Short-Term Photovoltaic Generation Forecasting Enhanced by Satellite Derived Irradiance

Kevin Bellinguer, Robin Girard, Guillaume Bontron, Georges Kariniotakis

SHORT-TERM PHOTOVOLTAIC GENERATION FORECASTING ENHANCED BY SATELLITE DERIVED IRRADIANCE

Kevin Bellinguer[1*]*, Robin Girard*[1]*, Guillaume Bontron*[2]*, Georges Kariniotakis*[1]

[1]*MINES ParisTech, PSL University, PERSEE - Centre for Processes, Renewable Energies and Energy Systems,*
CS 10207, rue Claude Daunesse, 06904 Sophia-Antipolis Cedex, France
[2]*Compagnie Nationale du Rhône, 2 Rue André Bonin, Lyon, France*
** kevin.bellinguer@mines-paristech.fr*

Keywords: PHOTOVOLTAICS, SHORT-TERM FORECASTS, SPATIO-TEMPORAL MODELS, SATELLITE IMAGES, SMART-GRID

Abstract

In a context of natural resources depletion, weather-dependent renewable energy sources play an increasingly important role in the energy mix. Yet, high shares of renewables can jeopardise the safe operation of power grid due to their variable nature. To address this challenge, it is essential to know the future amount of energy produced to balance production and consumption. This paper aims at investigating photovoltaic generation short-term forecasting and particularly spatio-temporal approaches. These approaches permit to exploit the spatial dependency of weather variables to provide valuable information regarding cloud movements. Thus, it is possible for a power producer to take advantage of dense PV plants networks by considering spatially distributed units as remote sensors. For low-density network, satellite-derived information observed in the vicinity of the power unit location offers an interesting alternative. To reduce the computational burden induced by this data source, feature-selection approaches are implemented. Usually, a correlation score is used to measure the dependence between lagged satellite-based time-series with the target feature (i.e. power production observations). However, this approach tends to provide redundant information (i.e. highly correlated pixels). To address this issue, we implement a minimal-Redundance Maximal-Relevance framework. Performance comparisons with state-of-the-art approaches are also performed.

Nomenclature

t	Launching time of the forecast
h	Forecasting horizon
x	Power unit of interest
$\boldsymbol{I}_t$	Irradiance-based quantity
$\boldsymbol{I}_t^{CS}$	Irradiance observed under clear-sky conditions
$\boldsymbol{P}_t$	Photovoltaic production
P_c	Installed capacity
$\boldsymbol{S}_t$	Satellite-derived surface irradiance
N_s	Number of satellite-derived surface irradiance features
N	Number of paired observations
f_{RF}	Random forest regression model used to infer the statistical relationship between inputs and output
A	Normalised PV production observed at time $t+h$
B	Normalised satellite-derived surface irradiance at a grid point at time t
$p_{A,B}$	Joint probability distribution of A and B
p_A	Marginal probability distribution of A
$\hat{\cdot}$	Forecast quantity
$\bar{\cdot}$	Normalised quantity

1 Introduction

Photovoltaic (PV) technology is one keystone of the global energy shift initiated to reduce anthropogenic greenhouse gases emissions; in Europe, on-grid PV plant installations increased from nearly 9 GW in 2018 up to 19 GW in 2019 [1]. This capacity growth is expected to continue over the next years due to costs reduction [2].

PV generation is characterised by a high variability and a limited predictability under non-clear sky conditions resulting from its dependence over weather. This results to several challenges for the secure and economic operation of the power systems especially in cases with high PV penetration. Weather variability impacts also economic profitability of renewable energy (RES). In a market-oriented environment, operators that sell energy have to pay financial penalties proportional to unplanned production fluctuations to the transmission system operator.

