Site adaptation of satellite-based GHI and DNI using clustered features and combined local-global modeling

Diego Rodrigues de Miranda[a*], Germán Ariel Salazar[b], Rodrigo Alonso-Suárez[c], Alexandre Costa[d], Renan Soares Siqueira Costa[e], Tsang Ing Ren[e], Olga de Castro Vilela[d]

[a] German Aerospace Center (DLR), Institute of Networked Energy Systems, Oldenburg, Germany; former affiliation: Center for Renewable Energy (CER), Federal University of Pernambuco (UFPE), Recife, Brazil
[b] Grupo de Estudio y Evaluación de la Radiación Solar (GEERS), CONICET, Universidad Nacional de Salta, Salta, Argentina
[c] Laboratorio de Energía Solar, Departamento de Física del Litoral, CENUR Litoral Norte, Universidad de la República, Salto, Uruguay
[d] Center for Renewable Energy (CER), Federal University of Pernambuco (UFPE), Recife, Brazil
[e] Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Recife, Brazil

* Corresponding author
E-mail address: diegorodriguesdemiranda@gmail.com (Diego Rodrigues de Miranda)

Abstract: Estimating solar irradiance over several years is essential for assessing large-scale photovoltaic and heliothermic projects. The usual way to achieve this is to adapt long-term satellite-based solar irradiance estimates for specific locations using short-term ground measurements. In this context, a novel site adaptation method is proposed for global horizontal irradiance (GHI) and direct normal irradiance (DNI) based on a two-level modelling strategy incorporating both clustered and unclustered features. These lead to global and local models, which are evaluated separately and in combination. Input variables are provided by the Copernicus Atmosphere Monitoring Service Radiation Service (CAMS) and ECMWF ERA5-Land. The method is tested at four locations – El Rosal and Salta (Argentina), Petrolina (Brazil), and Gobabeb (Namibia), addressing sites in the Southern Hemisphere with similar climates, but different latitudes and satellite viewing angles. A new non-supervised sky conditions classification method is proposed for the local models using the clear sky index, a variability index, and Kernel Density Estimation. This classification forms an integral part of the site-adaptation procedure by capturing local cloud-irradiance interactions. For GHI at 15-minute resolution, the method improves accuracy relative to CAMS at locations near the edge of the Meteosat Second Generation field of view satellites, with RMSE reductions of 26.2% (El Rosal) and 4.8% (Salta). For DNI in Petrolina and Gobabeb, the local and combination models outperform CAMS model, achieving RMSE reductions of 10.2% and 5.8%, respectively. Combination models present the more accurate results as they use the global, local, and CAMS outputs as input variables.

Keywords: site adaptation, CAMS Radiation Service, classification method, sky conditions, combination models

Variables

k_c	Clear sky index
k_t	Clearness index
k_n	Normal transmittance
VI	Variability index of Stein et al. (2012)
$\hat{y}_{corr\_QM}$	QM-corrected time series in the test set
$\hat{y}_{test}$	Modeled time series in the test set
$\hat{y}_{corr\_BD}$	Regression model output corrected by bias-variance
$\overline{y_{cal}}$	Mean of the observed time series in the calibration set
$\underline{\hat{y}_{cal}}$	Mean of the modeled time series in the calibration set
$\sigma_{y,cal}$	Standard deviation of the observed time series in the calibration set
$\sigma_{\hat{y},cal}$	Standard deviation of the modeled time series in the calibration set
ρ	Pearson correlation coefficient

Nomenclature

Acc	Accuracy
APOLLO_NG	AVHRR Processing scheme Over cLouds Land and Ocean Next Generation
BD	Bias and standard Deviation correction
CAMS	Copernicus Atmosphere Monitoring Service Radiation Service
CERES	Clouds and the Earth's Radiant Energy System
CI	Confidence interval
CM	Correlation Matrix
DHI	Diffuse Horizontal Irradiance
DNI	Direct Normal Irradiance
ECMWF	European Center for Medium-Range Weather Forecasts
ERA5	European Reanalysis
GHI	Global Horizontal Irradiance
KDE	Kernel Density Estimation
MERRA	Modern-Era Retrospective Analysis for Research and Applications
MSG	Meteosat Second Generation satellites
MLP	Multilayer Perceptron
MLR	Multilinear Regression
NSRDB	National Solar Radiation Database
PCA	Principal Component Analysis
QM	Quantile Mapping
SEL	Feature Selection
SSI	Surface Solar Irradiance
SS4	Skill Score proposed by Taylor (2001)
STDRatio	STandard Deviation Ratio

1. Introduction

In large-scale photovoltaic and heliothermic projects, accurate estimation of long-term surface solar irradiance (SSI) time series is necessary to define the project's bankability. The interannual variation of solar radiation requires at least ten years data for solar resource assessment (Habte et al., 2017). As long-term solar radiation measurements are expensive and rarely available at the exact project locations, statistical models are used to adapt long-term SSI time series estimated by satellite retrieval or reanalysis products. This procedure is known as site adaptation. During the overlapping period, short-term ground measurements (1–3 years) are combined with modeled datasets to derive statistical methods that are then used to adjust the modeled long-term time series (Polo et al., 2016), thereby forming the basis of the project's solar resource analysis.

The main historical solar irradiance databases are satellite-based models and reanalyses from Numerical Weather Prediction (NWP) models (Abreu et al., 2023). The European Reanalysis (ERA5) from the European Center for Medium-Range Weather Forecasts (ECMWF) and the Modern-Era Retrospective Analysis for Research and Applications (MERRA) from the National Aeronautics and Space Administration (NASA) are examples of reanalyses, while the Clouds and the Earth's Radiant Energy System (CERES), the National Solar Radiation Database (NSRDB), and the Copernicus Atmosphere Monitoring Service (CAMS) Radiation Service are examples of satellite-based databases. Satellite-based models generally provide higher accuracy than NWP to estimate global horizontal irradiance (GHI) and direct normal irradiance (DNI) (Salazar et al., 2020; Yang and Bright, 2020; Sarazola et al., 2023), and are largely employed for solar resource assessment (Huang et al., 2019; Fernández-Peruchena et al., 2020).

Different statistical approaches can be applied for site adaptation, such as bias correction, quantile mapping, and linear and non-linear regression models (Cebecauer and Suri, 2016; Gueymard et al., 2009; Schumann et al., 2011; Ines et al., 2006; Themeßl et al., 2012; Cannon et al., 2015; Ramgolam et al., 2020). The simplest regression model is the linear regression between the modeled and measured SSI (Aguiar et al., 2019; Tahir et al., 2020; Bangarigadu et al., 2021). Gueymard et al. (2012) and Bender et al. (2011) employed a multilinear regression of reanalysis and satellite-based solar data to perform site adaptation. Vernay et al. (2013) proposed using a Fourier transform to determine the dominant frequencies of the error between the measured and modeled clearness index (ratio between GHI and the extraterrestrial irradiance on a horizontal plane, also called top-of-atmosphere irradiance). Fernández-Peruchena et al. (2020) and Dhata et al. (2022) applied a multilinear regression followed by a quantile mapping correction and different distribution mapping techniques, respectively. Tiba et al. (2019), Narvaez et al. (2021), Salamalikis et al. (2022), Ruiz-Munoz et al., (2024), Zainali et al. (2024), and Ledesma and Salazar (2025) used different machine learning techniques (non-linear models) to perform the site adaptation (e.g., support vector machine, neural networks, random forest, AdaBoost, and XGBoost). Yang and Gueymard (2021) proposed a probabilistic site adaptation using gridded modeled irradiance inputs. Jadidi et al. (2020) proposed to estimate the GHI hourly distribution using a Bayesian approach. Babar et al. (2020) applied daily GHI averages from CERES and ERA5 to build a new dataset based on Random Forest regression trained using ground measurements from Norway and Sweden. Dhata et al. (2024a) used five deterministic models to generate probabilistic time-series assuming normal distribution. Recent developments have also integrated site adaptation within high-resolution satellite-based solar resource assessment frameworks as in Bayasgalan et al. (2024). The authors use visible channels from Himawari 8/9, ground albedo information and solar geometry in a physics-based calibration approach. Atmospheric extinction parameters are optimized to minimize the Root Mean Square Error (RMSE) between satellite-derived and ground-measured GHI. Although the approaches cited above improved site accuracy, most of them rely on global correction models applied to the entire time series that do not explicitly account for irradiance variability regimes and local phenomena.

Several clustering solar radiation time series approaches have been proposed in the literature, mostly using intra-hour data to classify hours and days in a set of solar irradiance regime categories (Pérez-Ortis et al., 2016). The methods include unsupervised techniques, where statistical models are applied to the dataset to define the clusters, and supervised techniques, where class information is provided as manually-defined targets to train a learning algorithm. Using unsupervised classification, Muselli et al. (2000) classified days of GHI time series as clear, overcast, and partly overcast with the aggregation Ward's method using the hourly clearness index. Moreno et al. (2017) applied the k-medoids algorithm and variability indices to classify different days from DNI time series. Lewis et al. (2021) proposed a classification algorithm for 15-min GHI time series based on the Fourier transform of the GHI signal and additional parameters derived from the irradiance series. Several other works using unsupervised algorithms were proposed to classify SSI time series (Soubdhan et al., 2009; Gastón-Romeo et al., 2011; Stein et al., 2012; Chicco et al., 2014; Fortuna et al., 2016) with models mainly based on threshold values of specific variables such as clearness index, air mass, GHI, and others (Duchon and O'Malley, 1999; Li and Lam, 2001; Harrouni et al., 2005; Kang and Tam, 2013; Trueblood et al., 2013; Reno and Hansen, 2016; Ruiz-Arias and Gueymard, 2023). A few works deal with supervised classification due to the extensive work in determining the ground truth (Calbó et al., 2001; Pagès et al., 2003; Schroedter-Homscheidt et al., 2018). Recently, Miranda et al. (2026) proposed a supervised classification algorithm based on a reference database of 277 hours of GHI measurements at minute resolution from the Carpentras BSRN station manually classified in eight different sky conditions. A combination of unsupervised and supervised classification methods has not been previously used for site-adaptation procedures. This double classification scheme was applied in solar energy short-term forecast (Alimohammadi and He, 2016; Jiménez-Pérez and Mora-López, 2016) but has not yet been addressed for site adaptation.

Polo et al. (2020) highlighted the limitations of global corrections, citing cases where improving GHI estimates using statistical techniques is challenging, particularly when the GHI satellite-based model is already highly accurate. Some studies have therefore introduced local models to improve statistical models by splitting the time series according to previous sky condition classifications (e.g. clear, variable or overcast), as reported by Polo et al. (2020), Fernández-Peruchena et al. (2020), Dhata et al. (2022), Salamalikis et al. (2022), Dhata et al. (2024b) and Pinto et al. (2026). However, although these studies apply local models, they rely on threshold-based rules or observed irradiance indices to define the sky conditions. Therefore, they do not connect the modelled variables with ground-based variability conditions using a supervised approach. The exception is Dhata et al. (2024b), which uses machine learning to estimate the threshold-based values of the observed clear sky index in order to split the time series into clear and cloudy conditions. However, the decision is not made using time series variability.

