Solar Radiation Forecasting with Statistical Models

Luis Mazorra-Aguiar and Felipe Díaz

Abstract Renewable energy electrical generation has experienced significant growth in the recent years. Renewable energies generate electrical energy using different natural resources, such as solar radiation and wind fields. These resources present an unstable behavior because they depend on different meteorological conditions. In order to maintain the balance between input and output electrical energy into the power system, grid operators need to control and predict these fluctuating events. Indeed, forecasting methods are completely necessary to increase the proportion of renewable energies into the system (Heinemann et al. in Forecasting of solar radiation: solar energy resource management for electricity generation from local level to global scale. Nova Science Publishers, New York, 2006 [17], Wittmann et al. in IEEE J Sel Top Appl Earth Obs Remote Sens 1:18–27, 2008 [46]). Reducing the uncertainty of natural resources, operators could reduce maintenance costs, improve the interventions in the intra-day market and optimize management decisions with nonrenewable energies supply. Many forecasting methods are used to obtain solar radiation forecasting for different time horizons. In this chapter, we will focus on several solar radiation forecasting statistical methods for intra-day time horizons using ground and exogenous data as inputs.

1 Introduction

Solar radiation forecasting could be used for different purposes with a wide range of methods. Depending on these purposes, forecasting models are based on different input parameters and used for several time horizons [22, 42].

L. Mazorra-Aguiar (✉) · F. Díaz
University Institute for Intelligent Systems and Numerical Applications in Engineering,
University of Las Palmas de Gran Canaria, Edificio Central del Parque Tecnológico,
Campus de Tafira, 35017 Las Palmas de Gran Canaria, Spain
e-mail: luis.mazorra@ulpgc.es

F. Díaz
e-mail: felipe.diaz@ulpgc.es

R. Perez (ed.), *Wind Field and Solar Radiation Characterization and Forecasting*,
Green Energy and Technology, https://doi.org/10.1007/978-3-319-76876-2_8

- For time horizons less than hour models based on ground-based sky images obtain very good results. These models offer high precision information about cloud cover variability using sky images with 180 cameras [10, 45].
- Satellite image models are considered a very useful tool to improve solar radiation for time horizons up to several hours ahead. Geostationary meteorological satellites obtain images from atmosphere and satellite models estimate solar radiation using these images. In recent years, these models obtain accurate results with temporary resolution less than an hour and spatial resolution around 1–5 km. A review of several satellite models is shown in Sect. 1.2.
- Statistical models obtain accurate results for time horizons up to hours ahead. These models are not good enough to estimate the cloud motion but the high correlation between ground solar radiation data series made them very good tools for solar forecasting over 1 hour. The bibliography offers different statistical models for solar radiation purposes, as autoregressive models (AR) and autoregressive moving average (ARMA) [4, 5], autoregressive-integrated moving average (ARIMA), or several machine learning techniques such as neural networks, support vector machines, or Gaussian process [6, 23, 26].
- For time horizons over 1 day ahead up to 15 days, numerical weather predictions (NWP) models estimate atmosphere conditions and give different meteorological variables as solar radiation. These models are based on physical models using differential equations and solved with numerical methods, see Sect. 1.3.

NWP models accuracy vary depending on the temporal resolution and the geographical area. Different works are presented in bibliography showing almost no deviation for clear sky days [17] and errors around 30–40% for different stations between Europe, U.S.A., and Canada [33–35]. NWP data have also been used in recent years for post-processing forecasting results with hourly ground measurements from 6 h ahead onwards [12]. On the other hand, satellite images could also provide information about cloud variability using cloud motion vectors and improve hourly forecasting [16, 33].

This chapter is focused on solar radiation forecasting for global horizontal irradiance up to 6 h ahead. The statistical models provide good forecasting results for short-time horizons with different temporal granularities (from 5 min to hourly data). Statistical models find a relation between input data and the desired forecast solar radiation data. Many references estimate this relation using past ground solar radiation data for the same time series as inputs. However, in recent years several works have pointed out the improvement obtained combining ground measurement data with exogenous data as inputs [11, 31, 48]. This chapter is intended to provide a procedure to use statistical models for solar radiation forecasting using ground measurements and exogenous data, such as NWP and satellite data. An automatic methodology is proposed for the selection of satellite pixels using Pearson's correlation values.

2 Ground Solar Radiation Data

As statistical models explained in this chapter are based on ground solar radiation measurements, it is important to establish a good quality series of data. Indeed, before applying forecasting models, a solar radiation data assessment and quality check procedure must be used, see Sect. 1.1.2.4.

GHI data series is not considered stationary because they are affected by several variabilities. One variability is completely predictable and it is caused by the annual and daily solar cycle. On the other hand, motion of clouds and atmospheric parameters such as aerosols or water vapor caused a nonpredictable variability. All statistical models suggested in this chapter work with stationary time series of data, so autocorrelation should be constant over the time [9].

To work with statistical models, separating solar geometry dependence from the nondeterministic influences generated by atmospheric phenomena is considered appropriate [13]. So, two different new variables have been introduced to get transformed solar radiation temporal series in stationary series, clearness index k, and clear sky index K_t^*.

Clearness index is calculated dividing the global solar horizontal radiation GHI from measurement data by exoatmospheric horizontal radiation GHI_0 in the same point, see Eq. 1. This index removes deterministic variability caused by solar cycle because exoatmospheric radiation is based on solar angles.

$$K_t^* = \frac{GHI}{GHI_0} \tag{1}$$

GHI_0 is calculated for every day of the year over an horizontal surface with a simple expression using slight variations of distance between the Sun and Earth.

$$GHI_0 = I_0 \varepsilon_0 \cos(\theta_{zs}) \tag{2}$$

$$\begin{aligned}\varepsilon_0 = 1.00011 + 0.034221 \cos\tau + 0.001280 \sin\tau + \\ +0.000719 \cos 2\tau + 0.000077 \sin 2\tau\end{aligned} \tag{3}$$

$$\tau = \frac{2\pi(n-1)}{365} \tag{4}$$

where I_0 represents solar energy received from sun in a specific surface outside of the atmosphere per unit of time. The solar constant is considered normally as $I_0 = 1367\ W/m2$. While ε_0 is the variation of the distance between the Sun and Earth over the year calculated with Eq. 3, and $cos\theta_{zs}$ is zenith angle. Finally, zenith angle equation is substituted in main equation, Eq. 5.

$$GHI_0 = I_0\varepsilon_0(\sin\delta\sin\Phi + \cos\delta\cos\Phi\cos\omega) \tag{5}$$

The second variable introduced allows us to remove seasonal and atmospheric variability from solar radiation data series. This index is called **Clear sky index** K_t^* and is widely used in the bibliography for facilitating the learning process of statistical methods. As the clear sky index includes a clear sky model, Eq. 6, some atmospheric conditions are included in this calculation and we obtain a stationary data series.

$$K_t^* = \frac{GHI}{GHI_c} \tag{6}$$

Clear sky models estimate solar radiation GHI_c in a surface taken into account a day without any clouds. Most of the clear sky models are based on different climatic variables that represent the conditions of the atmosphere in clear times, such as aerosol optical depths (AODs), water vapor, ozone, Linke turbidity factor, or pressure. AODs represent solar radiation attenuation for different wavelengths from the scattering and absorption of sunlight within an atmospheric column. AODs and water vapor could be obtained from AERONET measurement stations net [19], while ozone could be retrieved from World Ozone Monitoring Mapping provided by the Canadian Government [8]. MACC project also provides AODs, water vapor, and ozone data for the whole world from 2004, available in [43].

These kinds of models have been tested all over the world and good results were obtained compared with ground measurement for clear sky times [39, 47]. One of the most common clear sky models in solar energy community is Bird and Hulstrm model [2]. This model is easy to implement and use water vapor column in *cm*, two aerosol optical depths, for 380 nm and 500 nm respectively, and total ozone column for the point we are estimating clear sky radiation. Based on these data, Bird model estimates different variables, such as Rayleigh dispersion, absorption of ozone, oxygen, carbon dioxide, and water vapor or absorption and dispersion of aerosols.

Another example widely used in the solar energy field is a method based on the REST2 model [15]. First version of REST, developed by Gueymard, only estimated beam component of solar radiation for clear sky. Later, Gueymard developed REST2 as a dual-band model based on the CPCR2 model. REST2 includes spectral distribution of extraterrestrial radiation, solar constant, water vapor, Angstrom turbidity coefficient, and reduced NO_2 and ozone column as inputs. In [43], is also available data series for global horizontal irradiance (GHI), direct normal irradiance (DNI) and diffuse horizontal irradiance (DHI) calculated with McClear clear sky model [27]. McClear data are available from 2004 to current day $d-2$ with minute, hourly, daily, or monthly time step for whole world and with a spatial resolution of 1.125. This model is based on look-up tables and radiative transfer model libRadtran using atmospheric composition variables provided by the MACC projects over whole world, such as AOD at 550 and 1240 nm, water vapor, and ozone column.