To support RES integration within the power grid, accurate forecasting tools are required. Unlike wind production forecasting, which dates back to the 1980s [3], research on PV production forecasting is a more recent field of research. Nevertheless, the state of the art has developed rapidly in recent years; in this regard [4, 5] provide fairly complete literature reviews. The main source of information for short-term (i.e. from 15-min up to 6-h ahead) PV production forecasting (PVPF) models are past production measurements. Recent studies highlight that forecasts of the upcoming production (i.e. for several hours-ahead) benefit from spatio-temporal (ST) approaches. This family of model takes advantage of spatial dependencies that exist between weather state at the farm of interest and its surrounding so as to derive information regarding cloud motions. Reference [6] proposes a ST model based on geographically distributed PV units while [7] considered satellite-derived surface irradiance (SDSI). ST approaches can be easily applied by actors like aggegators that manage a portfolio of geographically distributed PV plants (e.g. a virtual power plant) and for which they dispose their production data. In the cases of very distant plants (spatial correlations are very small or nonexistent), or areas with low power units density (low ability to capture spatial effects through measurements, although the distance permits to detect correlations), geostationary satellite-based information offers more flexibility in the sense that it can fill the gap of sensors by covering large areas. Moreover, due to constant innovations in the satellite imagery field, satellite information is now delivered with high reliability and at sub-hourly frequencies (i.e. every 15 minutes and practically without delays in delivery), which promotes an operational use.

In the literature, one finds several approaches to deal with satellite-based information. One of the most widespread methods consists in extrapolating clouds displacement thanks to motion extraction techniques developed in the image processing field. Thus, a block-matching method applied to two successive images permits to identify positions of similar cloud structures and then to derive the displacement vectors. Cloud Motion Vectors (CMV) are then used to translate the most recent map by assuming that cloud structure remains unchanged over time [8]. CMV-based methods reveal interesting forecasting performances up to 2 hours ahead. Forecasting performances can be extended to further horizons by considering wind velocity computed by numerical weather prediction (NWP) models as displacement vectors. The main drawback of CMV approaches lies in their ineffectiveness in the case of local clouds formation [9]. More recent studies resort to statistical or deep learning approaches. For instance, [10] proposes a straightforward modelling chain, where a satellite image is flattened and fed to a support vector regression model which provides a forecast of PV production. In [11], the authors propose a deep neural network architecture, which extracts relevant features from three consecutive satellite images (thanks to the use of a convolutional neural network), which are then combined with meteorological data and fed into an artificial neural network to derive irradiance forecasts. To cope with the dimensional burden induced by satellite data set, it is a common practice to resort to simpler approaches: one can consider a set of well-chosen pixels fed to the forecasting model. To do so, a maximal relevance feature selection (MRFS) scheme is usually implemented: a correlation scores analysis performed between time-lagged satellite-derived information and PV production is used to select pixels having the highest scores. [9, 12] use the Pearson correlation score while [13, 14] consider the Mutual Information (MI) criterion for its ability to identify non-linear relationships.

In this paper, we focus on the question of how to perform an optimal pixel selection for feeding the PVPF model. It is fundamental to decrease the dimensionality of the problem, and thus in line with the principle of parsimony in forecasting while it opens the way to consider more easily additional sources of information (e.g. alternative sources of satellite images). In addition, these investigations are easy to replicate and do not require additional information such as NWPs. The present paper proposes the following original contributions:

- First, we observe that the MRFS scheme tends to select satellite pixels aggregated in some spatial regions. Therefore, selected features carry redundant information. To address this issue, we apply the minimal-redundancy-maximal-relevance (mRMR) incremental selection framework, introduced by [15]. This approach permits the selection of spatially distributed features.
- Then, the selected pixels are embedded in a PVPF model and confronted with traditional features selection processes based on Pearson correlation and MI.

The paper is organised as follows. First, Section 2 introduces the observational data used as case study. Then, Section 3 describes the methodology proposed to assess the features selection while Section 4 discusses the outcomes. Finally, Section 5 draws the conclusions of this study.

2 Case Study

We investigate a data set composed of PV production and SDSI observations. Both data sets have a temporal resolution of 15 minutes. The forecasting models are trained over the year 2015, while the year 2016 is used for out-of-sample testing purposes.

2.1 PV Power Production

The PV production measurements are provided by the Compagnie Nationale du Rhône which is France's leading producer of exclusively renewable energy. This data set is composed of 9 non-tracking grid-connected systems, with a capacity ranging from 1.2 up to 12 MWp. These 9 PV plants are located in the Rhône valley and especially along the Rhône River.

2.2 Satellite-Derived Surface Irradiance

We considered satellite images obtained from the HelioClim-3 database with the HelioSat-2 method [16]. Helioclim-3 images provide an estimation of the ground horizontal irradiance with a spatial resolution of $0.0625° \times 0.0625°$. Pixels constituting satellite-based images are converted into a set of time-series.

