-092X(93)90130-G.

Noia 1993b. 'Solar Irradiance Estimation from Geostationary Satellite Data: II. Physical Models'. *Solar Energy* 51 (6): 457–65. https://doi.org/10.1016/0038-092X(93)90131-7.

Olomiyesan, Boluwaji M., and Onyedi D. Oyedum. 2016. 'Comparative Study of Ground Measured, Satellite-Derived, and Estimated Global Solar Radiation Data in Nigeria'. *Journal of Solar Energy* 2016 (June): 1–7. https://doi.org/10.1155/2016/8197389.

Pelland, Sophie, George Galanis, and George Kallos. 2013. 'Solar and Photovoltaic Forecasting through Post-Processing of the Global Environmental Multiscale Numerical Weather Prediction Model: Solar and Photovoltaic Forecasting'. *Progress in Photovoltaics: Research and Applications* 21 (3): 284–96. https://doi.org/10.1002/pip.1180.

Perez, Richard, Sergey Kivalov, James Schlemmer, Karl Hemker, David Renné, and Thomas E. Hoff. 2010. 'Validation of Short and Medium Term Operational Solar Radiation Forecasts in the US'. *Solar Energy* 84 (12): 2161–72. https://doi.org/10.1016/j.solener.2010.08.014.

Perez, Richard, Robert Seals, Ronald Stewart, Antoine Zelenka, and Vicente Estrada-Cajigal. 1994. 'Using Satellite-Derived Insolation Data for the Site/Time Specific Simulation of Solar Energy Systems'. *Solar Energy* 53 (6): 491–95. https://doi.org/10.1016/0038-092X(94)90128-O.

Prado, Francisco. 2020. 'Forecasting Based on an Ensemble Autoregressive Moving Average - Adaptive Neuro - Fuzzy Inference System - Neural Network - Genetic Algorithm Framework', 13.

Rasmussen, Carl Edward. 2006. 'EVALUATION OF GAUSSIAN PROCESSES AND OTHER METHODS FOR NON-LINEAR REGRESSION', 138.

Remund, J, R Perez, and E Lorenz. 2009. 'COMPARISON OF SOLAR RADIATION FORECASTS FOR THE USA', 3.

Ren, Ye, P.N. Suganthan, and N. Srikanth. 2015. 'Ensemble Methods for Wind and Solar Power Forecasting—A State-of-the-Art Review'. *Renewable and Sustainable Energy Reviews* 50 (October): 82–91. https://doi.org/10.1016/j.rser.2015.04.081.

Rigollier, Christelle, Olivier Bauer, and Lucien Wald. 2000. 'ON THE CLEAR SKY MODEL OF THE ESRA — EUROPEAN SOLAR RADIATION ATLAS — WITH RESPECT TO THE HELIOSAT METHOD', 16.

Rim, Chang-Bok, Kum-Chol Om, Guoyu Ren, Su-Song Kim, Hyok-Chol Kim, and O. Kang-Chol. 2018. 'Establishment of a Wildfire Forecasting System Based on Coupled Weather–Wildfire Modeling'. *Applied Geography* 90 (January): 224–28. https://doi.org/10.1016/j.apgeog.2017.12.011.

Rodríguez-Benítez, Francisco J., Clara Arbizu-Barrena, Javier Huertas-Tato, Ricardo Aler-Mur, Inés Galván-León, and David Pozo-Vázquez. 2020. 'A Short-Term Solar Radiation Forecasting System for the Iberian Peninsula. Part 1: Models Description and Performance Assessment'. *Solar Energy* 195 (January): 396–412. https://doi.org/10.1016/j.solener.2019.11.028.

Rohani, Abbas, Morteza Taki, and Masoumeh Abdollahpour. 2018. 'A Novel Soft Computing Model (Gaussian Process Regression with K-Fold Cross Validation) for Daily and Monthly Solar Radiation Forecasting (Part: I)'. *Renewable Energy* 115 (January): 411–22. https://doi.org/10.1016/j.renene.2017.08.061.

Rossi, F., A. Lendasse, D. François, V. Wertz, and M. Verleysen. 2006. 'Mutual Information for the Selection of Relevant Variables in Spectrometric Nonlinear Modelling'. *Chemometrics and Intelligent Laboratory Systems* 80 (2): 215–26. https://doi.org/10.1016/j.chemolab.2005.06.010.