It is important to evaluate the accuracy of the clear sky model comparing with the measurement data for clear sky days. To evaluate these methods, it is necessary first to find out cloud-free solar radiation times (clear sky). Several methods could be found in the solar energy field to separate clear sky conditions from cloudy sky. Ineichen method detects clear sky hours by establishing a relation between global, beam, and diffuse, studying the stability of clearness index and the broadband aerosol optical depth [21]. In a similar way, Lefevre et al. [27] employ clearness index, corrected clearness index, direct normal clearness index, and diffuse fraction to detect clear sky instants. The method is described for 1 min. data and an adaptation for hourly data is used for Eissa to validate HelioClim-3 database in Egypt [14]. On the other hand, to detect individual times or period of times with clear or cloudy sky conditions only using GHI, Reno, and Hansen, [38] uses a moving window of period of times with 1 min. data series. This methodology detects clear and cloudy sky if data series meets certain conditions based on maximum value of GHI, mean value of GHI, and three different parameters to study the variability of each period. If the period studied in this window meets all the conditions, this period is considered clear sky weather. The limits of each condition should be established experimentally with ground measurement data in each location. Another methodology is also proposed to separate clear and cloudy sky conditions for GHI in periods of time [36, 37]. The model compares hourly data from ground measurement stations and clear sky model to detect whole clear sky days. For each day, the correlation coefficients matrix between ground data and clear sky data estimated by the model is calculated, Eq. 7. The determinant of this matrix should be lower than a threshold established experimentally once the data have been observed.

$$C = \begin{bmatrix} \rho_{GHI,GHI} & \rho_{GHI,GHI_c} \\ \rho_{GHI_c,GHI} & \rho_{GHI_c,GHI_c} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{GHI,GHI_c} \\ \rho_{GHI_c,GHI} & 1 \end{bmatrix} \tag{7}$$

$$\rho_{GHI_c,GHI} = \frac{Cov(GHI_c, GHI)}{\sigma_{GHI_c}\sigma_{GHI}} \tag{8}$$

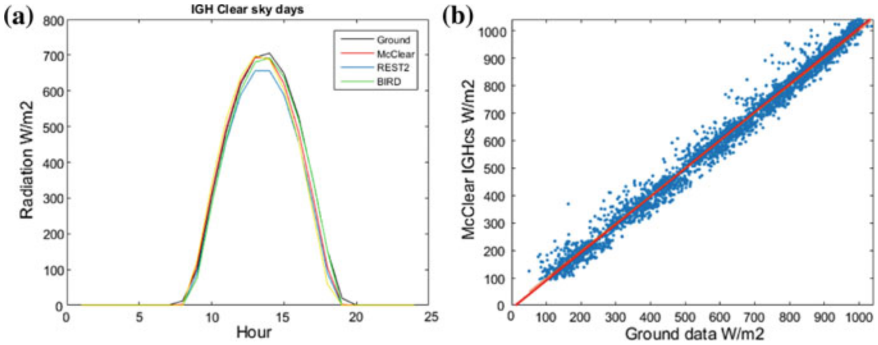


Fig. 1 Hourly GHI estimated with **a** different clear sky models and compared with ground measurement and **b** estimated with McClear model compared with ground measurement in a location in Canary Island, Spain

In Fig. 1a several clear sky models were tested for a single station in Canary Islands, Spain. All clear sky models reproduce very good results in terms of % rRMSE, variating from 4% with McClear model to 8% Bird model. While in Fig. 1b, GHI estimated with McClear compared with ground measurement data obtained in the same location is shown. McClear reproduces accurate results comparing hourly data.

3 Numerical Weather Prediction Model Data

NWP provided several atmospheric variables forecasting up to 15 days ahead. All these models are operated by 15 different meteorological agencies around the world. For global purposes, we can find the Global Forecast System (GFS) used by the US National Oceanic and Atmospheric Administration (NOAA) and Integrated Forecast System (IFS) operated by European Centre for Medium-Range Weather Forecast (EDMWF). On the other hand, some mesoscale models are available only for some zones around the world but offer better spatial resolution. In this case, we can find MM5 developed by Pennsylvania State University and National Centre for Atmospheric Research (NCAR) or WRF model. Accuracy of these different models change depending on the temporal scale and geographic area, as explained in Sect. 1.

Recently, several works have been published associating NWP models predicted data with a post-processing method to improve hourly ground solar radiation forecasting for time horizons hours ahead. Some other references establish a forecasting improving using NWP models data as inputs in different statistical methods.

In this chapter, it is explained the methodology for working on the second manner. NWP models data predicted for the next day are used as inputs in statistical models to improve solar forecasting. In this case, the methodology is described using the European Centre for Medium-Range Weather Forecast (ECMWF). ECMWF-provided data comes within 3 h intervals, so an interpolation of the value into hourly data was necessary. ECMWF provides information about several meteorological variables for different altitudes, however in this case, we only explained a methodology for using the following variables described by latitude, longitude, and time:

- Total Cloud Cover (TCC), with values between 0 and 1 using a cloud index.
- Surface Solar Radiation Downwards (SSRD), for accumulative values of J/m2 within two instants.

4 Satellite Solar Radiation Data

As proposed first with NWP data, satellite-derived data will be used to improve solar radiation forecasting accuracy with statistical models. The most important characteristic of satellite data is their great spatial resolution and possibility of introducing many information to see the evolution of the surroundings of the desired location. Indeed, the most important decision is the optimal selection of satellite pixels with the best information for the forecasting performance. The analysis of satellite data in a region surrounding the location where the solar radiation forecasting takes place is an important issue to establish an optimal selection.

4.1 Satellite Data Analysis

Satellite data offer solar radiation data with different spatial and temporary resolutions depending on the geographical area. These models provide GHI, DNI, and some great information about clouds and atmosphere conditions. Ineichen [20] provides an assessment study of several satellite-derived data for BSRN stations with hourly errors around 17% for global and 34% for direct normal irradiance. Anyway, depending on the location and climatic conditions, the uncertainties and deviation from ground measurement change significantly. Eissa [14] reports errors between 17 and 30% for different stations in Egypt, obtaining worst results for northern stations closer to the sea. Moreover, Mazorra [31] show errors with an average 12.2% rRMSE at C0-Pozo Izquierdo and 27.8% rRMSE at C1-Las Palmas, two stations in Gran Canaria island. The first station belongs to the southern area of the island with more occurrence of clear sky days, while the second station is situated in northern station with more cloudy days. Both works use satellite-derived hourly data from Helioclim3. On the other hand, Antonanzas [1] report around 4% rRMSE for a set of stations in Spain with yearly GHI data obtained from CMSAF database. In Gran Canaria island, for hourly data using CMSAF database with GHI an error was obtained from 15% in the south and 33% in the north, Fig. 2.

Calculate the error between satellite-derived data, both GHI or DNI, and ground measurement can show us the quality of the estimation. The more accuracy provided

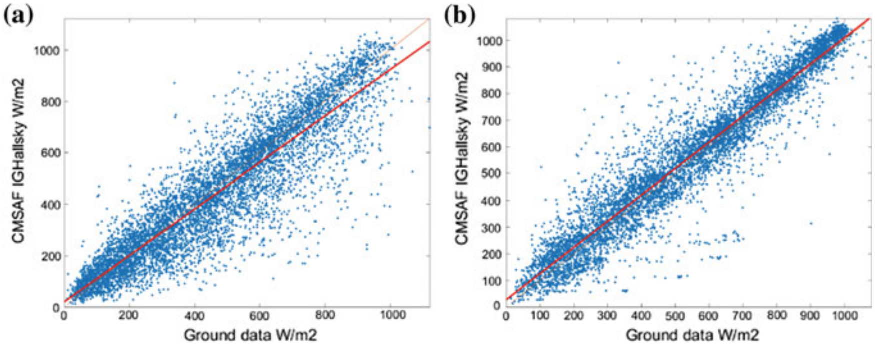


Fig. 2 CMSAF SIS hourly data comparison with ground data for northern station (**a**) and southern station (**b**) in Gran Canaria, Spain

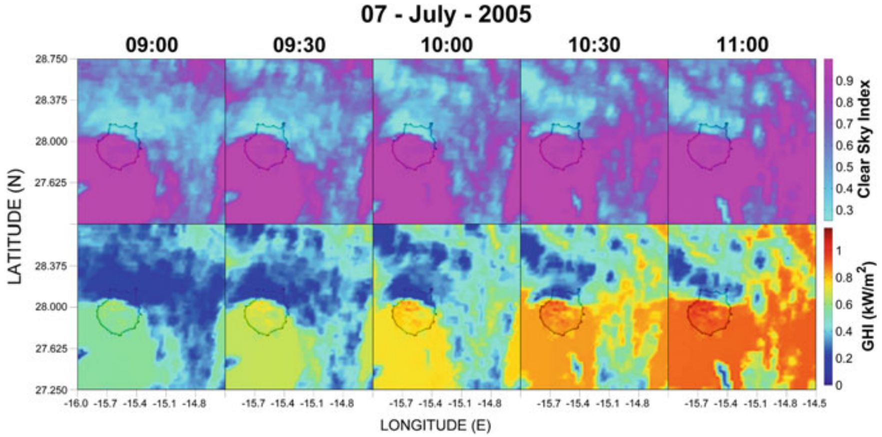
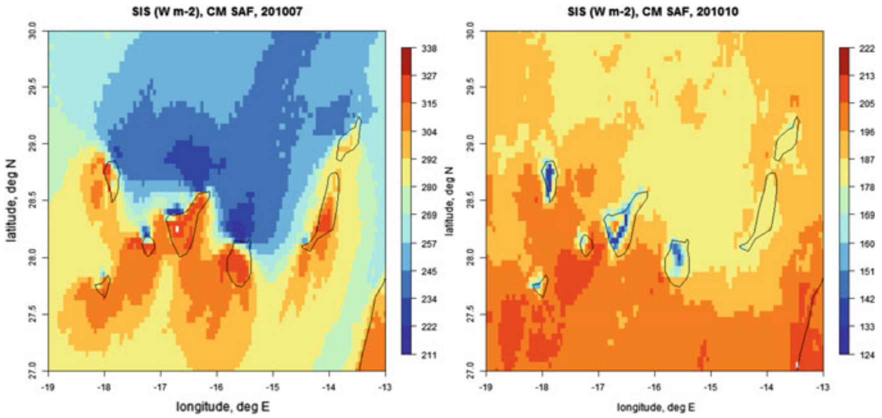


Fig. 3 Intra-day evolution 07-07-2005 of satellite-derived data each 30 min for Gran Canaria Island [31]

by the satellite derived, the more improvement in solar radiation forecasting will be obtained, instead of using only ground data. Moreover, an observation of satellite data all over the years could give an overview of the quality of the dataset. We can observe if satellite data represent specific climatic conditions in the location we are studying. Figure 3 shows that satellite data could estimate a particular behavior of known meteorological pattern in Canary Islands during summer. The northern part of the island presents an accumulation of clouds brought by the predominant trade winds. In this case, satellite-derived data used was obtained from the Helioclim-3 database version 5 (HC3v5). All this information has been processed by the Heliosat-2 method using images from the Meteosat geostationary [2, 3]. The selected area contains the entire island of Gran Canaria as well as a significant portion of sea at the north–east, motivated by the knowledge and influence of the trade winds in the Canary Islands. This area is defined, in decimal degrees, by the coordinates' latitude [+28.7500 to +27.2500], and longitude [−16.0000 to −14.5000], resulting in a grid of 61 × 55 pixels of information, where each pixel possesses a spatial resolution of 3 × 3 km^2.