3 Methodology

3.1 Data Stationarity

By nature, irradiance-based quantities are non-stationary (i.e. the statistical properties of the time-series change over time) due to astronomical and meteorological phenomena. A common way to make PV production process easier to analyse, consists in normalising the irradiance-based quantity, $\boldsymbol{I}_t$, with a clear sky model output $\boldsymbol{I}_t^{CS}$ (equation (1)). This is a model estimating irradiance with clear sky conditions, here, the MacClear model [17] is used.

$$\bar{\boldsymbol{I}}_t = \frac{\boldsymbol{I}_t}{\boldsymbol{I}_t^{CS}} \tag{1}$$

3.2 The PV Production Forecasting Model

The core of the forecasting approach developed in this paper is a random forest (RF) regression model [18]. We turn to this model because it is widely used in the RES forecasting domain and it tends to be considered as an advanced PVPF model. Furthermore, it has the ability to draw non-linear relationship between input and output features. It is an ensemble learning method composed of multiple regression trees grown in parallel. This combination of trees associated with bootstrap aggregating (i.e. bagging) approaches permits to overcome over-fitting and lack of accuracy issues which are inherent to tree models.

$$\widehat{\overline{\boldsymbol{P}}}^{\boldsymbol{x}}_{\boldsymbol{t+h|t}} = f_{RF}\left(\overline{\boldsymbol{P}}^{\boldsymbol{x}}_{\boldsymbol{t-1H:t}}, \overline{\boldsymbol{S}}^{\boldsymbol{1:N_s}}_{\boldsymbol{t-1H:t}}\right) \tag{2}$$

The generic forecasting model, represented by equation (2), is fed with normalised past PV production observations at the site of interest, $\overline{\boldsymbol{P}}^{\boldsymbol{x}}_{\boldsymbol{t-1H:t}}$ and normalised SDSI features $\overline{\boldsymbol{S}}^{\boldsymbol{1:N_s}}_{\boldsymbol{t-1H:t}}$. Past observations (i.e. from time $t-1H$ up time t) of both variables are included so as to take into account their evolution over time. Given the high dimensionality of SDSI (e.g. a map centred over the PV farm with a radius of 50km produces about 230 features while for 150km the features increase to 1870), raw satellite information can hardly be integrated within advanced PVPF models without experiencing high computational costs. Therefore, a pre-selection step is required so as to keep only the N_s most informative features.

3.3 Feature Selection Framework

Feature selection frameworks described hereinbelow aim at finding subsets of SDSI features providing relevant information regarding future PV production. To do so, we confront SDSI features observed at time t with PV production measured at time $t+h$ for each PV unit and for each forecasting horizon. The feature selection is performed via the learning set.

3.3.1 Maximal Relevance Feature Selection: In the literature it is a common practice to implement a MRFS scheme to determine the N_s features having the highest dependence with the PV production observations. Here, we focus on MRFS frameworks based either on the Pearson correlation score or on the MI criterion (defined at equation (3)). Pearson coefficient measures the linear correlation between two sets of data, while the MI permits to identify non-linear relationships.

$$I(A,B) = \sum_{a\in A}\sum_{b\in B} p_{A,B}(a,b)\log\left(\frac{p_{A,B}(a,b)}{p_A(a)p_B(b)}\right) \tag{3}$$

In the Figure 1 we depict a part of a satellite image around a PV plant (grey dot). The radius corresponds to 50 km. With the black dots we indicate the pixels selected through the feature selection algorithms. Left graph of Figure 1 highlights that MRFS tends to select features which are very close spatially. Thus, such an approach selects geographically close variables carrying redundant information, while making the forecasting model "blind" in some spatial direction (e.g. in this case, no information is provided regarding easterly cloud distribution).

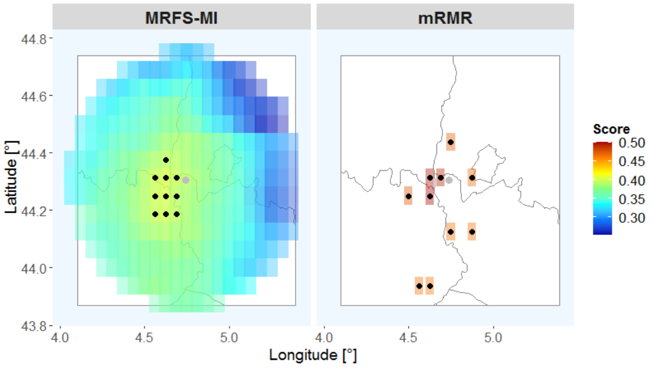

Fig. 1 MRFS-MI and mRMR score-based maps obtained considering a 1-hour forecasting horizon. Black dots stand for the position of the 10 most informative features while the grey dot represents the site location.