Salazar, Germán, Christian Gueymard, Janis Bezerra Galdino, Olga de Castro Vilela, and Naum Fraidenraich. 2020. 'Solar Irradiance Time Series Derived from High-Quality Measurements, Satellite-Based Models, and Reanalyses at a near-Equatorial Site in Brazil'. *Renewable and Sustainable Energy Reviews* 117 (January): 109478. https://doi.org/10.1016/j.rser.2019.109478.

Salloum, Maha, Mouhamet Diallo, Antoine Primerose, and Laurent Linguet. 2019. 'IMPROVING HOUR AHEAD SOLAR IRRADIANCE FORECAST USING ENSEMBLE METHOD IN FRENCH GUIANA', 5.

Schulz, Eric, Maarten Speekenbrink, and Andreas Krause. 2018. 'A Tutorial on Gaussian Process Regression: Modelling, Exploring, and Exploiting Functions'. *Journal of Mathematical Psychology* 85 (August): 1–16. https://doi.org/10.1016/j.jmp.2018.03.001.

Shakya, Ayush. 2016. 'Implementation of Solar Irradiance Forecasting Using Markov Switching Model and Energy Management System', 126.

Singh, Bismark, and David Pozo. 2019. 'A Guide to Solar Power Forecasting Using ARMA Models'. In *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*, 1–4. Bucharest, Romania: IEEE. https://doi.org/10.1109/ISGTEurope.2019.8905430.

Singh Doorga, Jay Rovisham, Kumar Ram Dhurmea, Soonil Rughooputh, and Ravindra Boojhawon. 2019. 'Forecasting Mesoscale Distribution of Surface Solar Irradiation Using a Proposed Hybrid Approach Combining Satellite Remote Sensing and Time Series Models'. *Renewable and Sustainable Energy Reviews* 104 (April): 69–85. https://doi.org/10.1016/j.rser.2018.12.055.

'Snapshot of Global PV Markets'. 2020, 20.

Sobri, Sobrina, Sam Koohi-Kamali, and Nasrudin Abd. Rahim. 2018. 'Solar Photovoltaic Generation Forecasting Methods: A Review'. *Energy Conversion and Management* 156 (January): 459–97. https://doi.org/10.1016/j.enconman.2017.11.019.

'Solar Power Europe'. 2019.

Solstice, Summer, Winter Solstice, Spring Equinox, and Fall Equinox. 2013. 'Geography Lab Lesson 4 – Solar Angles and Azimuths Terms to Learn : Solar Angle ; Zenith ; Solar Zenith Angle ; Declination DEFINTIONS ':

Soubdhan, Ted, Joseph Ndong, Hanany Ould-Baba, and Minh-Thang Do. 2016. 'A Robust Forecasting Framework Based on the Kalman Filtering Approach with a Twofold Parameter Tuning Procedure: Application to Solar and Photovoltaic Prediction'. *Solar Energy* 131 (June): 246–59. https://doi.org/10.1016/j.solener.2016.02.036.

Stoica, p, and Y Selen. 2004. 'Model_order Selection a Review of Information Criterion Rules.Pdf'.

Thorey, J., V. Mallet, C. Chaussin, L. Descamps, and P. Blanc. 2015. 'Ensemble Forecast of Solar Radiation Using TIGGE Weather Forecasts and HelioClim Database'. *Solar Energy* 120 (October): 232–43. https://doi.org/10.1016/j.solener.2015.06.049.

Thorey, Jean. 2018. 'Ensemble forecasting using sequential aggregation for photovoltaic power applications'.

Trieb, Franz, Christoph Schillings, Marlene O'Sullivan, Thomas Pregger, and Carsten Hoyer-Klick. 2009. 'Global Potential of Concentrating Solar Power', 12.

Troncoso, A., S. Salcedo-Sanz, C. Casanova-Mateo, J.C. Riquelme, and L. Prieto. 2015. 'Local Models-Based Regression Trees for Very Short-Term Wind Speed Prediction'. *Renewable Energy* 81 (September): 589–98. https://doi.org/10.1016/j.renene.2015.03.071.

Tsay, S. 1984. 'Order Selection in Nonstationary Autoregressive Models.Pdf'.