In the same way, satellite-derived data obtained from the Satellite Application Facility on Climate Monitoring (CM SAF) showed the same good results for representing the cloud cover during summer. CMSAF information has been processed using images taken from the Meteosat Second-Generation (MSG) geostationary satellite network with SEVIRI sensor on board and NOAA polar satellites with AVHRR sensor [40]. These data are converted into global solar radiation and direct normal irradiance using Heliosat method and the Magic approach, validated with BSRN ground stations and provided in SARAH-2 database [41, 44]. The selected area contains the entire Canary Islands as well as a significant portion of sea. This area is defined, in decimal degrees, by the coordinates latitude [+27.0000 to +30.0000], and longitude [−19.0000 to −13.0000], where each pixel possesses a spatial resolution of 5 × 5 km^2 (Fig. 4).



(a) july

(b) october

Fig. 4 CMSAF SIS monthly means of gridded satellite data for Canary Islands, between July and October 2010

4.2 Spatiotemporal Correlation Analysis

Statistical models usually forecast GHI data using only ground clear sky data as inputs. As explained, the aim is to improve statistical models hourly forecasting using different satellite information as well as ground data. Satellite-gridded data includes a huge amount of pixels, so statistical model computation would be very difficult using the whole radiation data. In order to introduce in these models the most representative information, one of the most important decisions is to select the optimal pixels from the total set. The variable used to establish the best satellite pixel is the Pearson correlation between ground data of each station and satellite data of the selected area [11, 31, 48]. Pearson correlation provides information about the weather relationship and establish a useful tool to enhance a prediction.

The higher Pearson correlation factor between a satellite pixel and the soil ground data at the studied location is, the more information about the surroundings provide this pixel. Indeed, pixels chosen to improve further prediction are those with the higher correlation factor. As proposed in [31, 48], clear sky index is the variable used for studying the correlation factor. To evaluate correlations between both the parameters, satellite and ground data sets, in different temporal moments a time lag is established. This time lag provides information about the best closest reactions in the area and gives us an important overview between ground data at the present moment and solar radiation from the surroundings in the past and possible incoming events, Eq. 9. It is suggested a selection of four time lags, for hourly data the time lags go from the same temporal moment to a 3 h earlier maximum. From 3 to 6 h, intercorrelation between satellite and ground dataset obtain values below 0.5, so the relation is not considered relevant in the studied cases.

$$C_{K_t^*}(i,j)_h = corr(K^*_{t,ground}(t), K^*_{t,satellite}(t-h))\ para\ h = 0, 1, 2\ \&\ 3 \qquad (9)$$

In Figs. 5 and 6, it is shown an example for two stations in Gran Canaria Island (Spain). Figure 5 represents the correlation between each pixel in the whole grid for each time lag with a station in the south. Each image corresponds to a correlation calculation using the whole time-lagged satellite grid and the ground measurements at the present time. In the same way, Fig. 6 shows the correlation with a northern station. In both the cases the results provided by the calculation resemble with the expected behavior. The higher correlated pixels belong to the part of the island surrounding each station and the correlation decreases while time lag increases. All these observations lead to dividing islands into two different zones and allows a better comprehension of islands' microclimate. North part of the island is heavily influenced by clouds created by trade winds and the complicated orography. On

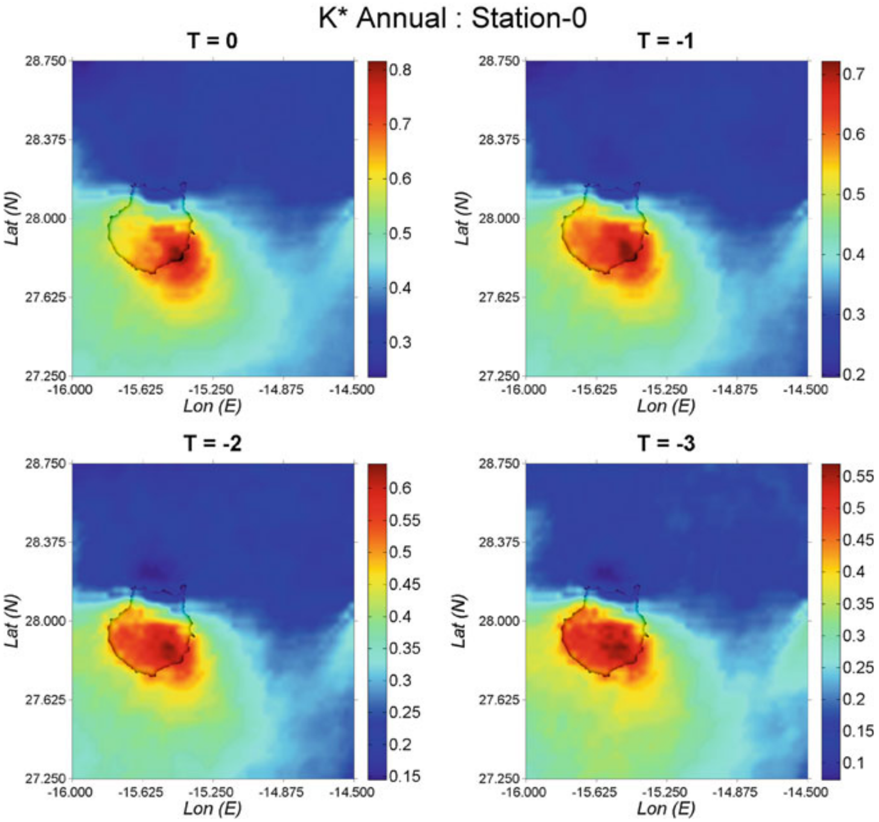


Fig. 5 Intercorrelation annual map for clear sky index between ground measurement and each satellite pixel around the measurements at station C0 Pozo Izquierdo for time lag h = 0, 1, 2, & 3 h [31]

the other hand, south sector remains protected from these clouds by the mountains. The empirical information about the island suggests that northern part possesses more cloudy days than south area and it is confirmed by the results obtained with Pearson's correlation. Indeed, correlated values estimated with satellite pixels and ground stations give us coherent information for selecting the best ones to improve solar forecasting.

Annual correlation is calculated using the whole year of each data set, both for satellite grid and ground measurement station. The results provide climatic conditions information from satellite grid radiation data in this area but it represents the statistical average on a whole year. Moreover, Zagouras et al. [48] propose a correlation using a temporal frame in order to evidence specific climatic conditions along the year. In this case, a correlation between satellite and ground data using different data sets for each meteorological quarters is shown, so the weather patterns where more coherent and accurate.

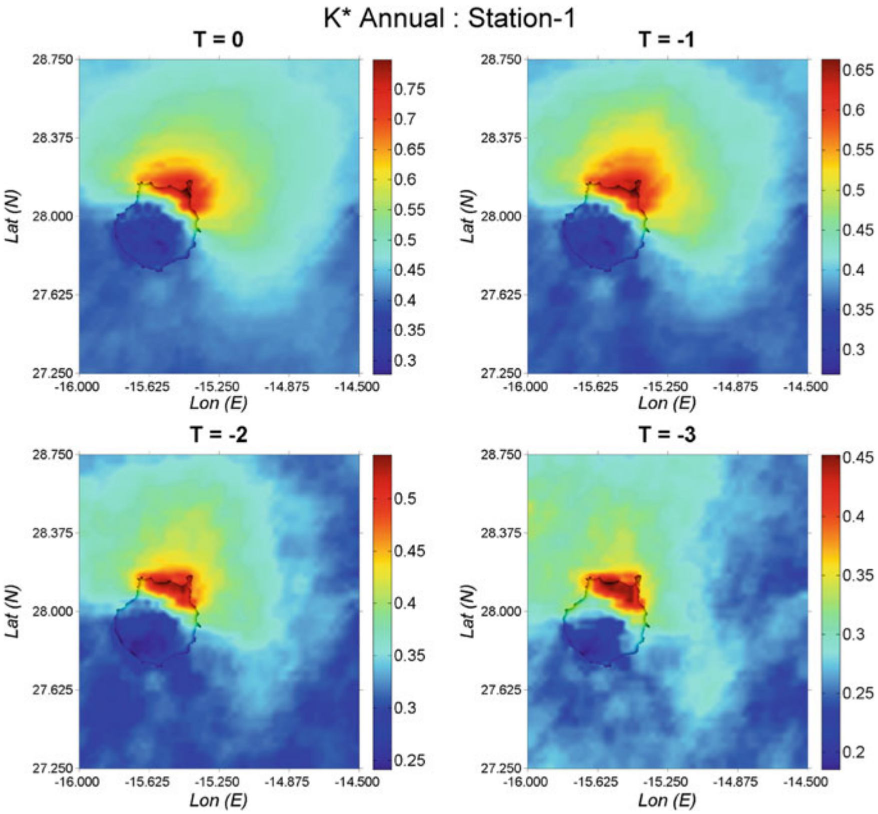


Fig. 6 Intercorrelation annual map for clear sky index between ground measurement and each satellite pixel around the measurements at station C1 Las Palmas for time lag h = 0, 1, 2, & 3 h [31]