3.3.2 Minimal-Redudant Maximal-Relevance Feature Selection: To tackle with the drawback of the MI-based approach described above, we implement the mRMR incremental selection framework proposed in [15]. To the authors knowledge, this method, which has been initially applied in the bioinformatics field, has never been tested within the PVPF domain.

The mRMR approach finds the most relevant and least redundant feature subset through an iterative process based on MI. To do so, we adopt a forward selection scheme which incrementally selects N_s features by identifying variables which possess high MI with the target variable (i.e. maximal dependency), while having a low correlation with already selected features (i.e. low redundancy).

The right graph of Figure 1 shows that the mRMR approach applied with SDSI features manages to identify grid points that are geographically distributed around the PV plant. Thus, this approach provides the PVPF model with sparse information which makes it more fitted to identify various weather dynamics. Of course the ultimate evaluation of the differences between the two approaches will be made upon the prediction performance of the model for each case.

4 Evaluation Results

4.1 Reference Model

To judge the performances improvement resulting from the ST approaches, we consider the RF model fed with only temporal-based inputs as our baseline model. This model is defined by equation (4):

$$\widehat{\boldsymbol{P}}^{\boldsymbol{x}}_{\boldsymbol{t+h|t}} = f_{RF}\left(\overline{\boldsymbol{P}}^{\boldsymbol{x}}_{\boldsymbol{t-1H:t}}\right) \tag{4}$$

4.2 *Evaluation Metrics*

To assess the forecasting performances of the different approaches under consideration, it is necessary to define some metrics to quantify differences between observations and forecast values. With a view to enable inter-comparison with other studies, we turn to widely used metrics, namely the normalised Root Mean Square Error (nRMSE) and the normalised Mean Absolute Error (nMAE), which are respectively described by equations (5) and (6). Scores are computed individually for the nine PV farms. Then, for ease of understanding and to place ourselves in a context of production aggregation, the scores are averaged. Nighttime data are discarded inasmuch as they do not offer relevant information.

$$nRMSE(h) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\frac{\widehat{\boldsymbol{P}}^{\boldsymbol{x}}_{\boldsymbol{t+h|t}} - \boldsymbol{P}^{\boldsymbol{x}}_{\boldsymbol{t+h}}}{P^{x}_{c}}\right)^{2}} \tag{5}$$

$$nMAE(h) = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{\widehat{\boldsymbol{P}}^{\boldsymbol{x}}_{\boldsymbol{t+h|t}} - \boldsymbol{P}^{\boldsymbol{x}}_{\boldsymbol{t+h}}}{P^{x}_{c}}\right| \tag{6}$$

To compare the forecasting performances of models investigated in this paper with respect to the reference model, the comparison skill score defined at equation (7) is used. A positive (negative) skill score means that the model has better (worse) forecasting performances than the reference model.

$$SS_{Model}(h) = \frac{Score_{Ref}(h) - Score_{Model}(h)}{Score_{Ref}(h)} \times 100\% \tag{7}$$

4.3 *Performances Assessment*

A sensibility analysis, performed over the number of SDSI features ($N_s \in [\![5, 50]\!]$) reveals that Peason-based, MI-based and mRMR-based selections perform better when considering the first 10 features with highest dependence scores. Beyond these values, the increase of features does not improve forecasting performances. In addition, similar investigation regarding the influence of the SDSI feature selection processes has been performed with an auto-regressive (AR) model. ARIMA models are a family of linear models that is well suited for short-term predictions and ST approaches [19]. It turns out that AR model fed with SDSI features is less parsimonious than the RF-based approach, while its performances are significantly lower. Due to page restrictions, details of these analysis are omitted.

Figure 2 represents forecasting performances of RF models fed with the best set of SDSI features provided by each feature selection process. First, we observe that the feature selection approaches have very little influence on forecasting horizons lower than 1 hour. As for MRFS schemes, the RF forecasting model fed with MI-based selected features slightly outperforms its counterpart, trained with features selected with the Pearson correlation score, for leading times higher than 2 hours. On the other hand, best forecasting performances are achieved by the mRMR selection scheme, both in terms of nRMSE and nMAE, for horizons higher than 1 hour ahead.