Existing site-adaptation frameworks rarely integrate unsupervised clustering and supervised classification within a unified modeling structure. Furthermore, they do not explicitly combine global and local regression through a dedicated training strategy. To the best of our knowledge, no study has yet investigated a solar irradiance site adaptation procedure that integrates: (i) variability-based clustering derived from minute-resolution data; (ii) supervised classification linking modelled variables to ground-based sky conditions; (iii) combination models merging global and local approaches; and (iv) a systematic assessment of data-splitting strategies to mitigate seasonal bias. The methodological development presented in this study is motivated by this research gap.

This work presents a novel site adaptation method using global, local, and combination models for adjusting long-term GHI and DNI time series from CAMS Radiation Service products (Schroedter-Homscheidt et al., 2024), based on Meteosat Second Generation (MSG) and Himawari satellites. Global models are applied to the complete time series, while local models are trained on subsets defined by sky condition classes. These classes are derived from a classification algorithm that splits the SSI time series into five clusters (clear, partly clear, variable, partly variable, and overcast sky) using the clear sky index, the variability index proposed by Stein et al. (2012), and their joint probability. Unlike previous site-adaptation works (Polo et al., 2020; Fernández-Peruchena et al., 2020; Salamalikis et al., 2022), which used threshold-based rules to derive sky conditions, this study first applies threshold-based filtering to classify basic cases in clear and highly variable skies, and then uses a k-means approach with fixed initial centroids to ensure consistent convergence across sites. This provides a more robust classification of sky conditions that captures local short-term variability and cloud effects. These classes are then used as targets in a supervised classification algorithm, enabling the modelled variables to reproduce the local effects of clouds on irradiance. Additionally, a dedicated combination set is introduced to train the combination models, merging global trends in the time series with the local effects of each sky condition. Finally, to mitigate the effects of seasonality on the site adaptation models, various data splitting strategies before training the models are tested: maintaining the sequence; alternating intervals; and randomizing the time series data.

The proposed site adaptation framework demonstrates substantial improvements over the original CAMS Radiation Service estimates for both GHI and DNI across the investigated sites. The combination models consistently provide the best performance, reducing systematic bias while improving RMSE relative to both the global and local models. In addition, the proposed sky-condition classification improves the estimation of SSI time series under different irradiance variability regimes. The site adaptation framework also reduces bias and RMSE in long-term monthly irradiation estimates compared with the original CAMS model, which is particularly relevant for long-term solar resource assessment and bankability analyses. These results demonstrate the practical value of combining unsupervised clustering, supervised classification, and global-local modeling within a unified site adaptation framework.

This article is organized as follows. Section 2 describes the measured and modeled data and provides a description of the sites under consideration. Section 3 describes the clustering algorithm and the site-adaptation techniques. It also presents the performance metrics and evaluation framework. Section 4 presents and discusses the results of the unsupervised and supervised classifications as well as the site adaptation methods, and Section 5 summarizes the main conclusions.

2. Ground and modeled datasets

Southern Hemisphere stations were selected (Tab. 1) covering different site conditions and satellite viewing angles, as this last feature is an important driver of satellite-based estimates' uncertainty (Laguarda et al., 2020). El Rosal and Salta are sites located on the border of the MSG field of view, thus susceptible to parallax and cloud representation problems due to the large view angle and pixel size. Different site altitudes are favored, and three of the stations have similar latitudes. In particular, El Rosal and Salta are close measuring sites, but with very different altitudes and climate conditions, located in the Andes and Salta city, respectively. While cloudy skies are frequent in Salta, cloudiness is seldom observed in El Rosal due to the Andes clouds blocking. Such a situation is especially challenging for MSG solar irradiance estimates, as cloud-contaminated large pixels can be assigned to very different site conditions, and parallax issues are enhanced. These two sites' locations provide an interesting dataset for testing site-adaptation methods. The other sites (Petrolina and Gobabeb) are better positioned concerning the MSG satellites. These stations also allow for testing the proposed site-adaptation method for DNI, a solar irradiance magnitude that is harder to estimate through satellite retrieval. With this choice, the proposed procedure is tested for GHI and DNI, and four different climates are addressed, following the updated Köppen-Geiger climate classification map (Peel et al., 2007). Petrolina and Gobabeb stations also measure the diffuse horizontal irradiance (DHI), allowing the comparison of the model results for this irradiance component. The ground measurements data passed for the Baseline Surface Radiation Network (BSRN) quality assurance procedure (Driemel et al. 2018; Long and Dutton, 2022) and additional quality tests as proposed by Miranda et al. (2022).

Variables from the CAMS Radiation Service (referred to as CAMS for simplicity) and the ERA5-Land reanalysis of ECMWF are used as input for the site adaptation. CAMS provides time series of GHI, DHI and DNI in all-sky and clear-sky conditions with a temporal resolution of 15 minutes. These are spatially interpolated to the point of interest and are available from 2004 until two days ago for any point inside the field of view (FOV) of the Meteosat satellites (which cover Europe, Africa, the Middle East and parts of South America), and from 2016 until two days ago for any point inside the FOV of the Himawari satellites (which cover the Asia-Oceania region). CAMS estimates SSI using the Heliosat-4 method (Qu et al., 2017), a fast parametrization of the libRadtran (Mayer et al., 2005 and 2010) radiative transfer model. It uses the McClear clear sky model module (Lefèvre et al., 2013; Gschwind et al., 2019) to input atmospheric parameters such as aerosol, ozone and water vapour. It also uses the cloud retrieving algorithm APOLLO_NG (Klüser et al., 2015; Schroedter-Homscheidt et al., 2022) to input physical cloud parameters derived from satellite imagery for the cloud module. CAMS data is available at its Atmospheric Data Store (ADS[1]). In total, 29 variables from CAMS Radiation Service expert mode were used: all-sky (4) and clear-sky (4) irradiance time series, all-sky irradiance time series with no bias correction (4), top-of-atmosphere irradiance (1), solar zenith angle (1), a variable to define the summer and winter split (1), total column of water vapor and ozone (2), aerosol optical depth for five different atmosphere characteristics (5), albedo variables from MODIS (4), cloud optical depth (1), cloud coverage of the pixel (1) and cloud type (1). Additionally, nine variables from ERA5-Land (wind speed, dew point, skin temperature, ambient temperature, surface pressure, surface solar radiation downwards, surface thermal radiation downwards, total precipitation, and evapotranspiration) were also employed. CAMS variables from the expert mode are in 1-min resolution and were averaged to 15-minute resolution, as MSG-derived cloud properties are in this time resolution. The site adaptation is also performed in the same time scale. Linear interpolation was employed to convert the hourly ERA5-Land variables to the 15-minute resolution. The results are also evaluated by comparing with the Solcast irradiance-based model, which considers the GOES-Satellite in the irradiance model, providing a better viewing angle for the stations in South America (Solcast, 2024).

Tab. 1. Meteorological stations used in this work (1-minute temporal resolution).

Station	El Rosal	Salta	Petrolina	Gobabeb (BSRN)
Country	Argentina	Argentina	Brazil	Namib
Latitude (°)	-24.393	-24.72872	-9.107	-23.5614
Longitude (°)	-65.7683	-65.40958	-40.442	15.04198
Altitude (m)	3355	1233	404	409
Satellite view angle	76.7°	76.4°	47.9°	32.3°
Climate classification	BSk	Cwb	BSh	BWh
Measurement period	2013 - 2016	2013 - 2015	2018 - 2022	2015 - 2020

[1]https://ads.atmosphere.copernicus.eu/datasets/cams-solar-radiation-timeseries

3. Classification and site adaptation methods

GHI all-sky measurements and GHI clear-sky from the McClear model (Lefèvre et al., 2013) in 1-minute resolution are employed to classify the sky conditions on 15-minute windows using a non-supervised classification method. The data are averaged on a 15-minute resolution, which is the target time basis for site-adaptation. Global models are applied to the whole data series, i.e. a single method is trained to site-adapt the output time series. Differently, in local models, the data is divided according to the sky condition using a supervised algorithm whose targets are the classes defined previously by the non-supervised approach. It must be noted that the classes are originally defined using ground measurements with the non-supervised technique, with another classification algorithm being trained to estimate the classes from the remote sensing information (which is available for several years and therefore can site-adapt the whole data series). The introduction of this classification ability of the sky conditions is key to this site-adaptation proposal. The variables from CAMS and ERA5-Land are used as input information for the supervised algorithm and the site-adaptation procedure. A feature selection/extraction procedure is applied before site-adaptation to minimize redundancy in the input dataset as described in section 3.4. A diagram of the process is shown in Fig. 1. Different dataset split strategies are tested to divide the data into calibration, combination, and test sets. The global and local models are trained with the calibration set, and the combination models are trained within a particular set named here as combination set, as this requires each standalone model to be adjusted previously.

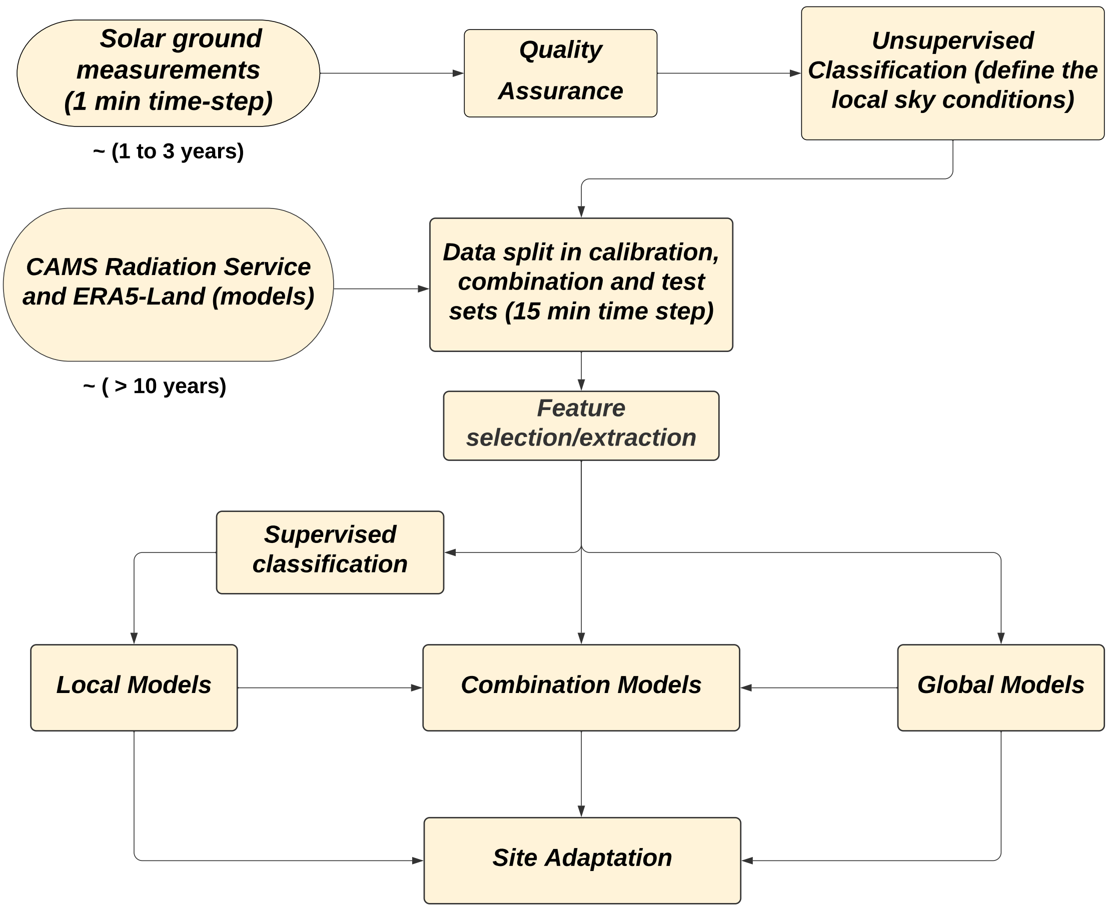


Fig. 1. Site adaptation methodological diagram.