Tso, Geoffrey K.F., and Kelvin K.W. Yau. 2007. 'Predicting Electricity Energy Consumption: A Comparison of Regression Analysis, Decision Tree and Neural Networks'. *Energy* 32 (9): 1761–68. https://doi.org/10.1016/j.energy.2006.11.010.

Vapnik, Vladimir N. 2000. *The Nature of Statistical Learning Theory*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4757-3264-1.

Verbois, Hadrien. 2020. 'Beyond Quadratic Error: Case-Study of a Multiple Criteria Approach to the Performance Assessment of Numerical Forecasts of Solar Irradiance in the Tropics'. *Renewable and Sustainable Energy Reviews*, 23.

Verbois, Hadrien, Robert Huva, Andrivo Rusydi, and Wilfred Walsh. 2018. 'Solar Irradiance Forecasting in the Tropics Using Numerical Weather Prediction and Statistical Learning'. *Solar Energy* 162 (March): 265–77. https://doi.org/10.1016/j.solener.2018.01.007.

Voyant, Cyril. 2014. 'Time Series Modeling and Large Scale Global Solar Radiation Forecasting from Geostationary Satellites Data'. *Solar Energy*, 12.

Voyant, Cyril, Fabrice Motte, Alexis Fouilloy, Gilles Notton, Christophe Paoli, and Marie-Laure Nivet. 2017. 'Kalman Filtering and Classical Time Series Tools for Global Radiation Prediction', 4.

Voyant, Cyril, Gilles Notton, Soteris Kalogirou, Marie-Laure Nivet, Christophe Paoli, Fabrice Motte, and Alexis Fouilloy. 2017. 'Machine Learning Methods for Solar Radiation Forecasting: A Review'. *Renewable Energy* 105 (May): 569–82. https://doi.org/10.1016/j.renene.2016.12.095.

Voyant, Cyril, Christophe Paoli, Marc Muselli, and Marie-Laure Nivet. 2013. 'Multi-Horizon Solar Radiation Forecasting for Mediterranean Locations Using Time Series Models'. *Renewable and Sustainable Energy Reviews* 28 (December): 44–52. https://doi.org/10.1016/j.rser.2013.07.058.

Voyant, Cyril, Ted Soubdhan, Philippe Lauret, Mathieu David, and Marc Muselli. 2015. 'Statistical Parameters as a Means to a Priori Assess the Accuracy of Solar Forecasting Models'. *Energy* 90 (October): 671–79. https://doi.org/10.1016/j.energy.2015.07.089.

World Meteorological Organization. 2008. *Guide to Meteorological Instruments and Methods of Observation*. Geneva: World Meteorological Organization.

Yagli, Gokhan Mert. 2019. 'Automatic Hourly Solar Forecasting Using Machine Learning Models'. *Renewable and Sustainable Energy Reviews*, 12.

Yagli 2020. 'Can We Justify Producing Univariate Machine-Learning Forecasts with Satellite-Derived Solar Irradiance?' *Applied Energy*, 15.

Yeom, Jong-Min, Ravinesh C. Deo, Jan F. Adamwoski, Taebyeong Chae, Dong-Su Kim, Kyung-Soo Han, and Do-Yong Kim. 2020. 'Exploring Solar and Wind Energy Resources in North Korea with COMS MI Geostationary Satellite Data Coupled with Numerical Weather Prediction Reanalysis Variables'. *Renewable and Sustainable Energy Reviews* 119 (March): 109570. https://doi.org/10.1016/j.rser.2019.109570.

Yokohata, Tokuta, James D. Annan, Matthew Collins, Charles S. Jackson, Hideo Shiogama, Masahiro Watanabe, Seita Emori, et al. 2013. 'Reliability and Importance of Structural Diversity of Climate Model Ensembles'. *Climate Dynamics* 41 (9–10): 2745–63. https://doi.org/10.1007/s00382-013-1733-9.

Zambrano, Andres Felipe, and Luis Felipe Giraldo. 2020. 'Solar Irradiance Forecasting Models without On-Site Training Measurements'. *Renewable Energy* 152 (June): 557–66. https://doi.org/10.1016/j.renene.2020.01.092.

ANNEX

Annex A

Forecasting errors for ground measured and satellite-derived data with four machine learning models according to time horizon and station of 2016.