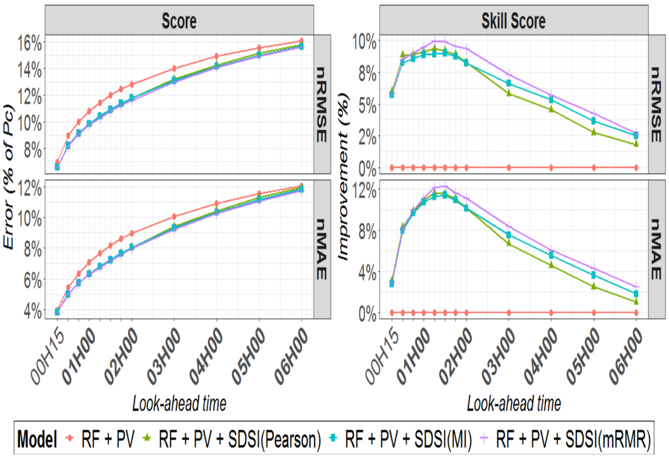

Fig. 2 Influence of the SDSI features selection process over the forecasting performance of the RF model.

As the nRMSE and nMAE differences are very low between the considered models, we implement the Diebold-Mariano (DM) test to judge the statistical significance of the differences.

The DM test compares the predictive accuracy of two forecast models. The time loss differential between the two forecasts is denoted by $d_{12,t} = |e_{1,t}| - |e_{2,t}|$ with $e_{i,t}$ being the forecast error. The two forecasts have equal accuracy if the expectation of the loss differential is zero (that constitutes the null hypothesis: $H_0 : E(d_{12,t}) = 0, \forall t$). Under the null hypothesis, the DM test follows the standard normal distribution (equation (8)) [20]. We suppose a significance level of 5%. As a result, DM statistics that fall outside the range defined by the 2.5% and 97.5% quantiles of the normal distribution (i.e. -1.96 and +1.96) enable the rejection of the null hypothesis.

$$DM_{12} = \frac{\bar{d}_{12}}{\hat{\sigma}_{\bar{d}_{12}}} \sim \mathcal{N}(0, 1) \tag{8}$$

Figure 3 highlights that for horizons between 45' and 2 hours, forecasts issued by the Pearson-based and the MI-based forecasting models are not statistically different. On the other hand, the difference between the forecasts delivered by the mRMR-based model and the MI-based approach are significant for horizons higher than 1 hour.

5 Conclusion and Perspectives

In this paper, a new features selection framework is applied to satellite-derived information in the context of PVPF. The proposed approach permits to select a subset of low-correlated variables, which ensure spatially distributed pixels around the power unit. Then, a performance comparison performed with two other features-selection schemes revealed that: (1) the selection scheme has little importance for very-short term horizons, (2) for higher horizons mRMR-based model outperforms

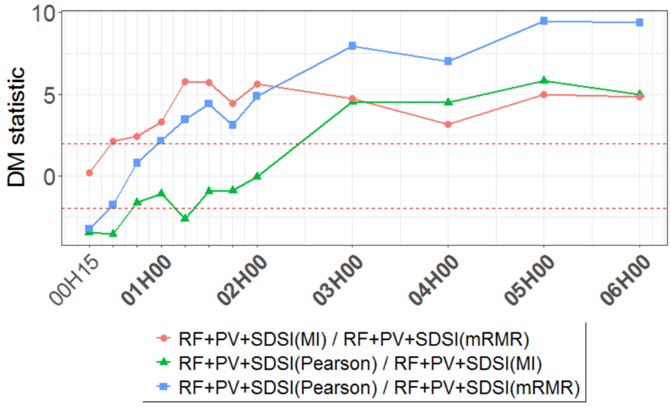

Fig. 3 DM statistic between the three SDSI feature selection frameworks studied for different forecast horizons. The red dotted lines stand for the borders delimiting the validation and rejection of the null hypothesis.

forecasting approaches based on the Pearson correlation coefficient and the MI criterion. The resulting features selection with the mRMR appoach is compatible with what is intuitively expected as result, that is to select pixels that are geographical distributed around the PV plant. Present works have been performed with irradiance-based information but future works should apply the proposed method to infra-red satellite images so as to improve forecast accuracy of the early morning.