3.1 Unsupervised Classification: definition of the sky conditions classes

The sky conditions classification is based on two indices that are relevant for identifying different SSI variability conditions: the clear sky index (k_c, which is the ratio of the measured GHI to the clear sky GHI) and the variability index (VI) of Stein et al. (2012). These two indices scatter in a characteristic 'arrow-shaped' pattern, as reported by Stein et al. (2012). This pattern facilitates the identification of distinct sky conditions using either threshold values (as done by Trueboold et al., 2013) or statistical clustering algorithms. VI quantifies the GHI variability within 15-minute time windows according to Eq. 1 (adapted from Stein et al., 2012) when applied to minute-by-minute data. High VI values indicate high variability, while low values indicate no variability associated with clear or overcast skies.

$$VI = \frac{\sum_{k=2}^{n} \sqrt{(I_{g,k} - I_{g,k-1})^2 + 1}}{\sum_{k=2}^{n} \sqrt{(I_{gClearSky,k} - I_{gClearSky,k-1})^2 + 1}} \tag{1}$$

where $I_{g,k}$ and $I_{gClearSky,k}$ are the measured GHI and the clear-sky estimates, respectively, and n is the number of 1-minute data samples within each 15-minute interval (e.g. $n = 15$).

The classification method combines threshold-based filtering and unsupervised clustering to identify the sky conditions. Data with more than 83° of solar zenith angle and k_c greater than 1.3 are not considered. The threshold-based filtering (step 1) is defined as follows: clear sky conditions (class 1) are empirically obtained by setting thresholds for the absolute difference between the average and standard deviation of the measured GHI and clear-sky GHI (Eq. 2) while variable conditions (class 4) are identified by VI values greater than the 95th percentile of VI (Eq. 3). These data are excluded from the clustering step, as they are already classified.

$$\left|\overline{I_g} - \overline{I_{gClearSky}}\right| < 20\ W/m^2\ \&\ \left|\sigma_{I_g} - \sigma_{I_{gClearSky}}\right| < 2\ W/m^2 \tag{2}$$

$$VI > P_{95}(VI) \tag{3}$$

where $\overline{I_g}$ and σ_{I_g} are the GHI average and standard deviation on 15-min resolution obtained from the GHI minute data, $\overline{I_{g,ClearSky}}$ and $\sigma_{I_{g,ClearSky}}$ are the clear sky GHI average and standard deviation on 15-min resolution obtained from the clear sky GHI minute data, and $P_{95}(VI)$ is the percentile 95 of the variability index on 15-min resolution as calculated by Eq. 1.

In a second step, the remaining data are analysed using a Kernel Density Estimation (KDE) to estimate the joint probability density of k_c and VI. Prior to the KDE estimation, the VI values used for classification are normalised by its maximum value to rescale the index between 0 and 1. The resulting 3D-scattering ($VI\ x\ k_c\ x\ KDE$) facilitates the identification of distinct clusters in the data, as shown in Fig. 2. These clusters are physically interpretable based on irradiance level (k_c) and variability. A peak of probability density is formed in the region with high k_c and lower VI (red circle), which is associated with a clear sky. The bottom of this peak is associated with a partly clear sky condition (magenta circle), and presents high k_c and low VI. The blue circle indicates overcast sky (low k_c and VI). The black circle on the right shows the highest VI values, indicating variable sky conditions. Finally, the middle of the graph shows an intermediate region between the partly clear, overcast, and variable sky classes, labelled 'partly variable sky' (green circle). The k-means clustering (Macqueen, 1967; Jain et al., 2000) is then applied using the initial centroids highlighted with 'X' markers in green. The initial centroid values are presented in Tab. 2 and were defined following an inspection of data from the four sites. Similar 3D-scattering patterns were observed in the analysed stations, supporting the robustness of the initial centroid choice and providing stable convergence for the k-means algorithm. Without a fixed initialization, the algorithm may converge to different local minima that do not correspond to the intended physically interpretable sky-condition categories. Only the initial centroids are fixed; the centroid positions remain free to evolve during the k-means optimization. The same centroids and number of clusters (five) are used for all stations to ensure methodological consistency. This indicates that the classification captures the intrinsic physical regimes of solar irradiance variability rather than site-specific characteristics.

Tab. 2. Initial centroids from the $VI\ x\ k_c\ x\ KDE$ diagram.

Class	Clear sky index (k_c)	Variability index (VI)	Kernel Density Estimation (KDE)
1 - Clear sky	0.940	0.025	0.690
2 - Partly clear sky	0.912	0.156	0.025
3 - Overcast sky	0.150	0.010	0.070
4 - Variable sky	0.766	0.937	0.007
5 - Partly variable sky	0.735	0.539	0.005

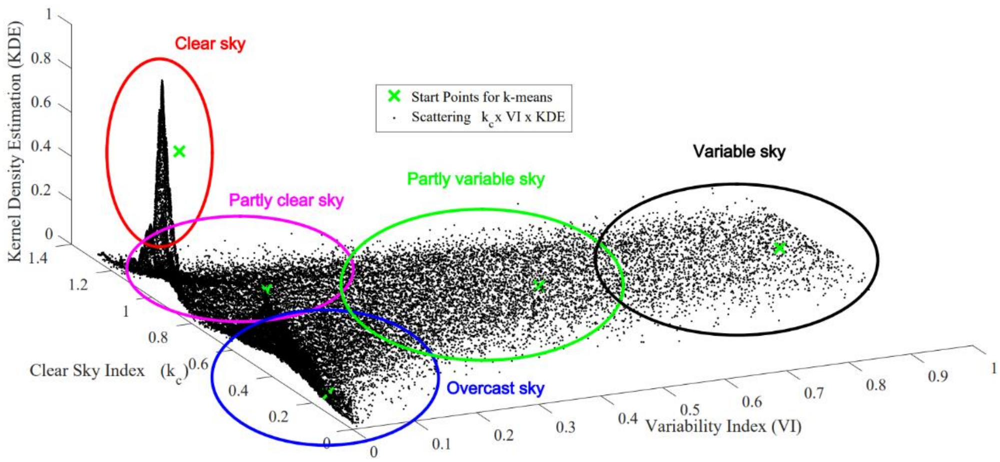


Fig. 2. Proposed classes in this work according to the scattering $VI\ x\ k_c\ x\ KDE$.

3.2 Data split in calibration, combination and test sets

The next steps require parameter adjustment, architecture definitions, and performance assessment. The class vector obtained in the previous step and the variables from the historical database are divided into calibration, combination, and test sets to apply the statistical models. The optimal data split for site adaptation training and performance assessment and the variability of the output of these procedures with different data splits are issues that have not been addressed previously. For evaluating the model sensibility to different data splits, an experiment was carried out by applying 42 subset settings divided into four main categories: 1) 10 settings (splits 1 to 10) for different positions of chronologically sorted sets, which are defined by a moving window and accounts for 70%, 10%, and 20% of total samples, respectively; 2) 12 settings (splits 11 to 22) intercalating daily periods (e.g. two first days for calibration, third for combination and forth for the test set); 3) 10 settings (splits 23 to 32) applying different sets sizes with shuffled daily periods of 15-minute time series data; 4) 10 settings (splits 33 to 42) applying different sets sizes with shuffled 15-min averaged data. The different settings are summarized in Tab. 3. As it will be shown afterwards, the results vary depending on the choice. This variability affects the metrics in absolute terms but not the relative merit of each algorithm, i.e. the algorithm's order is the same within each dataset split; hence, the conclusions are robust even under this situation. As the site adaptation aims to derive a stable mapping between ground measurements and modeled input features, the model is not using past irradiance values to predict future values. Therefore, randomization allows the calibration and test sets to contain representative samples of different seasons, solar zenith angles, and cloud conditions, which is particularly important for deriving a stable correction function for long-term resource assessment.

Tab. 3. Different settings applied to split the data.

Configuration	Splits	Calibration set	Combination set	Test set
1) Chronologically ordered	1 to 10	70%	10%	20%
2) Intercalated days	11 to 22	50%	25%	25%
3) Shuffling the 15-min data daily	23 to 26	~80%/~70%/~60%/~50%	~10%/~20%/~30%/~40%	10%
	27 to 29	~70%/~60%/~50%	~10%/~20%/~30%	20%
	30 to 31	~60%/~50%	~10%/~20%	30%
	32	~50%	~10%	40%
4) Shuffling all the 15-min timestamps	33 to 36	~80%/~70%/~60%/~50%	~10%/~20%/~30%/~40%	10%
	37 to 39	~70%/~60%/~50%	~10%/~20%/~30%	20%
	40 to 41	~60%/~50%	~10%/~20%	30%
	42	~50%	~10%	40%

3.3 Supervised Classification: learning to classify based on the modeled data

As the ground measurements cover a short period (1 to 3 years), the variables from the historical databases should learn the class vector obtained in the unsupervised classification to cluster the data for periods in which no ground measurements were available. In this paper, the supervised classification method chosen is the random forest (Breiman, 2001), after comparing its classification performance with other machine learning techniques (multilayer perceptron, convolutional neural network, gradient boosting, and support vector machine).