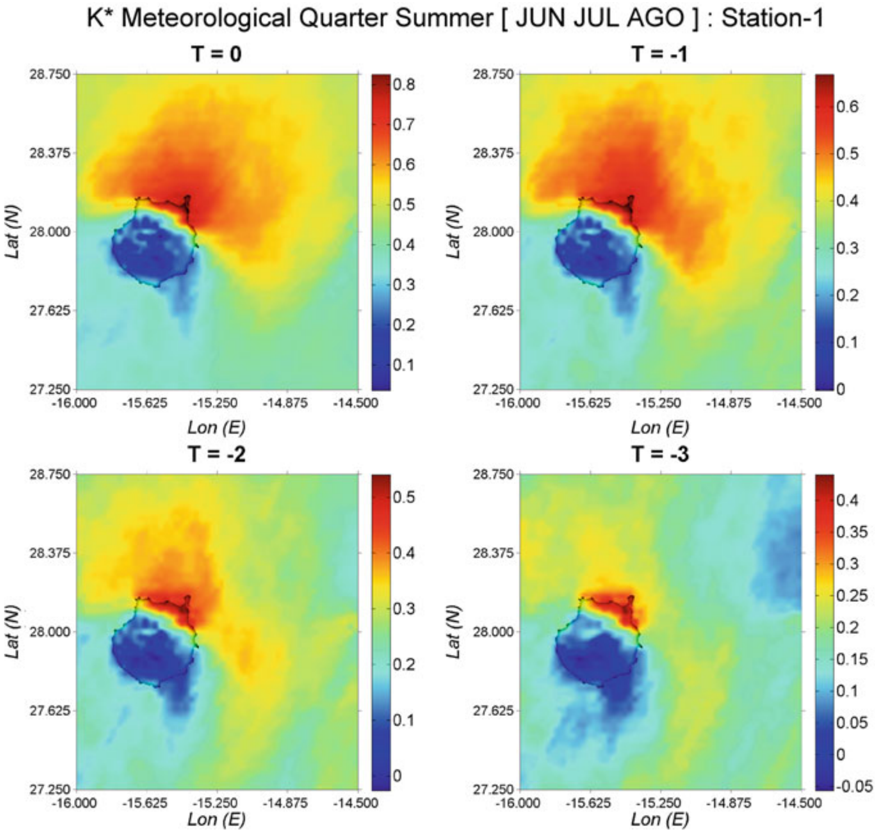


Fig. 7 Intercorrelation summer map for clear sky index between ground measurement and each satellite pixel around the measurements at station C1 Las Palmas for time lag h = 0, 1, 2, & 3 h [31]

For the example of northern station in Gran Canaria Island, Fig. 7 represents the correlation for summer data and Fig. 8 the correlation for autumn data. Both quarterly time-lagged correlation images offer climatic information consistency with empirical observations in all the seasons. In summer, due to the strong effect of trade winds, which creates a big area in the north of the island with a similar behavior. It is also important to remark the shelter provided but the orography, giving other climatic conditions to south of the Island. This shelter generates also a trail with similar climatic conditions on the sea, where we can also find clouds. In autumn and rest of the seasons, the Island is still divided into two regions but the higher correlations value correspond to pixels more concentrated around the ground station. In this season, the presence of winds is not so strong and there is not also influence in the surroundings.

The results obtained give us a similar behavior to the empirical climatic conditions in this region. Indeed, Pearson's correlation factor with time-lagged grid data give

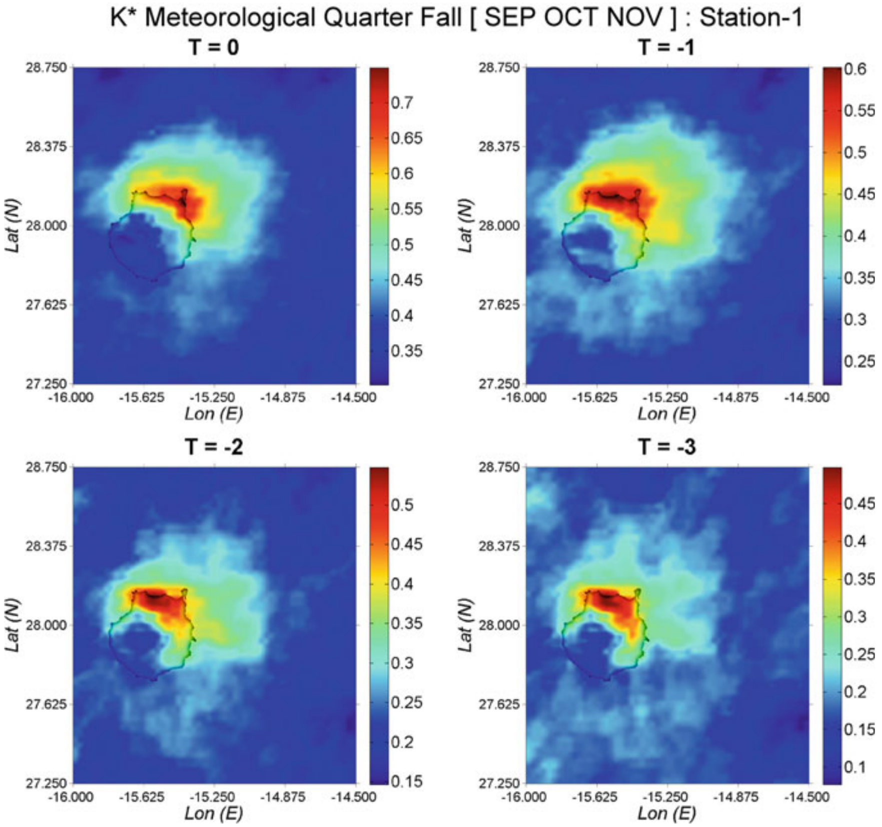


Fig. 8 Intercorrelation fal map for clear sky index between ground measurement and each satellite pixel around the measurements at station C1 Las Palmas for time lag h = 0,1, 2, & 3 h

us an important information for different time frames for selecting the most related pixels with the solar radiation we want to predict in the ground station. This technique will be used later to estimate the pixels to introduce in statistical models as inputs.

In the explained case, the time-lagged correlation was estimated using clear sky index data, both from satellite grid and ground data. In an attempt to get more additional data to enhance the prediction, Dambreville et al. [11] proposed a calculation using step by step clear sky index difference, Eqs. 10 and 11. This information should offer the direction of significant incoming patterns in the weather of the island, therefore, annual and quarterly analysis where made as in the previous case. These works were based on 15-min solar radiation data sets, instead of hourly data sets used by Mazorra et al. [31] and Zagouras et al. [48].

$$\Delta K_t^* = K_t^*(t+1) - K_t^*(t) \tag{10}$$

$$C_{\Delta K_t^*}(i,j)_h = corr(\Delta K^*_{t,ground}(t), \Delta K^*_{t,satellite}(t-h))\ para\ h = 0, 1, 2\ \&\ 3 \quad (11)$$

The results report a stretched area around the ground measurement station, indeed giving a more precise information about the most important pixels to improve the forecasting. However, these correlation values are not high, so the difference between selected and not selected pixels is not so relevant. The time-lagged images show a different behavior between east and west area (with higher correlated pixels in the west while lag increases), confirmed by the fact that wind mainly blows from west. Indeed, by selecting the most correlated pixels in each time lag we are giving information to the statistical models about incoming clouds.

5 Forecasting Statistical Models

The main purpose of this chapter is to explain a methodology to improve solar radiation forecasting for several hours ahead. As it was explained in Sect. 2, statistical models work with stationary data series. Hence, in this case the variable used in solar radiation forecasting models is the clear sky index. However, for calculating and discussing results and errors, the variable used is the global solar radiation GHI, estimated with Eq. 12.

$$G\hat{H}I = (\hat{K_t^*}) \cdot GHI_c \quad (12)$$

The general function used to connect input and outputs is Eq. 13.

$$\begin{aligned}\widehat{K_t^*}(t+h) = F[&K^*_{t,g}(t), \ldots, K^*_{t,g}(t-i), K^*_{t,e1}(t), \ldots \\ &K^*_{t,e1}(t-j), \ldots, K^*_{t,en}(t), \ldots, K^*_{t,en}(t-j)]\end{aligned} \quad (13)$$

where $\widehat{K_t^*}(t+h)$ is the clear sky index calculated for time horizon h, $K^*_{t,g}(t-i)$ is clear sky index from ground data set at the location for i past values and $K^*_{t,en}(t-j)$ corresponds to the n exogenous data with a j time lag. The number of exogenous data could vary depending on the selection of satellite pixels and NWP variables. One the most important decisions for the modeler is to choose the number of ground past data, number of satellite pixels, and NWP variables. Irrelevant inputs may unnecessarily increase model complexity and as a consequence may hamper the model performance. The general function F depends on the statistical model used and it is established during the training process. It is important to split measurement data set in training and testing sets. First one is used to establish the optimal function to relate input and output values, while testing set let us to calculate the accuracy of the model when new data are presented and controlled the overfitting.

5.1 Simple Forecasting Models

It is normal to compare any statistical model or methodology developed to improve solar radiation forecasting with a simple model. These simple models, naïve models, establish a reference forecasting limit that the new model should improve.

Two simple models for GHI hourly forecasting and different time horizons are suggested. Naïve models presented in this chapter only work with ground past values data, using clear sky index series. The first is the simple persistence (Pers) model [28], Eq. 14.

$$\widehat{K^*}(t+h) = K^*(t) \tag{14}$$

Persistence model is based on the assumption that atmospheric conditions remain invariant in two consecutive instants, indeed, that clear sky data for $t + h$ only depend on clear sky for the previous data. An easy improvement of this model is the smart persistence (smart pers). It consists of the forecast of the clear sky index for time horizon h using only the mean of h previous clear sky index with the ground data [18], Eq. 15.

$$\widehat{K^*}(t+h) = mean[K^*(t), ..., K^*(t-h)] \tag{15}$$

5.2 Linear Models

Statistical linear models have been widely developed for temporary series estimation. In this case, two different linear models are used for solar radiation forecasting using past ground data as inputs. The procedure explained in this chapter is based on linear models regression as described by Boland [4, 5] for solar radiation estimation in Australia using hourly and daily data.