Station	Model	h+1		h+2		h+3		h+4		h+5		h+6	
		GND	SAT	GND	SAT	GND	SAT	GND	SAT	GND	SAT	GND	SAT
nRMSE(%)													
RO	GPR	26,7	26,7	31,5	31	33,6	33,2	36	35,4	37,3	36,7	39	38,3
	SVM	27,2	25,4	32,1	31,3	34,4	34	37,1	36,5	38,4	38,1	40,6	40,2
	ENS_TRE	27,4	26,6	32,3	31,4	34,7	33,8	36,7	36,4	38	37,5	40,8	40
	TREES	27,9	28,4	32,7	32,4	35,2	35,1	37,7	37,2	38,5	38,1	39,4	39,4
SL	GPR	26,7	27,4	31,6	31,9	33,7	33,8	36,2	36,5	38,6	38,9	41,2	41,4
	SVM	26,6	26,6	31,4	31,9	33,8	33,8	35,9	36,2	38,1	38,8	40,1	41
	ENS_TRE	27,3	27,4	32,3	32,1	34,2	34,2	37	36,8	40	39,2	42,3	41,9
	TREES	28,2	28,6	33,4	33,1	35	35,2	37,1	37,1	40,1	40,5	41,4	41,8
MP	GPR	24,7	23,8	27,3	27	28,6	28,6	30,2	30,1	32,1	32,1	34,2	34,1
	SVM	24,4	23,8	27	26,9	28,3	28,5	29,6	29,8	31,7	32	34	34,3
	ENS_TRE	25,1	24,2	27,6	27,3	29,2	28,9	31,1	30,5	33	32,8	34,9	34,5
	TREES	26,2	25,4	28,8	28,2	29,6	29,8	31,2	31,2	33,3	33,8	35,3	35,1
SG	GPR	26,8	27,5	31,1	31,2	33,6	33,7	35,7	36	38,1	38,5	40,4	41
	SVM	26,9	27,6	31,3	31,7	33,7	34,6	36,2	36,6	38,6	38,9	41	42,1
	ENS_TRE	27,6	27,5	32	31,5	34,1	34,5	36,9	36,3	39,5	38,5	41,9	42,1
	TREES	28,4	28,7	32,6	32,6	35,4	35,5	37,1	37,8	39,5	40,2	41,6	42,4
IR	GPR	22,8	23,8	27,8	27,7	29,9	29,6	31,5	31,3	33,9	33,5	36,8	36,5
	SVM	23	23,8	28,1	27,8	30,5	29,7	32	31,3	34,9	34,6	38	38,3
	ENS_TRE	23,1	23,5	27,7	27,7	30,1	29,4	31,8	31,8	34,3	33,9	37,7	37,7
	TREES	23,9	25,3	28,7	28,8	31	30,7	33	32,7	35,4	34,1	38,3	38
KR	GPR	23,7	26,9	28	29,4	29,5	31	30,1	31,4	31,3	32,3	32,6	33,3
	SVM	23,9	26,5	27,8	28,9	29,5	30,2	29,9	30,8	31,8	33	33,2	33,4
	ENS_TRE	25,1	26,7	29,1	30,1	30,5	32,2	31,7	33	33,2	34,5	34,6	35,4
	TREES	25,6	27,7	29,7	30	31,5	32,3	31,9	33,3	32,6	34	34,4	35,1
nMAE(%)													
RO	GPR	19,3	19,7	23,4	23,2	25,2	24,9	26,8	26,5	27,5	27,3	28,6	28,2
	SVM	19,2	18,9	23,3	22,8	24,9	24,5	26,6	26,1	27	26,7	28,3	27,7
	ENS_TRE	19,7	19,6	23,9	23,5	26	25,2	27,2	27	28	27,7	29,7	29,2
	TREES	20,3	20,9	24,3	24,2	26,2	26,2	27,9	27,4	28,4	28	28,8	28,8
SL	GPR	19,8	20,4	23,9	24,4	25,9	26	28	28,2	29,9	30,2	31,9	32,3
	SVM	19,4	19,8	23,5	24	25,5	25,5	27,4	27,4	29,1	29,5	30,8	31,2
	ENS_TRE	20,4	20,5	24,4	24,4	26,2	26,2	28,7	28,4	30,9	31,1	32,7	32,2
	TREES	20,9	21,5	25,4	25,2	27	27	28,6	28,5	30,8	30,3	32	31,9
MP	GPR	19	18,3	21,5	21,1	22,5	22,4	23,7	23,5	25	25	26,6	26,5
	SVM	18,4	18	20,9	20,8	22	22,1	23	23	24,4	24,5	26	26,1
	ENS_TRE	19,2	18,5	21,7	21,4	23	22,5	24,4	23,6	25,7	25,3	27	26,6
	TREES	20,2	19,6	22,6	22	23,3	23,4	24,5	24,3	26	26,2	27,5	27,3
SG	GPR	19,7	20,5	23,1	23,3	25,1	25,2	26,8	27	28,5	28,7	29,7	30
	SVM	19,4	20,1	22,9	23,4	24,6	25,4	26,4	26,8	27,9	28,4	29,4	30,1
	ENS_TRE	20,2	20,3	23,6	23,5	25,3	25,7	27,2	27	28,9	28,3	30,2	30,3
	TREES	20,8	21,3	24,1	24,4	26,3	26,6	27,7	28,1	29	29,8	30,1	30,8
IR	GPR	15,8	16,8	19,6	19,9	21,2	21,2	22,1	22,4	23,3	23,3	24,8	24,8
	SVM	15,5	16,3	19,1	19,2	20,7	20,3	21,4	21,3	23,2	23	25	25,2
	ENS_TRE	15,8	16,3	19,3	19,3	20,9	20,7	22,1	22,3	23,6	23	25,3	24,8
	TREES	16,8	17,8	20,2	20,2	21,5	21,4	22,9	22,7	24,2	23	25,8	25,5
KR	GPR	17,1	20	21	22,1	22,3	23,3	22,5	23,5	23,2	23,9	24	24,6
	SVM	16,8	19,3	20,3	21,2	21,7	22,2	21,9	22,6	22,8	23,5	23,7	23,9
	ENS_TRE	18,3	19,7	22,5	22,5	23	24,1	23,7	24,4	24,5	25,2	25,4	25,6
	TREES	18,8	20,5	22	22,5	23,8	24,2	24,1	24,7	24,3	25,2	25,4	25,7