Random forest is an ensemble that combines different estimates from decision tree (DT) models. Ensemble models use different classifiers trained on various subsets of the training set to obtain a combined classification model. Because the ensemble combines different classification models, it is less prone to overfitting than using only one classification model, such as the DT, and also can perform well even in the presence of outliers or noise from the time series (Han et al., 2012; Aggarwal et al., 2015). All input variables from CAMS and ERA5-Land are used with the target being the classes from the unsupervised classification. A grid search was performed on the train set to search the best set of hyperparameters according to Tab. 4, as follows: the use or not of bootstrap, the maximum depth of the DT (*max_depth*), the number of features randomly selected at each split of the tree (*max_features*), the minimum number of samples each leaf should have (*min_samples_leaf*), the minimum number of splits each internal node should have (*min_samples_split*) and the number of DT (*n_estimators*). The hyperparameter optimization was performed using the GridSearchCV implementation from the scikit-learn library[2], with 3-fold cross-validation over the calibration set. The best model was selected based on classification accuracy. Parallel processing was employed to accelerate the grid search.

Tab. 4. Grid search for the best hyperparameters of random forest model.

Classification Model	Grid search
Random Forest	"bootstrap": [True, False], "max_depth": [10, 20, 30], "max_features": ['auto', 'sqrt'], "min_samples_leaf": [2, 4], "min_samples_split": [2, 5], "n_estimators": [200, 500, 700]

3.4 Feature selection/extraction

Feature selection algorithms (shortly named as SEL here) search for the best subset of input variables with minimum redundancy among the variables and maximum relevance between each variable and the target (Hall, 2000; Yu and Liu, 2004). This study applies a Pearson correlation-based filter for the feature selection, while principal component analysis was employed for the feature extraction. The feature selection procedure is based on threshold values of Pearson correlations (called here as Z) among the input variables (step 1) and between input variables and the target (step 2), as described in Tab. 5. The target is the clearness index for GHI (k_t – defined as the ratio between GHI and horizontal extraterrestrial irradiance) and normal transmittance for DNI (k_n – defined as the ratio between DNI and extraterrestrial irradiance). Different values for Z were chosen for each step and for local and combination models. The Z values at each step were heuristically defined by inspecting the data.

A principal component analysis (PCA) is used to conduct feature extraction. PCA is a well-known technique to reduce dimensionality in datasets with redundant variables (Hotelling, 1993). A linear transformation is applied to the original input variables, obtaining the principal components set, from which a smaller subset can be chosen to explain a certain percentage of the total variance of the input data (Jollife, 1986). In this work, the first principal components that explain 95% of the total original dataset variance are chosen. The input data is normalised subtracting the mean average and dividing it by the standard deviation before applying PCA.

The PCA extraction method is applied in the global, local, and combination models. In contrast, the wrapper/filter feature-selection algorithm (SEL) is applied independently for each local model and for each sky-condition class, as well as for each combination model. Consequently, the subset of selected variables is not fixed across all experiments, but varies according to the irradiance component, station, data split, and sky-condition class. To illustrate the final variable selection produced by the SEL algorithm, Appendix C presents representative examples for two stations considering a specific data split and irradiance component. Tab. 6 from Section 3.6 also specifies which site adaptation models use feature extraction (PCA) and which use feature selection (SEL).

Tab. 5. Two steps procedure of the Feature Selection Algorithm.

Feature Selection Algorithm (SEL)		
Steps	Local Models (selection applied in the calibration set)	Combination Models (selection applied in the combination set)
Step 1	1) Calculate the correlation matrix (CM) of input variables; 2) Find all correlations higher than Z1 = 0.99 in the CM; 3) Drop all the corresponding columns of these input variables.	1) Calculate the correlation between the input variables and the target; 2) Find the correlations lower than Z3 = 0.85; 3) Drop the respective variables.
Step 2	1) Calculate the correlation between the remaining input variables and the target; 2) Find the correlations lower than Z2 = 0.2; 3) Drop the respective variables.	1) Calculate the CM of the remaining input variables; 2) Find all correlations higher than Z4 = 0.999 in the CM*; 3) Drop all the corresponding columns of these input variables.

*Z4 was chosen high, because the input models used in the combination are highly correlated.

3.5 Performance metrics and evaluation framework

The models are evaluated with common metrices, such as the mean bias error (MBE), the root mean square error (RMSE), their normalised versions (MBEn and RMSEn, relative to the measured data average), the standard deviation ratio (STDRatio), the Pearson correlation (ρ) and the skill score (SS4) as proposed by Taylor (2001), as shown in Eqs. 4-8. The SS4 indicates the model's overall performance as it considers both the STDRatio and correlation. Finally, the accuracy (Eq. 9) is used to evaluate the results of the random forest classification model.

$$MBE = \frac{\sum_{N}^{i=1}(\hat{y}_i - y_i)^2}{N} \tag{4}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{5}$$

where $\hat{y}_i$ and y_i are the modeled and observed time series, respectively, and N is the number of points in the time series,

$$STDRatio = \sigma_{mod}/\sigma_{obs} \tag{6}$$

where σ_{mod} and σ_{obs} are the modeled and observed time series, respectively,

$$\rho = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2(\hat{y}_i - \bar{\hat{y}})^2}} \tag{7}$$

where $\bar{y}$ and $\bar{\hat{y}}$ are the observed and modeled time series average, respectively.

$$SS4 = \frac{(1+\rho)^4}{4(STDRatio + 1/STDRatio)^2} \tag{8}$$

$$Acc = (True\ Positive + True\ Negative)/(Total\ data\ amount) \tag{9}$$

3.6 Site adaptation

As previously stated, this work applies global, local, and combination models for site adaptation. Global models are those whose parameters are inferred based on minimizing a loss function over the entire time series of interest. Local models do so for subsets of this series, defined according to the local sky conditions through the random forest classification algorithm. Combination models take the outputs from global and local models to enhance the results over the combination set. Finally, all the strategies are evaluated in the test set.

The regression models employed for site adaptation are the multilinear regression (MLR) and the multilayer perceptron neural network (MLP). These models are chosen for their simplicity within their respective frames (linear and non-linear). The key element introduced in this proposal is the prior sky classification derived from the input data, which is then utilized to derive local and combination models. The input of the models are the variables from CAMS (29) and ECMWF (9), as described in Section 2, together with the GHI CAMS clearness index (k_{tCAMS} – defined as the GHI CAMS divided by the top-of-atmosphere irradiance) or the DNI CAMS normal transmittance (k_{nCAMS} – defined as the DNI CAMS divided by the extraterrestrial irradiance), depending on whether the site adaptation is performed for k_t or k_n, as well as the extraterrestrial irradiance. All input variables (40) for the global and local models are preprocessed by PCA or SEL, as described in section 3.4. The use of additional post-processing stages is also tested, by applying a quantile mapping (QM) correction or a bias-variance correction (BD). In the following, a brief description of the regression and post-processing models are presented:

Multilinear Regression: In the MLR, the input variables are employed to estimate the target variable linearly. The minimum squares method was used to find the best linear coefficients that minimize the mean square error (Wilks, 2013).

Multilayer Perceptron: The MLP is a feedforward neural network architecture consisting of an input layer, multiple hidden layers, and an output layer (Dawson and Wilby, 2001; Han et al., 2012; Aggarwal et al., 2015). In this work, the MLPs architecture were designed with three hidden layers containing 5, 4, and 3 neurons, respectively, and a single output neuron. A hyperbolic tangent activation function is used in all hidden layers. Training is performed using the Levenberg-Marquardt (LM) optimization algorithm (Gavin et al., 2020), which iteratively adjusts weights through a combination of Gauss-Newton and gradient-descent steps. The MLP was implemented using the MATLAB Neural Network Toolbox (feedforwardnet). The training process minimizes the mean squared error (MSE) between predictions and targets. The LM algorithm does not require a fixed learning rate, as it adaptively regulates the damping parameter (μ) to control the step size. The training was limited to 50 epochs.

To improve generalization and reduce sensitivity to weight initialization, the calibration dataset is randomly divided into 80% for training and 20% for validation. A total of 100 independent training runs are performed with different random initializations. The final model is selected based on the best SS4 performance over the validation subset. The MLP architecture and training configuration are fixed across all experiments. The combination and test datasets are reserved for model integration and independent evaluation, respectively.

Quantile Mapping: In QM, the inverse of the Empirical Cumulative Distribution Function (*ecdf*) of the observed data is applied to adjust the modeled data. In Eq. 10, $\hat{y}_{corr\_QM}$ is the QM-corrected time series of the modeled data in test set ($\hat{y}_{test}$), $ecfd_{mod}$ is the *ecdf* of $\hat{y}_{test}$ and $ecfd_{obs}^{-1}$ is the inverse transformation of the *ecdf* estimated with the measured time series on the calibration set.

$$\hat{y}_{corr\_QM} = ecfd_{obs}^{-1}(ecfd_{mod}(\hat{y}_{test})) \tag{10}$$

Bias-variance correction: The bias-variance (or bias-deviation) correction between model and observation is done according to Eq. 11, where $\overline{\hat{y}_{cal}}$ is the mean of the regression model in the calibration set, $\sigma_{y,cal}$ and $\sigma_{\hat{y},cal}$ are the standard deviations of the observed and modeled time series in the calibration set, $\overline{y_{cal}}$ is the mean of the ground measurements time series in the calibration set, and $\hat{y}_{corr\_BD}$ is the regression model with bias-variance correction in the test set.

$$\hat{y}_{corr\_BD} = (\hat{y}_{test} - \overline{\hat{y}_{cal}})\left(\frac{\sigma_{y,cal}}{\sigma_{\hat{y},cal}}\right) + \overline{y_{cal}} \tag{11}$$

Tab. 6 summarises all the site adaptation models that were evaluated in this study. The proposed combination models integrate the outputs of global and local approaches using a subset of the data dedicated to this purpose, referred to as the combination set. Global models can correct the overall bias of the time series, whereas local models focus on adjusting specific parts of the time series according to the sky conditions classes. The combination models aim to integrate both approaches, benefiting from the perspectives of the global and local models while reducing residual errors. They use the CAMS model (1), the global models (6) and the local models (8) as input. These inputs are then used to train the combination model, which is implemented either as an MLR or an MLP in order to learn how to combine the results of the individual models, as shown by Eq. 12, where $\hat{y}_{comb}$ denotes the combination model, $\hat{y}_{CAMS}$ denotes the CAMS model, $\hat{y}_{Gi}$ denotes the global models, $\hat{y}_{Li}$ denotes the local models, and f represents either a MLR or a MLP model. Therefore, the combination model is a second-stage regression model trained on the outputs of the CAMS, global, and local models. To reduce redundancy among the input variables, feature selection (SEL) is applied prior to training the combination models. This ensures that only the most relevant model outputs are retained, thereby improving model stability and generalization. As the input variables for the combination model can be highly correlated, feature selection was chosen as the preprocessing step.

$$\hat{y}_{comb} = f(\hat{y}_{CAMS}, \hat{y}_{G1}, \dots, \hat{y}_{G6}, \hat{y}_{L1}, \dots, \hat{y}_{L8}) \tag{12}$$

A key feature of the combination framework is that it uses independent data (from the combination set) to train the combination models. This data is not used to train the global or local models. This ensures that the combination models are trained using data that has not been seen before, which prevents overfitting and enables the model outputs to be adjusted more robustly. This approach is conceptually similar to stacking, where multiple base learners are combined using a higher-level model. The target variable for all the models was the clearness index (k_t) for GHI and normal transmittance (k_n) for DNI.