- Autoregressive models (AR), a regression linear model based only on ground clear sky past data to forecast solar radiation for time horizon h, Eq. 16.

$$\widehat{K_t^*}(t+h) = \sum_{i=0}^{p-1} \left[\Phi_{i+1} K_t^*(t-i)\right] + \varepsilon_{t+h} \tag{16}$$

- Autoregressive moving average (ARMA), based on two linear models, an autoregressive model (AR) and a moving averages model (MA). This model estimates solar radiation forecast using a linear combination of different numbers of past data and error, Eq. 17.

$$\widehat{K_t^*}(t+h) = \sum_{i=0}^{p-1} \left[\Phi_{i+1} K_t^*(t-i)\right] + \varepsilon_{t+h} + \sum_{j=0}^{q-1} \left[\Theta_{j+1} \varepsilon_{t-j}\right] \tag{17}$$

In both the equations, $\widehat{K_t^*}(t+h)$ represents the solar radiation forecasting for time horizon h in terms of clear sky index, ε is a white noise and $K_t^*(t-i)$ are the ground clear sky past data from the measurement station used as inputs in AR model. For the AR model, Φ_{i+1} for $i = 1, 2, ..., p$ displays the autoregresive parameters and established the relation between clear sky past ground data and output data. While, for the MA model Θ_{j+1} for $j = 1, 2, ..., q$ shows moving average parameters that accompany the errors in MA regression. Both kind of values, Φ and Θ, are obtained during the training process. The methodology used to obtain both parameters is least square regression resulting from comparison of the set of past data used as input and future data that you want to predict.

The order p for the AR model shows the number of past data used to predict. One of the most important decisions during the training process is the model complexity. In this case, the optimal order p is obtained by calculating the partial autocorrelation function (PACF) and the Bayesian information criterion (BIC). Indeed, the model is defined by AR(p) depending on the number of past inputs used to obtain the best forecast data. On the other hand, the optimal order q for the MA model is obtained by calculating the autocorrelation function (ACF) and define the number of errors used during the prediction. In case of ARMA model, the optimal p and q orders should be established during the training process and the model is defined as ARMA(p,q).

These models are widely used for solar radiation forecasting because of the flexibility for working with temporary series depending on model orders. In many cases, very good results are described using AR and ARMA models with low-order parameters [7], which means not a long number of past clear sky values as inputs.

5.3 Artificial Neural Networks

Machine learning techniques have been described as very useful models for solar radiation forecasting. The method of machine learning explained in this chapter is artificial neural networks (ANN) [3]. However, many other techniques, such as Gaussian process or support vector machines, have been described in many papers with very good results and the methodology would be similar. ANNs is a statistical model that establishes a relation between a group of inputs and outputs during a training process. The model is based in a group of units, called neurons, that generate an output and received inputs from a group of input data or from other units. The units are connected between them by an associated weight. Each unit, neuron, receives the sum of different variables affected by these connection weights and produces an output. The output is obtained using a nonlinear activation function of transfer function to limit its amplitude and the input sums. The activation function used in this case is the hyperbolic tangent function, Eq. 18.

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{18}$$

The neural network used in this case is the multilayer perceptron (MLP), as described in many engineering and forecasting applications. MLP consists of a group of input data, which makes up the input layer, connected by weights with at least one layer of neurons, called hidden layer, finally connected with the output layer neurons. The input layer is not neurons because it only contains the meteorological input data. On the other hand, neurons in hidden layer present a nonlinear transfer function (hyperbolic function), while the final output neuron uses a linear activation function. The output layer consists of a single neuron with the solar radiation (in terms of clear sky data) data for the time horizon we want to forecast, $\hat{K}_t^*(t+h)$. Each variable in the input layer is connected with each neuron in hidden layer by a first group of weights. All hidden layer outputs are also connected with the single output by a second group of weights, Eq. 19.

$$\hat{K}_t^*(t+h) = \sum_{j=1}^{H} \omega_{sj}^2 f_j [\sum_{i=0}^{T-1} (\omega_{ji}^1 K_t^*(t-i) + \omega_0^1] + \omega_0^2 \tag{19}$$

where $\hat{K}_t^*(t+h)$ is the forecast solar radiation for time horizon h, ω_{sj}^2 is the group of weights that connect the output of hidden layer neurons with general output, ω_{ji}^1 is the group of weights that connect each input i with each hidden unit j, $K_t^*(t-i)$ represents the inputs variables for the ANN and ω_0^1 and ω_0^2 are the biases for hidden and output layers. Input variables $K_t^*(t-i)$ could be only ground measurement past data or also other meteorological, satellite, or NWP data.

5.3.1 Backpropagation Training Process

Both groups of weights, ω^1 and ω^2, associated to each connection between input, hidden, and output layer are modified during the training process. The optimal group of weights is obtained by minimizing a cost function. The mean square error between the target forecast data $K_t^*(t+h)$ and the estimated data obtained with ANN is one the most common methods $\hat{K}_t^*(t+h)$, Eq. 20.

$$E(\omega) = \frac{1}{2} \sum_{i=1}^{N} [\hat{K}_i^*(t+h) - K_i^*(t+h)]^2 \tag{20}$$

The backpropagation algorithm is the optimizing method used to minimize the cost function. In this algorithm, first the ANN weight vectors are randomly initialized. During the training process, the weights ω_k are changed with each iteration by calculating a new group ω_{k+1} by minimizing $E(\omega)$ with a gradient descent process, Eq. 21. Where η is the learning parameter. Scaled conjugate gradient gives us an optimal solution to estimate gradient direction and learning parameter in each iteration. In this way, we get the optimal solution faster.

$$\omega_{k+1} = \omega_k - \eta \frac{\partial E}{\partial \omega_k} \tag{21}$$

5.3.2 Regularization Techniques

The network architecture is one of the most important issues to obtain the optimal accuracy of ANNs to approach continuous functions. If ANNs obtain very good results with the training data set and approximate the noise of the function, poor accuracy will be obtained when new data are presented. This problem is called overfitting. So, ANNs structure will determine the possibility of the function to be useful with a general data set. It is widely described the use of regularization techniques to avoid overfitting problem [3, 25]. This complexity control has been treated with different regularization techniques, as pruning methods [24], regularization coefficients, or Bayesian regularization framework [30]. Classical regularization techniques need to estimate the regularization coefficients using a cross-validation method. Control model complexity reduces the computational load and find inputs without any influence to improve forecasting, because their associated weights are pruned.

In this case, the number of hidden units and inputs are decided by using Bayesian regularization framework. This method controls the complexity of the model. Bayesian framework considers a probability density function over the weight space. Indeed, the optimal group of ANNs weight values agree to the maximum probability density function. In practice, Bayesian framework [29, 30] introduces two hyperparameters, α and β, to the cost function in order to control the model complexity, Eq. 22. Term E_ω in the cost function induces a decay in unnecessary weights, so at the end of training process it is possible to prune weights under a certain value. Bayesian framework permits to estimate hyperparameters at the same time that we are training our network.

$$S(\omega) = \frac{\beta}{2} E_D + \frac{\alpha}{2} E_\omega \tag{22}$$

$$E_\omega(\omega) = \frac{1}{2} \sum_{j=1}^{m} (\omega_j^2) \tag{23}$$

where m is the number of parameters of the whole ANN structure. Bayesian framework permits to estimate hyperparameters at the same time that we are training our network. So, not only overfitting is controlled but also it is studied the complexity of the model to reduce hidden and input units. As described in [25], Bayesian framework approach uses an iterative procedure to estimate hyperparameter's optimal values, α and β, and optimal group of weights ω_{MP}. This iterative procedure takes place only in the training dataset.

1. Hyperparameters α y β are initialized using small values and vector of weights is randomly set using a Gaussian distribution. In this iteration number k, estimated

weights ω^k and defined hyperparameters α^k and β^k give us first an ANN output and calculate the error function $S^k(\omega)$, Eq. 22.

2. The optimal vector of weights ω_{MP}^{k+1} is obtained in this step using an optimization algorithm, as scaled conjugate gradient. The number of iterations in this step depends on the convergence criterion decided for backpropagation process. With this optimal weight, we estimate cost function for iteration $k+1$, E_ω^{k+1} y E_D^{k+1}.
3. In this step hyperparameters, α^{k+1} and β^{k+1}, are recalculated using the following steps:

   a. $\gamma^{k+1} = \sum_{p=1}^{m} \left(\frac{\lambda_p}{\lambda_p + \alpha^k} \right)$, where λ_p are eigenvalues of error Hessian matrix without regularization term, $H = \beta^k \nabla\nabla E_D$.

   b. $\alpha^{k+1} = \frac{\gamma^{k+1}}{2E_\omega^{k+1}}$.

   c. $\beta^{k+1} = \frac{N\gamma^{k+1}}{2E_D^{k+1}}$.

4. Repeat step 2 using new parameters ω^{k+1}, α^{k+1}, and β^{k+1}, calculated in the previous step until reaching the convergence criterion.

These steps are repeated until the regularized error is equal to half of the number of data points. The theory states that $S(w) = N/2$ when $\alpha = \alpha_{MP}$ and $\beta = \beta_{MP}$. It is also possible to study the hyperparameters α and β and parameter γ in each iteration k and decide the convergence when they are almost constant.

Bayesian framework gives also the possibility to study the number of inputs and hidden units, model complexity. To study the number of inputs, we study the weights associated to each input to decide the influence in the final result. We can divide weights into different sets, one group for weights associated to each input, one for second layer of weights (connect hidden units with output), and one for each layer biases. Each group is controlled for an independent hyperparameter α_g. This technique is called automatic relevance determination (ARD). As different hyperparameters are assigned to each group of weights, during the training process it is possible to determine the most relevant inputs. Weights associated to a large α_g are supposed to be small. In this case, input related to this weight and hyperparameter is not relevant for network results and can be eliminated.