nMBE(%)													
RO	*GPR*	0,4	-0,7	1,2	0,3	2,5	1,4	4,2	3,2	5,2	4,1	6,7	5,4
	SVM	1,7	-0,4	3,3	1,9	4,7	3,3	6,9	6,3	8,5	7	10,6	9,6
	ENS_TRE	0,5	-0,9	1,7	0,4	2,5	1,2	4,8	3,5	5,1	3,5	5,8	4,4
	TREES	1	-0,6	1,4	0,5	2,6	1,3	4	3	4,1	4,5	6,6	5,1
SL	*GPR*	-0,6	-0,5	-0,6	-0,5	-0,5	-0,3	-0,6	-0,1	-1	-0,4	0,2	0,7
	SVM	0,1	0,5	1,1	2,1	2,1	3,2	3,1	3,6	2,3	3,5	4	4,2
	ENS_TRE	-1,1	-0,5	-1,1	-0,6	-1,5	-0,7	-1,5	-0,5	-1,6	-0,1	-0,6	0,3
	TREES	-0,8	-0,2	-1,1	-0,4	-1	0	-1,8	-0,4	-2,4	-0,4	-0,4	0,9
MP	*GPR*	-1,3	-1,4	-1,8	-1,9	-1,4	-1,4	-0,9	-1,1	-1,2	-1,4	-0,8	-0,8
	SVM	0,1	-0,9	0	-0,2	0,5	0,5	0,4	0,3	1,2	1	1,9	1,2
	ENS_TRE	-2,1	-1,9	-2,4	-2,6	-2,3	-1,9	-2,2	-1,6	-2	-1,9	-1,9	-1,7
	TREES	-1,9	-1,9	-3	-2,8	-2,8	-2,6	-2	-2,3	-2,5	-2,1	-2,4	-1,6
SG	*GPR*	0,8	1,5	1	1,6	1,6	2,1	1,8	2,5	2	2,6	1,9	2,3
	SVM	1,9	2,1	2	3,3	2,8	4,7	3,7	4,8	4	5,4	5,5	5,5
	ENS_TRE	0,5	1,6	0,8	1,4	1	1,9	1,4	2,3	1,9	2,4	1,7	2,3
	TREES	0,9	2	0,6	1,3	1,7	2,3	1,6	2,5	2,2	3,3	2,7	3,3
IR	*GPR*	2,2	2,6	2,8	2,7	3,4	3,2	3,7	3,5	4,9	4,9	7,9	7,9
	SVM	3,8	4	5,8	4,9	7,1	6	8,3	6,5	9,8	8,8	13,3	11,9
	ENS_TRE	2,1	2,9	2,7	2,6	3	3,4	2,9	3,3	3,9	4,9	7,1	8
	TREES	2,2	2,8	2,8	3	3,5	3,7	3,6	3,2	4,8	5,7	8,2	8,5
KR	*GPR*	-2,1	-5	-3,1	-4,4	-3,3	-4,2	-3	-3,9	-2,4	-3,7	-2	-2,6
	SVM	-1	-3,2	-0,8	-2,4	-0,9	-2	0,3	-1	0	-0,7	0,7	-0,5
	ENS_TRE	-2,5	-5,2	-3,9	-4,4	-3,4	-4,7	-4,1	-5,2	-3,6	-4,9	-3,1	-4,2
	TREES	-2,4	-4,4	-2,7	-4,5	-3,4	-4,1	-3,8	-4,8	-3,2	-4,2	-3,5	-3,6