6 Acknowledgements

The authors would like to thank Transvalor for providing the SDSI data set.

7 References

[1] "IRENA RE Time Series." [Online]. Available: https://public.tableau.com/profile/irena.resource#!/vizhome/IRENARETimeSeries/Charts

[2] IRENA, "Future of Solar Photovoltaic: Deployment, investment, technology, grid integration and socio-economic aspects (A Global Energy Transformation: Paper)," Tech. Rep., 2019.

[3] G. Kariniotakis, *Renewable Energy Forecasting: From Models to Applications*. Elsevier - Woodhead Publishing, 2017.

[4] J. Antonanzas, N. Osorio, R. Escobar, and al., "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, pp. 78–111, Oct. 2016.

[5] S. Sobri, S. Koohi-Kamali, and N. A. Rahim, "Solar photovoltaic generation forecasting methods: A review," *Energy Conversion and Management*, vol. 156, pp. 459–497, Jan. 2018.

[6] X. G. Agoua, R. Girard, and G. Kariniotakis, "Short-Term Spatio-Temporal Forecasting of Photovoltaic Power Production," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 538–546, Apr. 2018.

[7] L. M. Aguiar, B. Pereira, P. Lauret, and al., "Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting," *Renewable Energy*, vol. 97, pp. 599–610, Nov. 2016.

[8] S. Cros, O. Liandrat, N. Sébastien, and N. Schmutz, "Extracting cloud motion vectors from satellite images for solar power forecasting," in *2014 IEEE Geoscience and Remote Sensing Symposium*, Jul. 2014, pp. 4123–4126.

[9] L. Mazorra Aguiar, B. Pereira, M. David, F. Díaz, and al., "Use of satellite data to improve solar radiation forecasting with Bayesian Artificial Neural Networks," *Solar Energy*, vol. 122, pp. 1309–1324, Dec. 2015.

[10] D. P. Larson and C. F. M. Coimbra, "Direct Power Output Forecasts From Remote Sensing Image Processing," *Journal of Solar Energy Engineering*, vol. 140, no. 021011, Feb. 2018.

[11] Z. Si, Y. Yu, M. Yang, and P. Li, "Hybrid Solar Forecasting Method Using Satellite Visible Images and Modified Convolutional Neural Networks," *IEEE Transactions on Industry Applications*, vol. 57, no. 1, pp. 5–16, Jan. 2021.

[12] K. Bellinguer, R. Girard, G. Bontron, and al., "Short-Term Photovoltaic Generation Forecasting Using Multiple Heterogenous Sources of Data," in *36th European Photovoltaic Solar Energy Conference and Exhibition*, Oct. 2019, pp. 1422–1427.

[13] T. Carriere, C. Vernay, S. Pitaval, and al., "A Novel Approach for Seamless Probabilistic Photovoltaic Power Forecasting Covering Multiple Time Frames," *IEEE Transactions on Smart Grid*, pp. 1–1, 2019.

[14] K. Bellinguer, R. Girard, G. Bontron, and al., "Short-term Forecasting of Photovoltaic Generation based on Conditioned Learning of Geopotential Fields," in *2020 55th International Universities Power Engineering Conference (UPEC)*, 2020, p. 6.

[15] Hanchuan Peng, Fuhui Long, and C. Ding, "Feature Selection based on Mutual Information Criteria of Max-dependency, Max-relevance, and Min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[16] P. Blanc, B. Gschwind, M. Lefèvre, and al., "The HelioClim Project: Surface Solar Irradiance Data for Climate Applications," *Remote Sensing*, vol. 3, no. 2, pp. 343–361, Feb. 2011.

[17] M. Lefèvre, A. Oumbe, P. Blanc, and al., "McClear: A new model estimating downwelling solar radiation at ground level in clear-sky conditions," *Atmospheric Measurement Techniques*, vol. 6, pp. 2403–2418, 2013.

[18] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[19] R. J. Bessa, A. Trindade, and V. Miranda, "Spatial-Temporal Solar Power Forecasting for Smart Grids," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 1, pp. 232–241, Feb. 2015.

[20] F. X. Diebold, "Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests," *Journal of Business & Economic Statistics*, vol. 33, no. 1, pp. 1–1, Jan. 2015.