Tab. 6. Models for site adaptation.

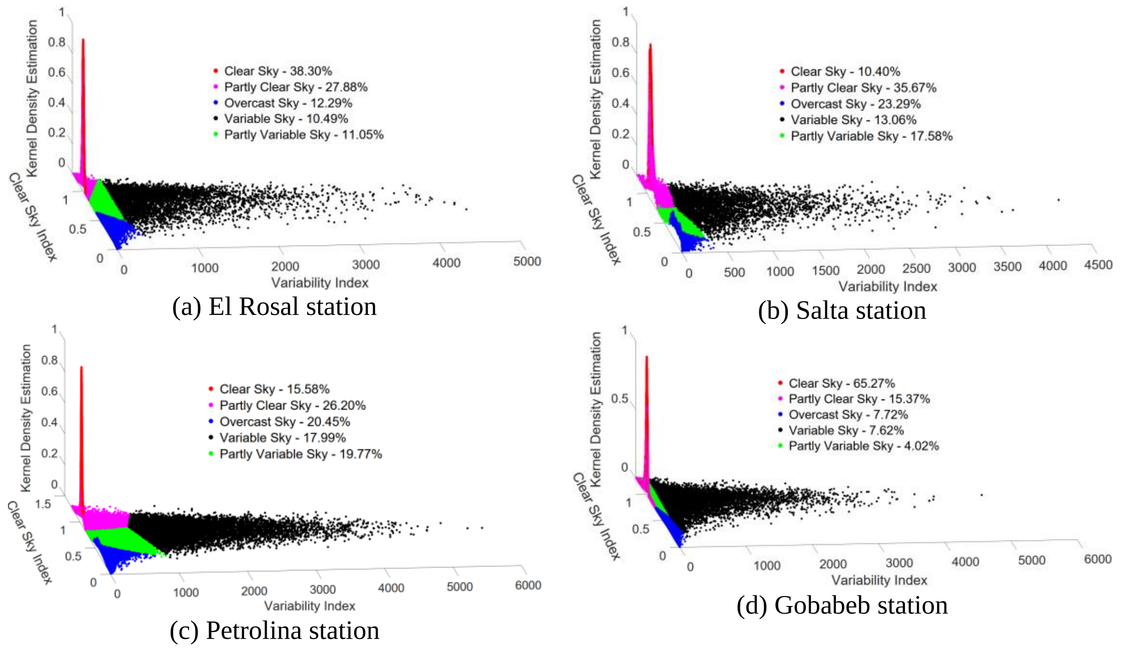
REFERENCE MODELS	
CAMS	GHI, DNI, and DHI from CAMS model
Solcast	GHI, DNI, and DHI from Solcast model
GLOBAL MODELS	
MLR	MLR with input variables preprocessed using PCA
MLR_QM	MLR with input variables preprocessed using PCA and postprocessing using QM
MLR_BD	MLR with input variables preprocessed using PCA and postprocessing using bias-deviation correction
MLP	MLP with input variables preprocessed using PCA
MLP_QM	MLP with input variables preprocessed using PCA and postprocessing using QM
MLP_BD	MLP with input variables preprocessed using PCA and postprocessing using bias-deviation correction
LOCAL MODELS	
MLR_SEL_Loc	5 MLRs (five sky conditions) with input variables preprocessed using SEL in each cluster
MLR_SEL_BD_Loc	5 MLRs with input variables preprocessed using SEL and postprocessing using bias-deviation correction in each cluster
MLR_PCA_Loc	5 MLRs with input variables preprocessed using PCA in each cluster
MLR_PCA_BD _Loc	5 MLRs with input variables preprocessed using PCA and postprocessing using bias-deviation correction in each cluster
MLP_SEL_Loc	5 MLPs with input variables preprocessed using SEL in each cluster
MLP_SEL_BD_Loc	5 MLPs with input variables preprocessed using SEL and postprocessing using bias-deviation correction in each cluster
MLP_PCA_Loc	5 MLPs with input variables preprocessed using PCA in each cluster
MLP_PCA_BD_Loc	5 MLPs with input variables preprocessed using PCA and postprocessing using bias-deviation correction in each cluster
COMBINATION MODELS	
MLRComb	MLR trained on combination set using as input variables the CAMS model (GHI or DNI time-series) and the global and local model outputs preprocessed by SEL
MLRComb_BD	MLR trained on combination set using as input variables the CAMS model (GHI or DNI time-series) and the global and local model outputs preprocessed by SEL and postprocessing using bias-deviation correction
MLPComb	MLP trained on combination set using as input variables the CAMS model (GHI or DNI time-series) and the global and local model outputs preprocessed by SEL
MLPComb_BD	MLP trained on combination set using as input variables the CAMS model (GHI or DNI time-series) and the global and local model outputs preprocessed by SEL and postprocessing using bias-deviation correction

4. Results and Discussion

In this work, site adaptation is composed of a classification stage and an adjustment stage, both of which require evaluation. The assessment of the unsupervised and supervised data classification methods is presented in subsections 4.1 and 4.2 respectively, while the complete site-adaptation results are presented in subsection 4.3. A performance breakdown of the site adaptation methods per sky condition is shown in subsection 4.4.

4.1 Unsupervised classification results

The results of the threshold-based filtering and the k-means algorithm together are shown in Fig. 3. There is a peak of the probability density in the clear sky region with high k_c and small VI (class 1 in red). The overcast conditions are located in the region of low k_c and low VI for all stations, while the partly variable sky represents mixed conditions situated in the middle of classes 2, 3, and 5. As Gobabeb station is in the Namib desert, the percentages of partly variable, variable, and overcast skies are lower, in particular for the partly variable sky, for which only a small portion is seen in the graph. Petrolina and Salta presents a small peak in the blue region (overcast sky), due to the more balanced percentages distribution among sky conditions. Overall, the k-means output shows similar patterns across stations, showing that the classification algorithm is robust in these different climates. Fig. 4 shows an example of the final 15-min window classification plotted with GHI measured data for El Rosal and Petrolina in minute resolution. In general, the classes correspond with the expected behavior of GHI. The black lines representing variable sky contain most of the cloud enhancement cases and the magenta lines show only short-term variability in the GHI signal, but remain close to the clear-sky line for most of the time. Misclassification can be observed at border regions or during transition period between two sky conditions, as is the case of partly clear and partly variable sky classes.



(a) El Rosal station

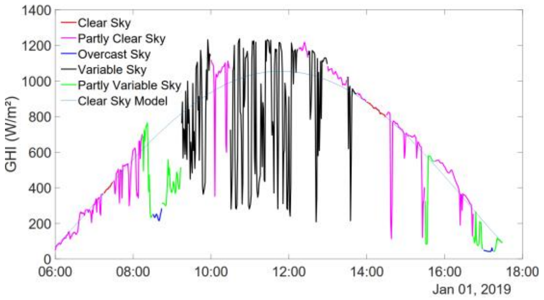
(b) Salta station

(c) Petrolina station

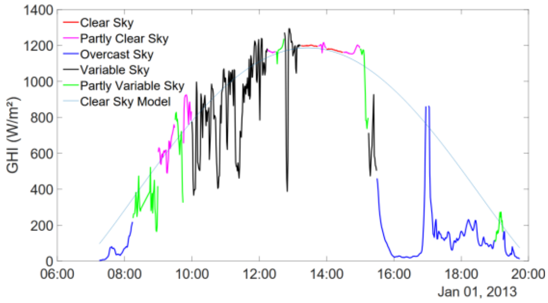
(d) Gobabeb station

Fig. 3. Output of k-means algorithm for all stations.

Fig. 5 presents the probability density curves of the clearness index in minute resolution estimated by KDE. The clear sky and partially clear sky classes present a peak in the higher k_t region (red and magenta lines, respectively), while the overcast sky class shows a peak in the lower k_t region (blue line). The variable sky class (black line) presents a bimodal distribution, with one of the modes at low k_t and the other at high k_t, except for El Rosal station. The partially variable sky class (in green) does not present a specific shape in the different stations. The final percentages per sky condition are presented in Tab. 7. Stations El Rosal and Gobabeb have a predominance of clear sky conditions with more than 66% and 80% of the classes, respectively, being classes 1 or 2. Petrolina and Salta stations have sky conditions more distributed among all classes. Note the different behavior between Salta and El Rosal stations, which are close sites experiencing very different sky conditions during the year. The clustering results and data split percentages in each class are consistent with the station climate classification given in Tab. 1.

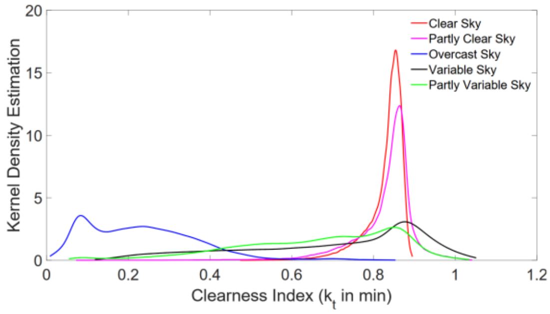


(a) GHI El Rosal station.

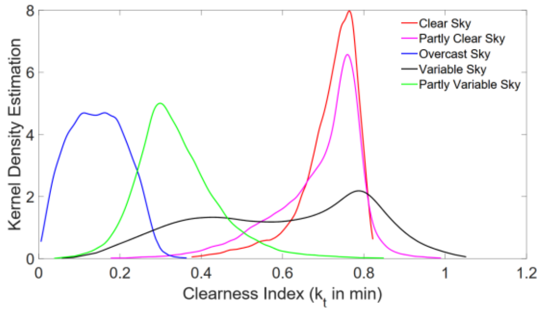


(b) GHI Petrolina station.

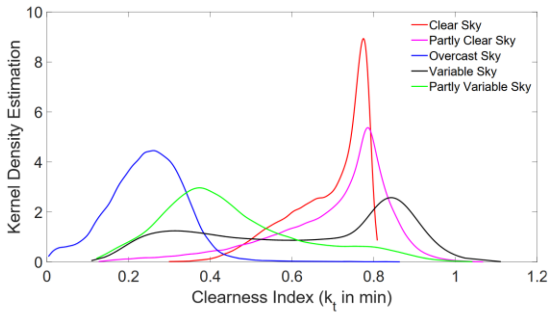
Fig. 4. 15-min classes plotted with GHI measurements at minute resolution.