To control the number of hidden units, Bayesian framework estimates the probability for each model, called evidence of the model. Different ANNs are trained using several numbers of hidden units and the network with the highest evidence provides us the best one [24, 30, 32]. To calculate evidence of each model the final expression is Eq. 24 that calculate the log of evidence. Where N is the number of inputs, m is the total number of parameters, γ is the number of well-determined parameters (weights not close to zero), and $|A|$ is the determinant of the Hessian matrix of the total (regularized) error function $S(w)$.

$$logP(M_i|D) = -\alpha_{MP}E_{\omega}^{MP} - \beta_{MP}E_{D}^{MP} - \tfrac{1}{2}log|A| + \tfrac{m}{2}log\alpha_{MP} + \\ +\tfrac{N}{2}log\beta_{MP} + \tfrac{1}{2}log\left(\tfrac{2}{\gamma}\right) + \tfrac{1}{2}log\left(\tfrac{2}{N-\gamma}\right) \quad (24)$$

6 Numerical Statistical Models Implementation

Once the theoretical approach of statistical models has been explained, in this section it is described the implementation using different data sets (ground measurement, satellite-derived, and NWPs data). In case of linear models and ANNs, one of the most important decisions is the model complexity. Following sections explain how to work with both statistical models in order to choose the optimal model complexity and number of inputs. As it is necessary to split data sets into training and testing set, both groups should represent the same climatic conditions and seasonal events to work with similar relations between input and output data (for example, one whole year for each set).

6.1 Linear Models Complexity and Results

The model complexity is one of the most important issues to take into account by the modeler. The complexity of a linear model consists of the number of inputs for the AR model and the number of error terms for the MA model. This complexity is settled by estimating the order q and p of the model. If the model uses a great number of unnecessary parameters, the general accuracy could be worse. As explained in Sect. 5, to study the model complexity we use the sample of partial autocorrelation function (PACF), the sample of the autocorrelation function (ACF) and the Bayesian information criterion (BIC).

Partial autocorrelation function (PACF) sets the correlation between two instants of time series with a ρ delay. The sample PACF for the different time lags gave us the number of past values relevant for the forecasting. The maximum order p of the model is established within a range of 95% of this sample, Fig. 9.

Following the same criterion, the maximum q order is selected using the sample autocorrelation function (SACF). Once the maximum orders have been decided, for AR models we calculated several simulations using order p from 1 to maximum for all time horizons in order to select the optimal number of parameters. In case of ARMA model, we calculate different situations using all possible combinations with order p and q from 1 to maximum.

Finally, to decide the best option between all simulations the Bayesian information criterion (BIC) and the error of the model %rRMSE (with testing data set) give us the optimal solution. For each time horizon AR model, optimal solution is obtained with different p orders.

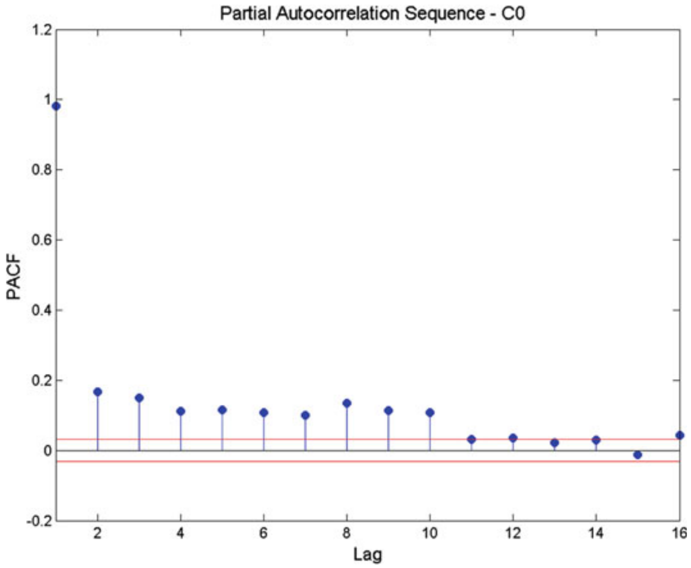


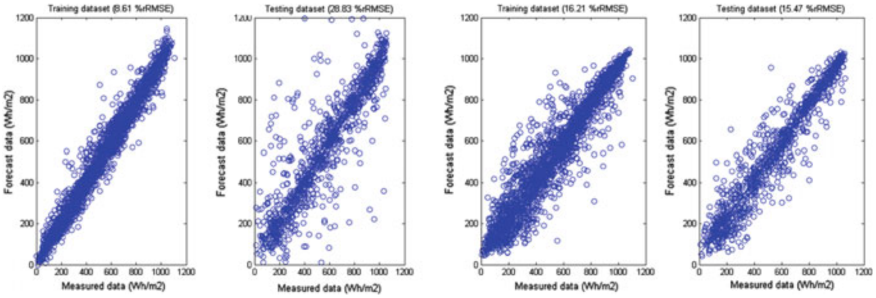
Fig. 9 Sample partial autocorrelation factor (SPACF) using solar radiation clear sky index. In this case, the maximum order was selected in $p = 12$

In the same way, ARMA model optimal solution was established calculating BIC and %rRMSE for all scenarios. In most cases, optimal model shown by BIC gives us different results of p and q orders. However, in many cases when compared with the optimal solution obtained with BIC to a simple ARMA model using $p = 2$ and $q = 1$, there is not a substantial improvement in terms of error %rRMSE. The optimal solution could need different orders for each time horizons and a huge number of input data (i.e., p order around 11), while ARMA(2,1) is a very simple model using only past input data to obtain a solar radiation forecasting.

6.2 ANNs Optimal Selection Using Ground Data

ANNs complexity in one of the most important issues to obtain the optimal forecasting accuracy. As explained in Sect. 5.3.2 we focus in selecting the number of inputs and hidden units. Bayesian framework gives us the possibility of selecting the number of inputs with ARD technique and the number of hidden units calculating the log of evidence.

Moreover, Bayesian framework controls the overfitting of the model [24]. Figure 10 shows the final result obtained with training and testing datasets forecasting using classical NN (a) and bayesian NN (b). In the first case, it is possible to observe a major dispersion in testing set because the model has overfitted the



(a) Classical ANN

(b) Bayesian NN

Fig. 10 Measured data versus forecasted data for the training and testing (right) datasets using classical ANN (**a**) and Bayesian NN (**b**)

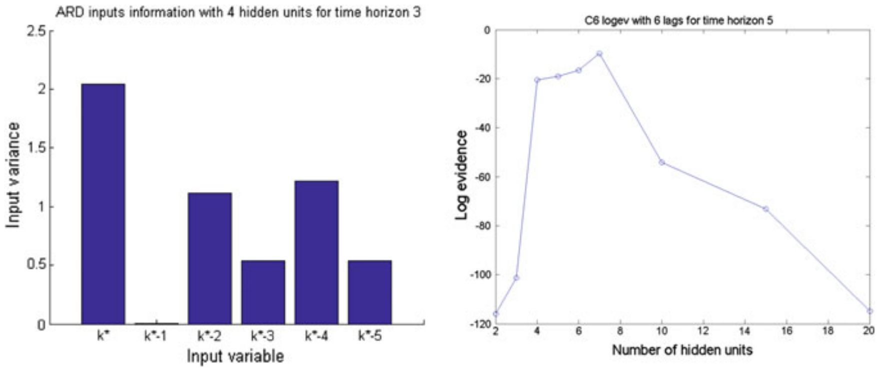
training set (overfitting problem). While with Bayesian NN the dispersion in training and testing sets remain almost similar, overfitting problem is not present.

In case of using only ground data, the number of inputs of each model corresponds to the number of past ground measurement using to forecast the solar radiation for time horizon h. ARD assigns a different hyperparameter α_g to each group of weights associated with one input. At the end of the training session, the weights with a large α_g are close to zero. In this case, the corresponding input is considered not relevant for the network and can be eliminated. In practice, each hyperparameter is represented in a figure with his variance. Inputs with a low bar comparde with other hyperparameters associated to the rest of inputs is considered irrelevant and could be eliminated. Figure 11a shows the result obtained with six past clear sky index inputs. Sometimes, pruned inputs are considered irrelevant with ARD technique, as the second input in Fig. 11, do not reproduce more accurate results and it is better to use all inputs. It is advisable to check the general error of the model when we prune these inputs.

The number of hidden units is settled once it is decided the number of inputs that give us the optimal results. Bayesian framework calculates the log of evidence between several ANNs with different number of hidden units, Eq. 24, to establish the optimal one. The ANNs with the higher Log of evidence is considered the best one. As in Fig. 11b, most of results show low number of units. As explained with ARD technique, log of evidence give us information about the best number of units but it is recommended to calculate the error of several models around the best one to establish the optimal number.

6.3 Exogenous Data Optimal Selection

The aim of using exogenous data is to improve ANN's hourly forecasting obtained only with ground data. Exogenous data used in this case are NWP's data and a grid of



(a) ARD (b) Log of Evidence

Fig. 11 ARD information for six ground data (**a**) and log of evidence for different hidden units (**b**)

satellite-derived data. These exogenous data will be added to the number of ground measurement inputs obtained in Sect. 6.2.

NWPs data suggested are total cloud cover (TCC) and surface solar radiation downwards (SSRD) obtained for the location of study. Both data are the one day ahead prediction for the hour we want to predict, $k_t^*(t+h)$, estimated by a NWP model.

Moreover, it is proposed to use also satellite-derived data in order to include information of the surroundings to the ANN. Satellite-gridded data includes a huge amount of pixels, so ANN computation would be very difficult using the whole radiation data. In order to introduce the most representative information obtained from satellite data, one of the most important decisions is to select optimal pixels from the total set. The variable used to establish the best satellite pixel is the Pearson correlation between satellite-gridded data and ground data, [11, 31, 48].

This Pearson correlation is calculated for each station between ground data at the present time and satellite pixel with time lags. During the training proces, we used time lags from $t = 0$ to a maximum of 3 h obtaining four time-lagged images. This correlation quantifies a relation between ground data and satellite for different time lags. After 3 h, the correlation between present ground data and past satellite data is not representative. Consequently, Pearson correlation gives us information about meteorological event incoming from the surroundings included in satellite images.

In that way, we can select pixels from the surroundings that represent the highest relation with ground station data for different time lags. ANN improves solar forecasting depending on the satellite information we use as inputs. The selection of optimal group of pixels is one the most important issues in this field. Dambreville et al. [11] proposes to use Pearson correlation of clear sky index variation between satellite and ground data with a time lag from 15 to 60 min. each 15 min. It is suggested to use a fixed number of pixels from each time-lagged image to improve solar forecasting with a linear statistical model. While Zagouras et al. [48] choose the 100 most correlated pixels from all time-lagged images. They work with hourly data and choose the best pixel using clear sky index Person correlation between satellite and ground data. Number of optimal pixels in each time- lagged image is settled with the genetic Algorithm.