Annex B

Post processing WRF forecasting errors according to station and time horizon for ground-measured data and satellite-derived data, three months of 2016.

Station	Model	h+1		h+2		h+3		h+4		h+5		h+6	
		GND	SAT	GND	SAT	GND	SAT	GND	SAT	GND	SAT	GND	SAT
nRMSE(%)													
RO	*WRF*	39,3		38,5		38		37,8		40,5		41,7	
	KALMAN	38,5	38,3	37,6	37,6	37,3	37,5	37,2	38,4	39,8	39,9	41,1	41,6
SL	*WRF*	38,8		38,5		38,3		39,2		40,7		41,7	
	KALMAN	38,1	36,6	37,6	36,7	37,4	36,7	38,5	38	40,2	39,9	41,6	41,3
MP	*WRF*	34,9		34,1		34,6		35		36,2		40,1	
	KALMAN	33	32,3	32,6	32,6	32,5	32,4	33,5	33,5	34,7	35,4	38,5	39,4
SG	*WRF*	40,9		40,8		41,3		42,5		45		45,5	
	KALMAN	40,1	40,8	39,9	41,8	40,2	43,2	41,1	42,5	44,1	44,9	45,3	46,6
IR	*WRF*	44,3		44,4		44,6		45,3		48,1		52,3	
	KALMAN	42,7	41,2	42,6	41,4	42	41,7	43,4	42,3	45,2	44,6	46,9	49,7
KR	*WRF*	37,6		37,6		40,5		41,1		41,3		41,7	
	KALMAN	34,3	34,3	35,4	35,6	37,6	38,2	38,2	37,3	36,3	34,7	36	36,3
nMAE(%)													
RO	*WRF*	29,1		28,6		28,1		27,8		29		29,5	
	KALMAN	28,7	28,2	28,3	27,7	27,6	27,5	27,1	28,1	28,3	28,6	28,8	30,1
SL	*WRF*	28,8		28,5		28,4		29,2		30,5		31,6	
	KALMAN	28,7	27,6	28,4	27,5	28,4	27,7	29	28,6	30,7	30,2	31,9	31,5
MP	*WRF*	26,2		25,6		26		26,2		26,8		29,5	
	KALMAN	25,1	24,3	24,8	24,9	24,6	24,3	25,4	25,2	26,3	26,1	29,2	28,6
SG	*WRF*	30,5		30,7		31,1		31,6		33,5		33,6	
	KALMAN	30,1	30,1	30,4	30,8	30,5	32,2	30,5	31,7	32,7	32,6	32,9	33,6
IR	*WRF*	33,2		33,6		34		34,9		37,4		41,5	
	KALMAN	32,2	32,1	32,3	32,3	32,3	32	32,4	32,9	33,4	34,8	34,9	39,5
KR	*WRF*	28,5		28,5		28,8		27,9		26,4		26,6	
	KALMAN	28,5	26,9	28,6	26,7	28,6	26,7	27,7	26,5	25,7	25,2	25,9	26,2
nMBE(%)													
RO	*WRF*	10,4		11		11,8		12,7		14,6		15,2	
	KALMAN	6,1	4,5	6,8	5,3	7	5,7	7,5	4,9	8,8	6,1	9	6,4
SL	*WRF*	4,7		4,7		4,6		5,5		6,8		6,7	
	KALMAN	3	1,9	2,9	0,9	2,8	1,8	2,8	3,6	3,1	3	4,2	3,4
MP	*WRF*	12,6		11,6		11,7		11,3		11,8		12,3	
	KALMAN	6,8	5,6	5,9	3,9	5,2	4,1	5,1	2,8	5,1	0,7	5,7	2,9
SG	*WRF*	11,8		11,5		12,6		14,3		14,6		14,5	
	KALMAN	7,1	7,2	5,8	7,5	6,3	9,1	7,6	9,3	7,9	10,8	7,8	9,8
IR	*WRF*	13,9		14,7		15,8		17,5		22,3		31,3	
	KALMAN	6,5	3,7	6,8	5,9	7,7	9,4	8,7	11,2	10,8	15,7	15,3	23,4
KR	*WRF*	6,3		7,8		8,5		9,3		11		14,1	
	KALMAN	3,2	2,3	4,6	4,1	4,5	4,8	5,1	5,1	6	6,2	7,9	9,1