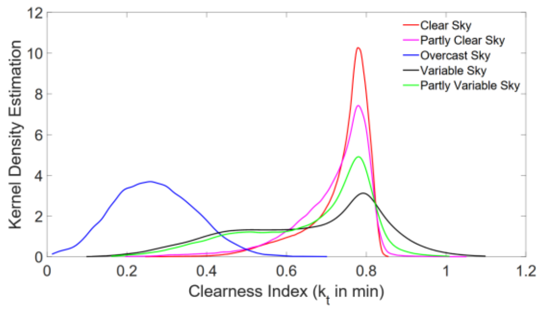
(a) $KDE\ x\ k_t$ in minute resolution for each class of El Rosal station.



(b) $KDE\ x\ k_t$ in minute resolution for each class of Salta station.



(c) $KDE\ x\ k_t$ in minute resolution for each class of Petrolina station.



(d) $KDE\ x\ k_t$ in minute resolution for each class of Gobabeb station.

Fig. 5. Clearness index (k_t) density curves for each class.

Tab. 7. Final percentage of unsupervised classification for all stations.

Class	El Rosal	Salta	Petrolina	Gobabeb
1 - Clear sky	38.3%	10.4%	15.6%	65.3%
2 - Partly clear sky	27.9%	35.7%	26.2%	15.4%
3 - Overcast sky	12.3%	23.3%	20.4%	7.7%
4 - Variable sky	10.5%	13.0%	18.0%	7.6%
5 - Partly variable sky	11.0%	17.6%	19.8%	4.0%

4.2 Supervised classification results

The random forest model was trained for each of the 42 different divisions, using only the calibration set for training and the test set to evaluate the results. Other machine learning models were also tested, for example eXtreme Gradient Boost (XGBoost), which showed slightly lower accuracy performance compared to Random Forest over all stations. Therefore, the Random Forest model was kept for the supervised classification. Fig. 6 presents the accuracy of the random forest on the test set for each data split strategy. The accuracy is higher for the strategy of shuffling all 15-minute timestamps than for the other strategies, with average values of 69%, 72%, 65%, and 83% for El Rosal, Salta, Petrolina, and Gobabeb, respectively. These accuracies should be interpreted with caution, as the temporal autocorrelation of sky conditions may lead to optimistic estimates when neighboring timestamps are distributed across both the calibration and test sets in the fully randomized strategy.

A clear difference in dispersion can be observed among the boxplots of the four split strategies. The chronologically ordered splits exhibits the greatest variation, which can be attributed to imbalanced classes in the calibration and test subsets, resulting from seasonal sampling bias and different atmospheric conditions. In contrast, data-shuffling approaches, whether by intercalating or shuffling days, or randomizing 15-minute timestamps, result in less variance through different data splits, as they ensure a more homogeneous distribution of atmospheric conditions and class frequencies. In particular, fully randomized 15-minute timestamps prior to data splitting led to higher accuracy. This approach breaks the temporal autocorrelation of solar irradiance time series and reduces dependence on the solar zenith angle, ensuring a more homogeneous distribution of classes in the calibration and test sets. The reduced dispersion observed for strategies involving intercalation or data shuffling indicates that the Random Forest model is robust with respect to different data partitions with varying amounts for training and testing (see Tab. 3), and can generalize well when class distributions are preserved. Overall, except for some splits, the relative results between stations are consistent. Since the application of the proposed method is in resource assessment for project bankability, which relies on retrospective analysis of long-term time series, the complete randomization of the time series is considered acceptable. This is not the case for forecasting applications, where temporal consistency must be preserved. This is the first insight provided in this work that the dataset splits influence performance, an overlooked issue in solar resource assessment.

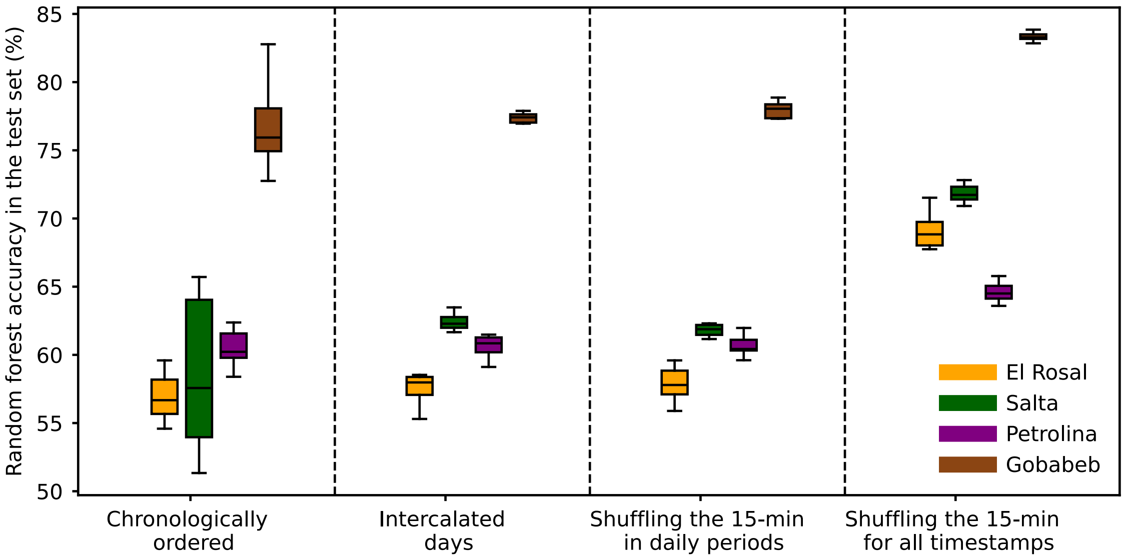


Fig. 6. Accuracy of the random forest model for all stations in the different configuration of data splits for calibration, combination, and test sets.

4.3 Site adaptation

The criteria for selecting the more accurate site adaptation model and the best data split strategy are set out in Tab. 8. Fig. 7 presents the MBEn and RMSEn of the more accurate global, local, and combination models in each data split for El Rosal and Salta stations. These top-performing models in each category (global, local, and combination) are recognized for consistently achieving the highest rankings based on the criteria listed in Tab. 8. These are the MLP_QM and MLP_BD (global model for El Rosal and Salta, respectively), MLR_PCA_Loc (local model) and MLRComb_BD (combination model). The identification of these models allows the comparison of the three strategies based on their best-achieved performance. The local and combination models show better results than CAMS and the global models at El Rosal station. There is a remarkable gain in all metrics due to the site-adaptation procedures at this site. It shall be noted that the satellite estimates have a weak performance in El Rosal, caused by a sudden climate transition in space and a large satellite view angle. By inspecting the RMSEn, an improvement with the local and combination models can be observed (those including clustering techniques), especially when data shuffling is used in the dataset splits. For Salta station, the gain of the site-adaptation procedures is lower, as the satellite estimates have a better performance. Similar to El Rosal, the MBEn metric shows more stability across the splits than the RMSEn. Again, the combination models achieve the best RMSEn, as seen in the data shuffling division. The local models came second, showing improvement by introducing data clustering in the methodology.

Tab. 8. Criteria* to choose the best models and split in the test set.

Model category	Model evaluation for each of 42 splits
Global, local, and combination models	1) The four models with the highest SS4 are selected; 2) Among these four models, select the two with STDRatio closest to one in absolute value; 3) Among these two models, the best one will be that with the lowest RMSEn.
Choice for the more accurate split (after choosing the more accurate model per category)	1) Select the ten splits whose model results present high SS4; 2) Among these ten splits, select the five with STDRatio closest to one in absolute value; 3) Among these five splits, select the three with RMSEn closest to one in absolute value; 4) The best split from the three previous steps will be that with the lowest MBEn.

*The criteria rank the models first by the SS4 and then by the STDRatio and RMSEn. This sequence of statistics was chosen to highlight first the model's ability to reproduce the phase and frequency structure of the target signal (SS4) and then the difference in standard deviation (STDRation) and dispersion (RMSEn). In the case of the split evaluation, it is considered the MBEn in the last step, as it has more results to rank (42 splits).

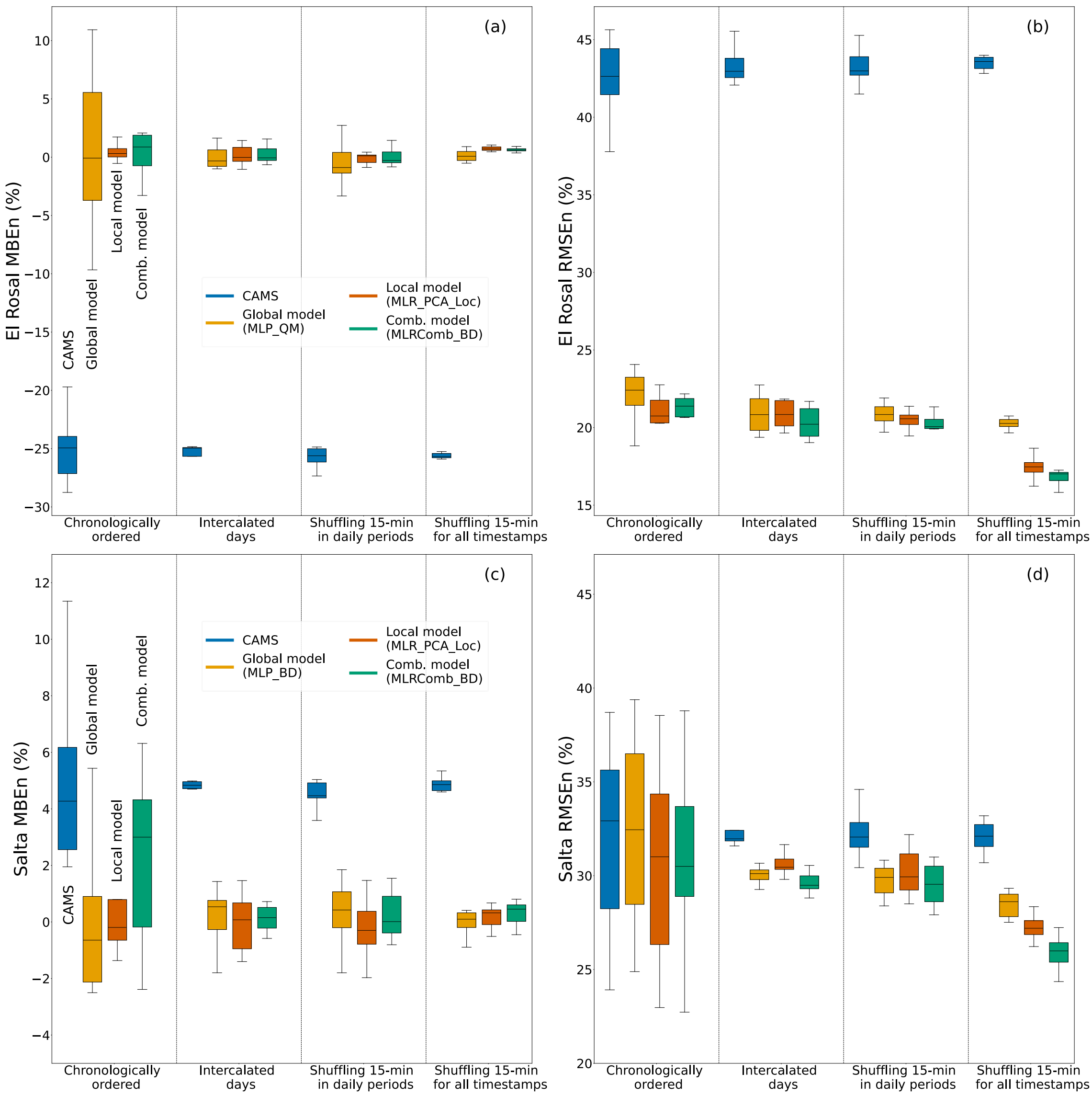


Fig. 7. Results for the best global, local, and combination models for GHI considering different data split strategies in El Rosal and Salta stations (15-min resolution).