Mazorra et al. [31] considered a maximum number of 30 satellite-derived radiation data. During the training process, six different tests at each station based on selecting different pixels were made. For the first test, the number of pixels over 0.5 correlation values were retrieved for each image and later the distribution of pixels between the four time-lagged images was computed. The considered 30 satellite pixels are the highest correlated for each image according to this percentage distribution (e.g., in Test-1 56% of pixels at time lag t = 0; 30% at $t = -1$; 8% at $t = -2$ and 6% at $t = -3$ should be selected, which means to retrieve 17, 9, 2, and 2 pixels respectively). As most of the optimal pixels were taken from first two time-lagged images, it was considered five tests using different percentage distributions of pixels. Test-2 and Test-3 estimated a new distribution taking into account more pixels from the other two images. While Test-4 only added satellite pixels only from time lags, $t = 0$ and $t = -1$. Finally, Test-5 formulates the same procedure as Test-1 but calculates a different distribution for every quarterly group of images for each station. Test-6 composes a new distribution using the best previous percentage distribution but selecting pixels from quarterly images.

The huge amount of satellite-derived data makes the computation difficult, so a median filter for each 3 × 3 satellite pixels is applied. Consequently, a superpixel was created computing GHI median value of every 3 × 3 group of pixels, Fig. 12.

To improve the previous work that use a different distribution of pixels for each test, it is suggested an automatic methodology. The estimation of the optimal number of pixel is based on the same Pearson correlation calculation. Instead of selecting a fix limit correlation value (0.5) to generate the distribution, it was considered different tests changing this limit. The percentile of the whole Pearson correlation distribution for all time-lagged images was established as the limit. In that way, it is possible to change the distribution of pixel from the four images depending on the percentile considered. During the training process, different percentiles from 0.1 to 0.9 were

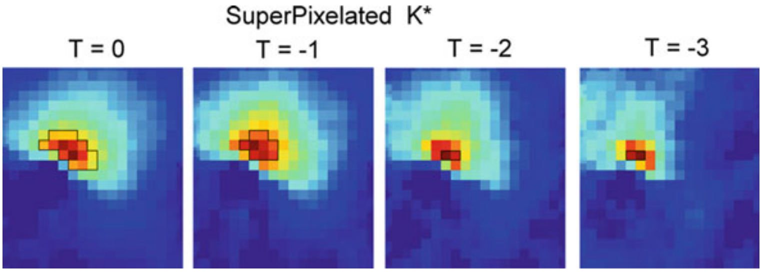


Fig. 12 Superpixel (3 × 3) selection at station in Gran Canaria (Spain) for time-lagged correlation images, t = 0, 1, 2, & 3 h. Black area shows selected superpixels

suggested. For the first one, more pixels from first images (time lag 0 and 1 h) are selected, while for the second one more pixels from latest images were extracted.

Once selected a different test, ANNs are trained using ground past data and satellite data for each test. Comparing forecasted hourly GHI with measured data for the testing dataset using the relative root mean square error, the best ANN's architecture and the optimal satellite information is selected for each case. In each location, this procedure should be repeated for the different time horizons' solar radiation forecasting.

6.4 Solar Radiation Forecasting Results

Statistical forecasting models explained in this chapter should be validated with measurement data. It is possible to find several error metrics suggested in specialized bibliography to establish the accuracy of each model. The error of the models are calculated with testing data set, because it is necessary to evaluate the capacity of each model to generalize the results with unknown data. All metrics are expressed in terms of GHI (W/m^2) even if clear sky model was the variable used during the training season. The most common error metrics are root mean square error (RMSE), mean absolute error (MAE), or mean bias error (MBE) and their relative metrics calculated dividing by the generally measured mean for the testing data set. It is also widespread the use of SKILL metric. This metric calculates the difference of each model with a simple model used as a reference. In this case, it is explained SKILL error metric compared with persistence model. Indeed, this value gives us how the described model improves a simple persistence model. This chapter shows some examples obtained using two measurement stations in Gran Canaria (Spain), calculating the accuracy of each model using %rRMSE, Eq. 25, and SKILL, Eq. 26. The models in terms of these error metrics are the following:

- Persistence Model—*Pers*
- Smart Persistence Model—*Smart Pers*
- Autoregressive Moving average with orders (2,1)—*ARMA(2,1)*
- Artificial Neural Networks with ground data—*NN*
- Artificial Neural Networks with ground data, satellite data and NWP data—*NN+ECMWF+SAT* Model

$$RMSE_{modelo} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{GHI}_{g,i} - GHI_{measure,i})^2} \tag{25}$$

$$SKILL(\%) = \left(1 - \frac{RMSE_{modelo}}{RMSE_{persistence}}\right)x100 \tag{26}$$

Table 1 RMSE for time horizons $h = 1 ... 6$ in two stations in Gran Canaria (Spain)

Stations	Models	1 h	2 h	3 h	4 h	5 h	6 h
C0	Persistence	92.47	128.04	149.88	168.08	176.17	177.26
	Smart persistence	92.47	124.64	140.10	144.44	141.50	138.69
	ARMA(2,1)	85.78	109.30	119.77	126.64	129.32	130.38
	NN	88.20	113.39	125.13	127.12	131.68	130.03
	NN+ECMWF+SAT media IGH = 543.10 Wm^{-2}	84.00	106.17	110.51	114.93	118.89	120.43
C1	Persistence	118.95	167.03	195.15	213.39	224.71	228.18
	Smart persistence	118.95	169.11	190.69	195.34	190.21	182.18
	ARMA(2,1)	111.44	145.14	159.90	167.17	170.65	171.49
	NN	110.63	143.90	157.06	162.11	162.09	162.88
	NN+ECMWF+SAT media IGH = 433.79 Wm^{-2}	104.75	134.37	142.82	145.41	147.31	147.88

Table 1 shows the results in terms of RMSE in (W/m^2) and the ground measurement mean for the testing dataset. While, Fig. 13 describes the results in terms of %rRMSE. Both of them give the results for time horizons between 1 and 6 h ahead and for two ground measurement stations. First station (C0) is located to the south of the island and presents better results because the weather is more stable along the year with more presence of clear sky day. On the other hand, C1 station is on the north of the island a presents more cloudy and unstable days during the year, so error metrics are worse. As it is obvious, all the models obtain better results for shorter time horizons and get worse results while increase time horizon. In case of persistence simple models this growth is much more pronounced, while ARMA and NN with or without exogenous data control the error for large time horizons. Even if smart persistence presents an improvement compared to persistence, in larger time horizons it presents still some problems. ARMA(2,1) and NN only use ground measurement past data as inputs get similar results in terms of RMSE for both the stations and time horizons. Both the models improve significantly as simple models. Moreover, the inclusion of exogenous data in NN as inputs improve also the model and obtain the best accuracy for both stations and time horizons.

In Fig. 14 it is possible to see the SKILL(%) parameter for both the stations and all time horizons. In this figure, it is shown the different combinations of NN inputs in order to discuss the importance of everyone: NN only with ground data, NN with ground and satellite, NN with ground and NWP data, and NN with ground and all exogenous data. The SKILL forecast increases with time horizon, which means that the more far ahead in time, the better results we get with ANN+ECMWF+SAT method compared with persistence model. For both the stations the best model is the neural networks with ground, satellite and NWP data as inputs. Moreover, it is also observable that satellite data (NN+SAT) give better results for the first three time horizons, from 1 to 3 h, while NWP data (NN+ECMWF) is the best model from time horizon 4 to 6 h. It could also be interesting the results separating testing data sets in

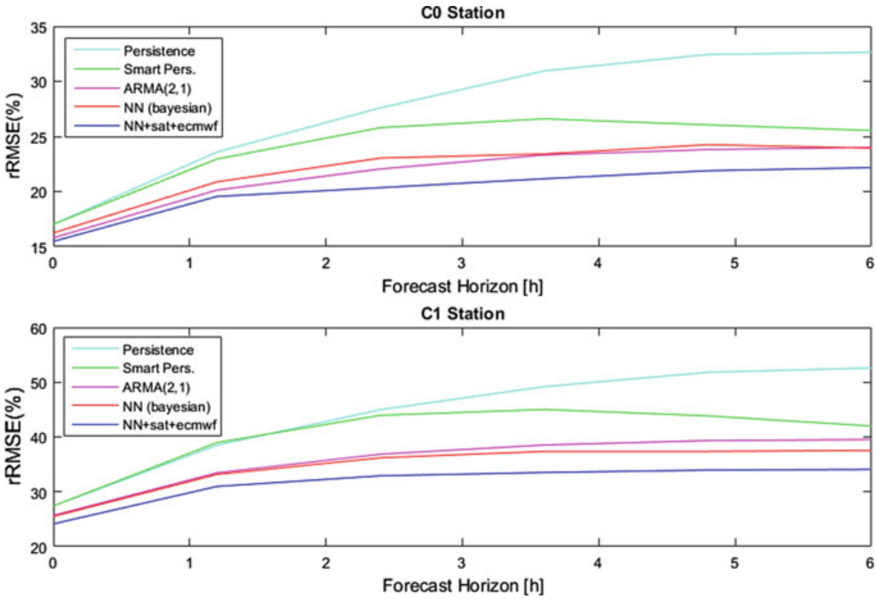
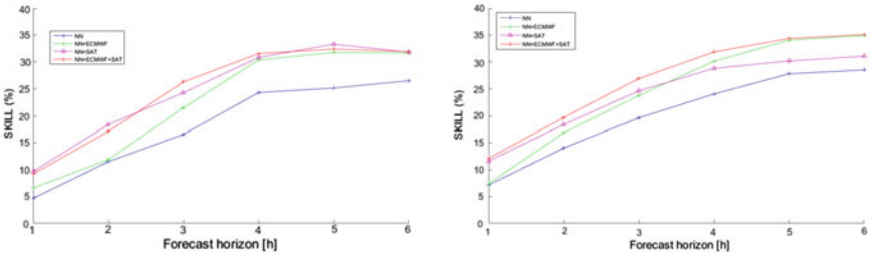


Fig. 13 %rRMSE results using testing data set for two different stations in Gran Canaria (Spain) with several forecasting models. C0 Station (up) & C1 Station (down)



(a) C0 Station

(b) C1 Station

Fig. 14 SKILL(%) results for two different stations in Gran Canaria (Spain) using exogenous data with ANNs

the different seasons of the year or type of days (i.e., cloudy or sunny days). In this way, it is possible to establish a different model depending on the weather conditions or the time of the year.