Annex C

Forecasting errors of ensemble method using satellite-derived data according to station and time horizon for three months of 2016.

Station	Model	h+1	h+2	h+3	h+4	h+5	h+6
nRMSE(%)							
RO	SP	33,5	38,2	42,3	45,7	48,8	49,7
	ARMA	26,9	30,3	33,0	35,5	37,7	37,8
	GPR	25,6	29,3	32,1	33,7	35,2	35,9
	SVM	26,3	29,6	32,2	34,2	35,9	35,4
	KALMAN	37,9	36,9	36,9	37,7	39,3	40,8
	ENS	24,3	28,8	31,4	33,0	33,9	34,5
MP	SP	30,0	36,0	42,2	49,2	55,4	60,9
	ARMA	23,8	27,0	29,1	31,0	32,1	35,4
	GPR	24,4	26,6	28,1	29,9	31,9	34,8
	SVM	24,4	26,8	28,6	30,3	32,5	35,4
	KALMAN	32,1	32,3	32,1	33,3	35,3	39,2
	ENS	23,7	25,5	27,2	29,8	30,5	33,2
SG	SP	33,6	38,0	42,3	47,6	51,2	56,4
	ARMA	29,5	32,7	35,6	38,1	40,4	41,9
	GPR	27,4	31,1	33,8	36,2	38,8	41,0
	SVM	28,3	31,7	34,6	36,7	39,6	41,9
	KALMAN	40,6	41,3	42,5	41,8	44,2	45,9
	ENS	26,9	27,6	29,3	32,1	35,2	38,7
IR	SP	31,8	38,4	44,5	49,6	54,7	60,7
	ARMA	26,4	31,8	35,0	35,9	39,1	41,0
	GPR	23,9	28,5	31,1	31,4	34,8	38,0
	SVM	24,2	29,1	32,2	32,5	37,0	41,2
	KALMAN	39,6	39,8	40,5	41,0	43,4	48,0
	ENS	23,7	27,0	29,5	30,2	32,2	35,8
KR	SP	22,7	26,4	29,3	31,1	32,4	35,5
	ARMA	20,4	23,9	25,7	25,8	26,5	28,5
	GPR	19,9	22,5	24,3	24,1	23,9	24,9
	SVM	20,2	22,5	24,3	24,0	24,2	25,2
	KALMAN	26,5	26,6	26,6	26,3	25,0	26,0
	ENS	17,4	18,9	20,6	21,9	22,3	23,9