The more accurate model for El Rosal is MLRComb_BD in split 40 (RMSEn of 16.9% and correlation of 0.94), as shown in the Taylor's diagram of Fig. 8. The CAMS model presents a correlation of 0.74 and an RMSEn of 43.1%. In contrast, the global models obtain correlation values close to 0.9, while the local and combination models obtain values close to 0.94. The Taylor diagrams for other stations are shown in Appendix A.

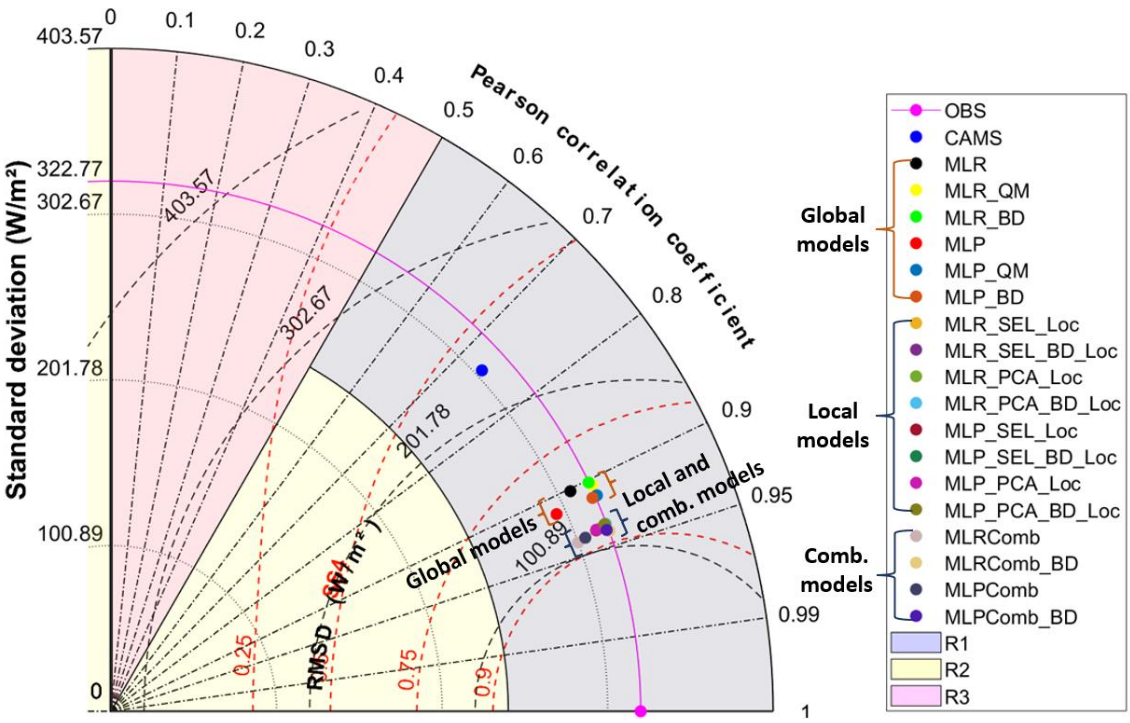


Fig. 8. Taylor diagram for split 40 in El Rosal station.

Figs. 9 and 10 provide the site-adaptation results for GHI and DNI in Petrolina and Gobabeb. The quantitative results provided below follow the metrics found in the last split strategy (shuffling the 15-min for all timestamps). In both stations, the CAMS MBEn overestimation (Petrolina) and MBEn underestimation (Gobabeb) are corrected and site-adapted unbiased estimates are obtained in all models, aside from some divisions in the chronologically ordered strategy. This is notably corrected for DNI, for which the MBEn of CAMS estimates are around +22% for Petrolina and -7% for Gobabeb. For GHI, the MBEn of CAMS estimates is lower but also notable, being +5% and -3%, respectively. RMSEn results are similar for all models, with improvements compared to the original CAMS estimates. These improvements are higher for DNI than for GHI, and the combination models perform slightly better than the others. It is observed that the gain of the combination models is higher for DNI. The results are consistent in both stations, even though the climate is very different. This difference is clear when the mean metrics' values are inspected. For instance, in Petrolina, the RMSEn change comparing original CAMS and combination models is from 24% to 22% for GHI (8.3% gain) and from 51% to 42% for DNI (17.6% gain), and in Gobabeb is from 9% to 7% for GHI (22.2% gain) and from 18% to 13% for DNI (27.8% gain). From these two stations, the gains found at Gobabeb are the highest, explained mostly by local sky conditions under clear sky (see the climate classification of Tab. 1). However, this gain is lower compared to the one observed at El Rosal station, which is importantly affected by an error-prone cloud's observation with a large satellite view angle, pixel size, and the high Andes ground albedo.

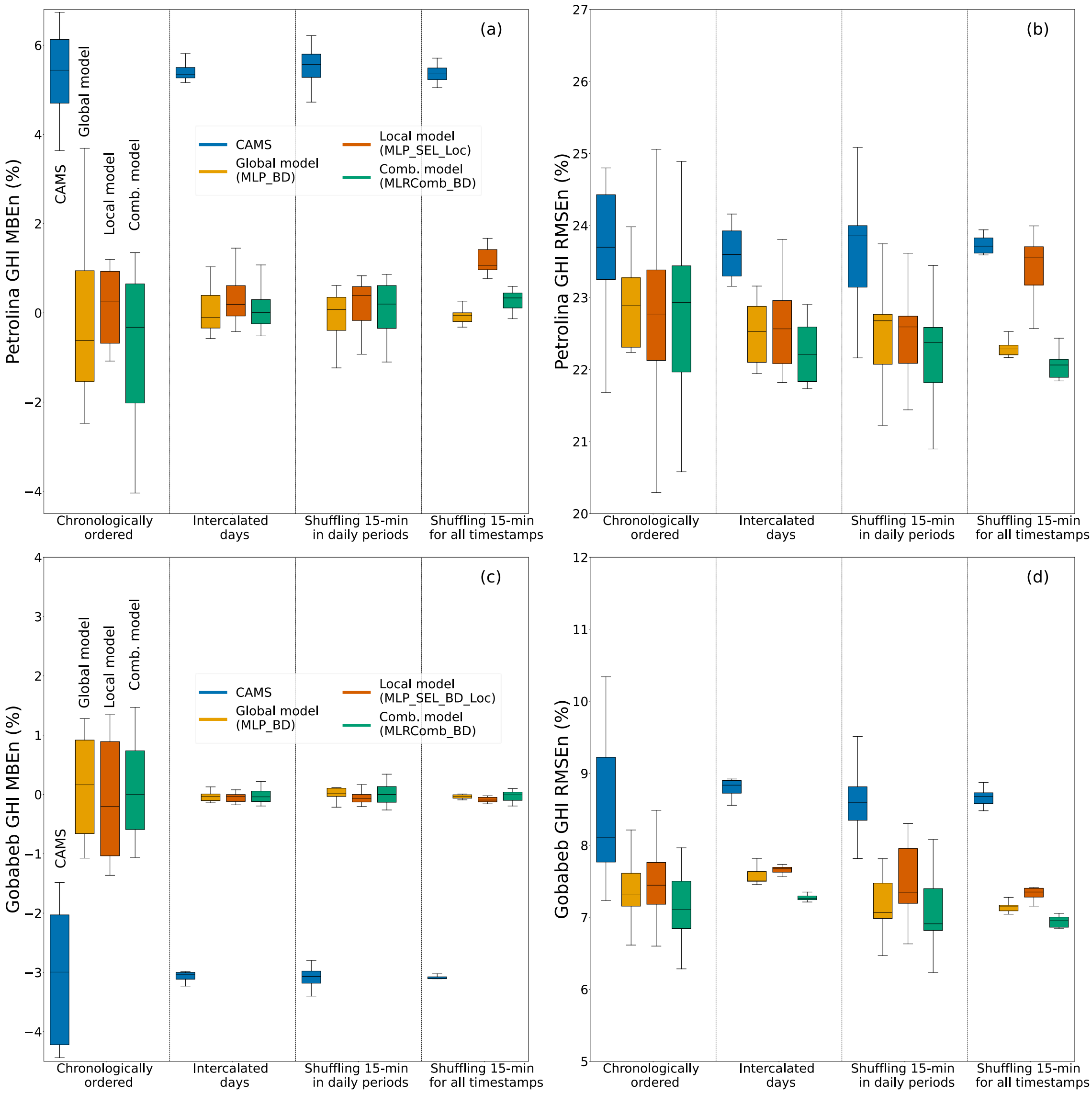


Fig. 9. Results for the best global, local, and combination models for GHI considering different data split strategies in Petrolina and Gobabeb stations (15-min resolution).

Tab. 9 presents the splits in which the more accurate performance is observed, following the criteria of Tab. 8 (last row), and the respective percentage of data in calibration, combination, and test sets. These more accurate splits use the shuffling strategy and have the major percentage of data for calibration.

Tab. 9. Split with the more accurate results and respective percentage of data used in the models.