7 Conclusions

The main conclusion of this chapter is that ANN and ARMA model present very good results in solar radiation hourly forecasting compared with persistence simple models. On the other hand, when exogenous data, as satellite data and NWP data, are introduced to the ANN as inputs we obtain an important improvement. In this case, we used solar radiation from several pixels around the measurement station with time lagged from $t = 0$ h to $t = -3$ h compared to present time. NWP data used to improve solar radiation forecasting are 24 h ahead the prediction of total cloud cover and surface solar radiation forecasting for the time step we want to forecast. One of the most important decision in order to obtain more accurate results is to find the optimal satellite pixels. A huge number of pixels without relevant information for solar radiation forecasting causes a high computation cost with ANNs and worse estimation errors. Pearson's correlation between ground and satellite data give us critical information to select optimal satellite pixels. The architecture of neural networks influences the final result of the estimation. The Bayesian methods explained in this section are considered an adequate tool to estimate the number of inputs and hidden neurons. With this method, it is possible to avoid overfitting problem and obtain accurate prediction results with both training and testing data.

References

1. Antonanzas-Torres F, Cañizares F, Perpiñán O (2013) Comparative assessment of global irradiation from a satellite estimate model (cm saf) and on-ground measurements (siar): a spanish case study. Renew Sustain Energy Rev 21:248–261
2. Bird RE, Hulstrom RL (1981) Simplified clear sky model for direct and diffuse insolation on horizontal surfaces. Solar Energy Research Inst., Golden, CO (USA), Technical report
3. Bishop CM (1995) Neural networks for pattern recognition. Oxford university press
4. Boland J (1995) Time-series analysis of climatic variables. Sol Energy 55(5):377–388
5. Boland J (2008) Time series modelling of solar radiation. Springer
6. Bosch J, Lopez G, Batlles F (2008) Daily solar irradiation estimation over a mountainous area using artificial neural networks. Renew Energy 33(7):1622–1628
7. Box G, Jenkins G (1998) Time series analysis, forecasting and control. Wiley
8. Canada's E (2015) World ozone monitoring mapping. http://es-ee.tor.ec.gc.ca/e/ozone/ozoneworld.htm/
9. Chatfield C (2013) The analysis of time series: an introduction. CRC press
10. Chow CW, Urquhart B, Lave M, Dominguez A, Kleissl J, Shields J, Washom B (2011) Intra-hour forecasting with a total sky imager at the uc san diego solar energy testbed. Sol Energy 85(11):2881–2893
11. Dambreville R, Blanc P, Chanussot J, Boldo D (2014) Very short term forecasting of the global horizontal irradiance using a spatio-temporal autoregressive model. Renew Energy 72:291–300
12. Diagne M, David M, Boland J, Schmutz N, Lauret P (2014) Post-processing of solar irradiance forecasts from wrf model at reunion island. Sol Energy 105:99–108
13. Diagne M, David M, Lauret P, Boland J, Schmutz N (2013) Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. Renew Sustain Energy Rev 27:65–76

14. Eissa Y, Korany M, Aoun Y, Boraiy M, Abdel Wahab MM, Alfaro SC, Blanc P, El-Metwally M, Ghedira H, Hungershoefer K et al (2015) Validation of the surface downwelling solar irradiance estimates of the helioclim-3 database in egypt. Remote Sens 7(7):9269–9291
15. Gueymard CA (2008) Rest2: high-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation-validation with a benchmark dataset. Sol Energy 82(3):272–285
16. Hammer A, Heinemann D, Lorenz E, Lückehe B (1999) Short-term forecasting of solar radiation: a statistical approach using satellite data. Sol Energy 67(1):139–150
17. Heinemann D, Lorenz E, Girodo M (2006) Forecasting of solar radiation: solar energy resource management for electricity generation from local level to global scale. Nova Science Publishers, New York
18. Hoff TE, Perez R (2012) Modeling pv fleet output variability. Sol Energy 86(8):2177–2189
19. Holben BN, Eck T, Slutsker I, Tanre D, Buis J, Setzer A, Vermote E, Reagan J, Kaufman Y, Nakajima T et al (1998) Aeroneta federated instrument network and data archive for aerosol characterization. Remote Sens Environ 66(1):1–16
20. Ineichen P (2014) Long term satellite global, beam and diffuse irradiance validation. Energy Procedia 48:1586–1596
21. Ineichen P (2016) Validation of models that estimate the clear sky global and beam solar irradiance. Sol Energy 132:332–344
22. Kostylev V, Pavlovski A et al (2011) Solar power forecasting performance–towards industry standards. In: 1st International workshop on the integration of solar power into power systems Aarhus, Denmark
23. Lauret P, David M, Fock E, Bastide A, Riviere C (2006) Bayesian and sensitivity analysis approaches to modeling the direct solar irradiance. J Sol Energy Eng 128(3):394–405
24. Lauret P, Fock E, Mara TA (2006) A node pruning algorithm based on a fourier amplitude sensitivity test method. IEEE Trans Neural Netw 17(2):273–293
25. Lauret P, Fock E, Randrianarivony RN, Manicom-Ramsamy JF (2008) Bayesian neural network approach to short time load forecasting. Energy Convers Manag 49(5):1156–1166
26. Lauret P, Voyant C, Soubdhan T, David M, Poggi P (2015) A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. Sol Energy 112:446–457
27. Lefevre M, Oumbe A, Blanc P, Espinar B, Gschwind B, Qu Z, Wald L, Schroedter-Homscheidt M, Hoyer-Klick C, Arola A et al (2013) Mcclear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. Atmos Meas Tech 6(9):2403–2418
28. Lorenz E, Heinemann D (2012) Prediction of solar irradiance and photovoltaic power—Comprehensive Renewable Energy. Elsevier, Oxford, pp 239–292
29. MacKay DJ (1992) A practical bayesian framework for backpropagation networks. Neural Comput 4(3):448–472
30. MacKay DJ (2003) Information theory, inference and learning algorithms. Cambridge university press
31. Mazorra Aguiar L, Pereira B, David M, Daz F, Lauret P (2015) Use of satellite data to improve solar radiation forecasting with bayesian artificial neural networks. Solar Energy
32. Penny WD, Roberts SJ (1999) Bayesian neural networks for classification: how useful is the evidence framework? Neural Netw 12(6):877–892
33. Perez R, Kivalov S, Schlemmer J, Hemker K, Renné D, Hoff TE (2010) Validation of short and medium term operational solar radiation forecasts in the US. Sol Energy 84(12):2161–2172
34. Perez R, Lorenz E, Pelland S, Beauharnois M, Van Knowe G, Hemker K, Heinemann D, Remund J, Müller SC, Traunmüller W et al (2013) Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. Sol Energy 94:305–326
35. Perez R, Moore K, Wilcox S, Renné D, Zelenka A (2007) Forecasting solar radiation-preliminary evaluation of an approach based upon the national forecast database. Sol Energy 81(6):809–812
36. Petruccelli JD, Nandram B, Chen M (1999) Applied statistics for engineers and scientists. Prentice Hall New Jersey

37. Polo J, Zarzalejo LF, Salvador P, Ramírez L (2009) Angstrom turbidity and ozone column estimations from spectral solar irradiance in a semi-desertic environment in Spain. Sol Energy 83(2):257–263
38. Reno MJ, Hansen CW (2016) Identification of periods of clear sky irradiance in time series of ghi measurements. Renew Energy 90:520–531
39. Reno MJ, Hansen CW, Stein JS (2012) Global horizontal irradiance clear sky models: implementation and analysis. SANDIA report SAND2012-2389
40. Schulz J, Albert P, Behr HD, Caprion D, Deneke H, Dewitte S, Durr B, Fuchs P, Gratzki A, Hechler P et al (2009) Operational climate monitoring from space: the eumetsat satellite application facility on climate monitoring (cm-saf). Atmos Chem Phys 9(5):1687–1709
41. Scientist CS (2016) Annual product quality assessment report 2015. EUMETSAT Satellite Application Facility on Climate Monitoring
42. Sengupta M, Habte A, Kurtz S, Dobos A, Wilbert S, Lorenz E, Stoffel T, Renné D, Gueymard C, Myers D et al (2015) Best practices handbook for the collection and use of solar resource data for solar energy applications. NREL
43. Transvalor MP (2014) Soda solar radiation data http://www.soda-pro.com/. Accessed 2017
44. Trentmannn J, Huld T (2016) Meteosat-east solar surface irradiance data records. EUMETSAT Satellite Application Facility on Climate Monitoring
45. Urquhart B, Ghonima M, Nguyen D, Kurtz B, Chow C, Kleissl J (2013) Sky imaging systems for short-term forecasting. J Elsevier, Waltham, Massachusetts, Kleissl
46. Wittmann M, Breitkreuz H, Schroedter-Homscheidt M, Eck M (2008) Case studies on the use of solar irradiance forecast for optimized operation strategies of solar thermal power plants. IEEE J Sel Top Appl Earth Obs Remote Sens 1(1):18–27
47. Younes S, Muneer T (2007) Clear-sky classification procedures and models using a world-wide data-base. Appl Energy 84(6):623–645
48. Zagouras A, Pedro HT, Coimbra CF (2015) On the role of lagged exogenous variables and spatio-temporal correlations in improving the accuracy of solar forecasting methods. Renew Energy 78:203–218