nMAE(%)							
RO	SP	24,3	27,9	30,1	32,3	33,4	32,8
	ARMA	20,0	23,0	25,7	27,9	29,4	29,7
	GPR	18,5	21,9	24,2	25,3	26,1	26,3
	SVM	18,7	21,7	23,5	24,7	25,4	25,3
	KALMAN	27,9	27,4	27,2	27,9	28,5	29,8
	ENS	17,2	19,1	20,4	22,2	23,6	24,8
MP	SP	22,2	26,8	31,0	35,5	40,1	44,3
	ARMA	18,2	21,4	22,8	24,4	25,4	28,3
	GPR	18,2	20,4	21,2	22,7	24,3	26,4
	SVM	18,0	20,1	21,4	22,8	24,2	26,5
	KALMAN	24,1	24,8	24,1	25,1	26,1	28,5
	ENS	16,1	18,8	19,9	21,2	23,2	25,5
SG	SP	24,8	28,1	31,4	34,6	36,3	39,2
	ARMA	22,6	25,6	27,7	29,2	31,2	32,5
	GPR	20,0	23,7	26,0	27,7	29,6	30,9
	SVM	20,6	23,8	25,7	27,6	29,7	31,3
	KALMAN	30,0	30,7	31,9	31,5	32,6	33,5
	ENS	18,4	20,9	24,3	25,8	26,2	28,2
IR	SP	24,4	29,6	34,3	37,6	40,1	43,7
	ARMA	20,4	24,7	27,0	27,9	29,6	30,3
	GPR	17,2	20,7	22,4	22,8	24,1	26,2
	SVM	17,0	20,5	22,3	22,6	24,5	27,2
	KALMAN	31,1	31,2	31,3	32,3	34,0	38,4
	ENS	15,8	19,3	20,2	20,8	22,5	25,0
KR	SP	32,5	37,6	42,9	46,6	49,3	53,5
	ARMA	27,5	31,2	33,9	34,3	35,2	37,4
	GPR	27,4	30,5	33,1	33,2	33,5	34,4
	SVM	27,9	30,8	33,7	33,8	35,0	35,5
	KALMAN	36,9	36,9	37,5	36,8	34,2	35,8
	ENS	21,8	23,4	25,4	27,3	28,4	30,8

nMBE(%)							
RO	SP	-10,7	-10,5	-10,6	-11,0	-11,6	-13,2
	ARMA	-1,4	0,4	2,0	3,7	5,2	4,4
	GPR	-1,9	-1,6	-0,8	0,4	2,0	1,9
	SVM	-3,7	-1,7	-0,4	1,7	2,3	3,6
	KALMAN	4,2	5,0	5,4	4,7	6,1	6,1
	ENS	-3,2	-2,5	-2,1	-1,0	0,8	0,8
MP	SP	-10,9	-13,1	-14,2	-16,9	-19,3	-22,3
	ARMA	-1,1	-0,8	0,7	0,5	0,0	-0,2
	GPR	-3,5	-3,7	-2,4	-2,3	-3,2	-3,0
	SVM	-3,1	-2,4	-1,1	-1,8	-1,0	-1,1
	KALMAN	5,5	3,7	3,9	2,8	0,7	2,7
	ENS	-2,9	-4,0	-2,8	-2,6	-3,3	-2,7
SG	SP	-7,4	-5,9	-4,5	-3,2	-3,8	-4,5
	ARMA	0,7	1,3	1,9	2,5	2,5	2,1
	GPR	0,0	-0,9	-0,1	0,3	-0,2	-1,7
	SVM	-1,2	-0,5	1,1	1,3	1,2	-0,7
	KALMAN	7,2	7,5	8,8	9,2	10,8	9,7
	ENS	-1,3	-1,9	-1,0	-1,7	-1,3	-3,0
IR	SP	-6,9	-8,7	-10,8	-13,2	-13,0	-11,0
	ARMA	0,5	3,0	4,5	4,5	6,1	8,4
	GPR	0,1	1,1	2,5	2,4	5,1	7,6
	SVM	0,5	2,1	4,5	4,1	7,7	10,5
	KALMAN	2,5	4,7	6,6	7,4	8,8	9,1
	ENS	0,9	1,4	2,4	2,3	5,3	8,0
KR	SP	-11,9	-12,3	-13,8	-15,3	-17,2	-18,7
	ARMA	-2,9	-0,4	0,3	1,1	1,2	2,9
	GPR	-2,4	-1,5	-1,3	-1,1	-1,6	-0,9
	SVM	-1,9	-1,1	-1,0	0,8	1,4	1,8
	KALMAN	1,8	3,9	4,6	4,8	6,0	8,9
	ENS	-3,3	-3,5	-2,1	-2,2	-2,9	-1,9