Stations	Splits	Calibration	Combination	Test
Gobabeb	33	80%	10%	10%
Petrolina (DNI and DHI)	34	70%	20%	10%
Salta and Petrolina (GHI)	36	50%	40%	10%
El Rosal	40	60%	10%	30%

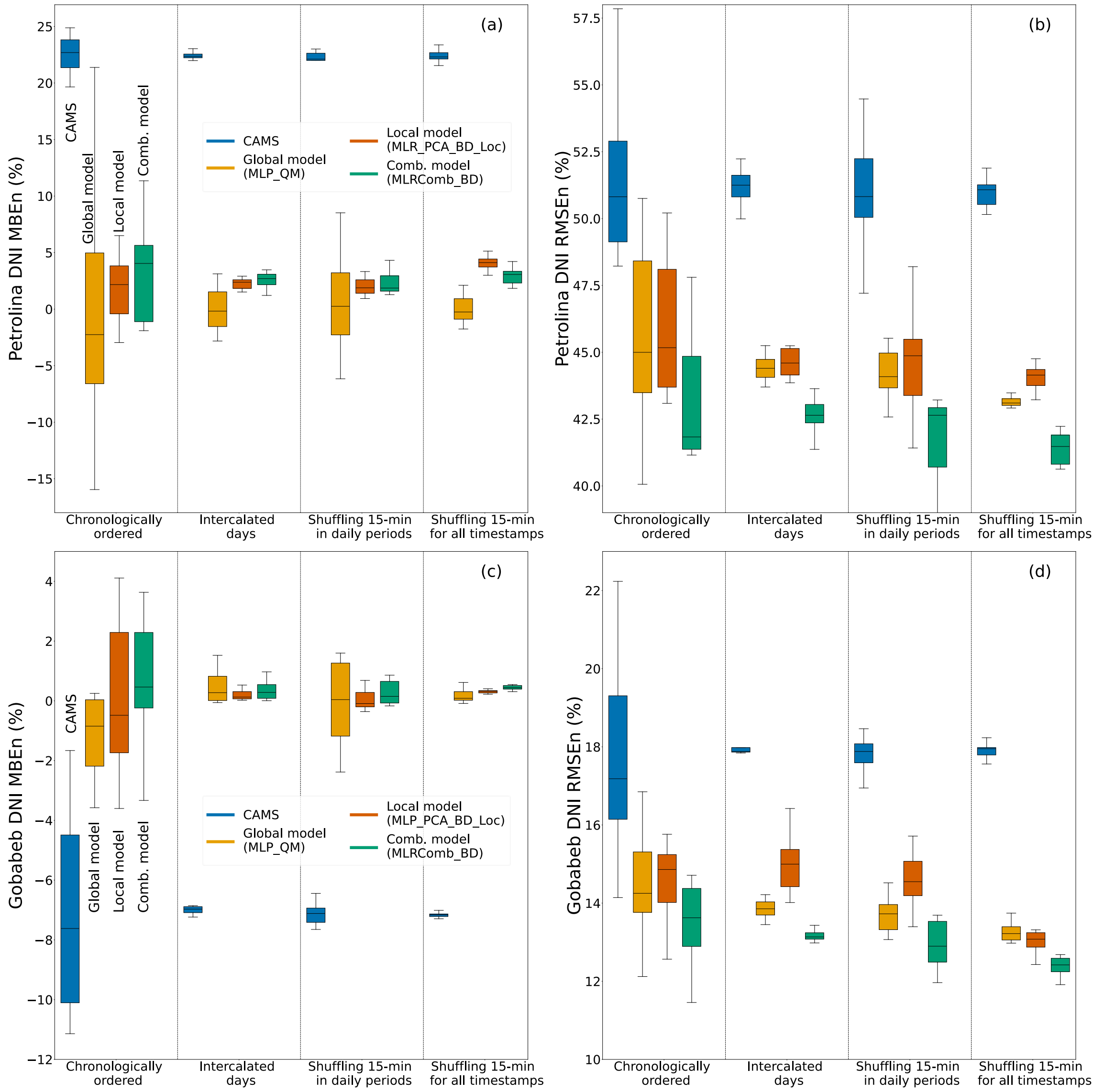
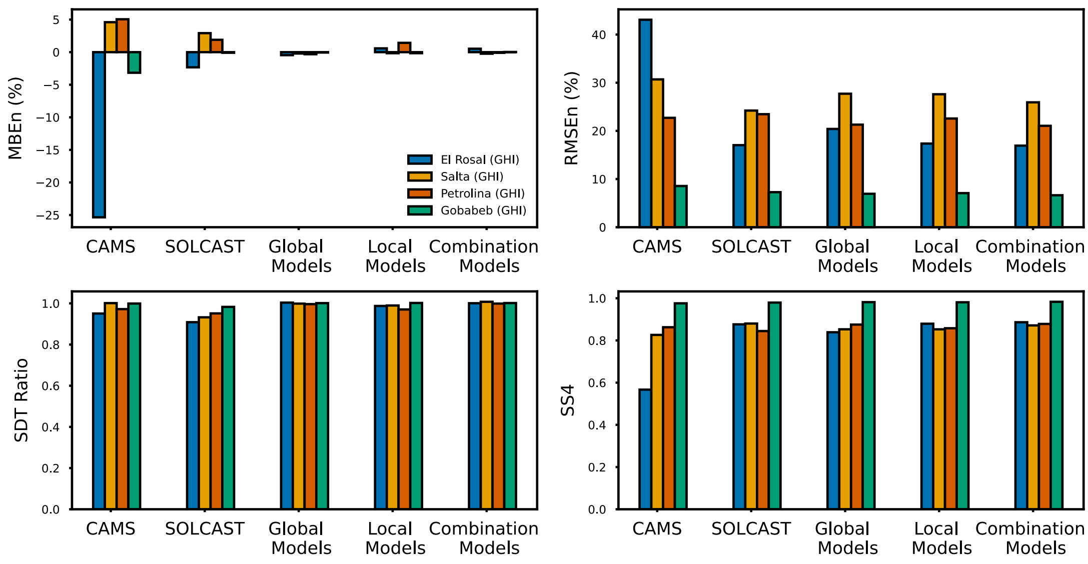
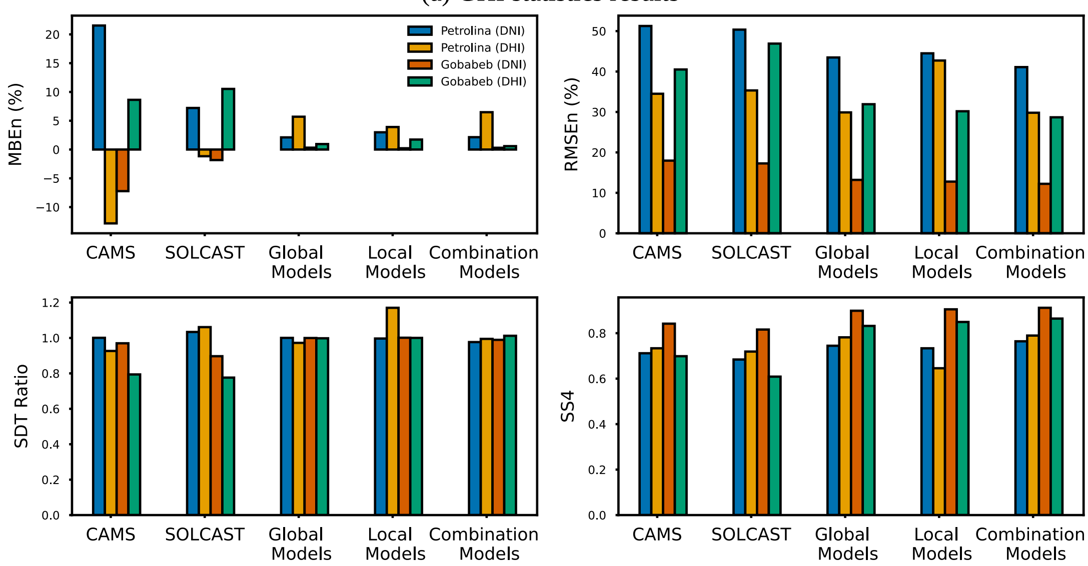


Fig. 10. Results for the best global, local, and combination models for DNI considering different data split strategies in Petrolina and Gobabeb stations (15-min resolution).

Tab. 10 and Fig. 11 show the statistical results comparing original CAMS and Solcast estimates with the best site-adaptation method for each model. The reported metrics in Tab. 10 also shows 95% Confidence Intervals (CI), which were estimated using a non-parametric bootstrap approach where the observations and model results were 2000 times resampled (with replacement). The 2.5th and 97.5th percentiles of each MBEn and RMSEn bootstrap distributions were used to estimate the 95% CI. For El Rosal, the combination model achieves an outstanding result when compared to the CAMS model, with a decrease in RMSEn of 26.2% and an increase in SS4 from 0.57 (CAMS) to 0.89 (combination model). In Salta station, the site adaptation models present better results than CAMS; however, the improvement is lower than El Rosal station. Petrolina and Gobabeb are the sites where the site-adaptation obtains the lower relative gain in RMSEn for GHI (1.7% and 2.0% in RMSEn reduction compared with CAMS, respectively). Although the absolute gain is lower in Gobabeb, the starting RMSEn value is quite low, so it is relatively more difficult to obtain an improvement. In the other metrics (STDRatio, correlation, and SS4), Gobabeb has the lowest GHI site-adaptation gain. Comparing these two stations, the DNI site-adaptation improvements are higher than for CAMS DNI model in all metrics. Combination models provide the best performance in most metrics for both solar radiation magnitudes. For completeness, results for diffuse horizontal irradiance (DHI) in the case of Petrolina and Gobabeb stations are shown, which were obtained by the difference between the global and direct horizontal plane irradiance. The more accurate results for DHI are also in the combination models, which present lower RMSEn and higher correlations, especially for Gobabeb, where it was possible to decrease the RMSEn by 11.8% compared to the CAMS model. These results are also compared with Solcast model, which considers GOES-East satellite images over South America (not the MSG satellite images, as the CAMS model), therefore providing a better SSI estimation for this continent. For GHI, the best site adaptation and Solcast have similar MBEn (on average, 0.1% and 0.6%, respectively) and RMSEn (on average, 17.6% and 18%, respectively). In change, DNI and DHI site adaptation results are more accurate than Solcast model, with an average decrease in RMSEn of 7.2% and 11.9%, respectively. The MBEn of the best site adaptation model is also closer to zero than Solcast, except DHI at Petrolina station. Scatter plots in the test set between the ground measurements against CAMS dataset and the best site-adaptation method are shown in Appendix B for GHI, DNI, and DHI.



(a) GHI statistics results



(b) DHI and DNI statistics results

Fig. 11. Results for the best global, local, and combination models compared with CAMS and Solcast models (15-min resolution).

Tab. 10. Site adaptation results for the more accurate split in the test set (15-min resolution).

Station	Models	MBE in W/m² (MBEn)	MBEn 95% CI	RMSE in W/m² (RMSEn)	RMSEn 95% CI	SS4
El Rosal (GHI) Split 40	CAMS	-166.7 (-25.4%)	[-25.9, -24.8]	283.2 (43.1%)	[42.5, 43.6]	0.57
	SOLCAST	-15.3 (-2.3%)	[-2.6, -2.1]	112.1 (17.0%)	[16.6, 17.5]	0.88
	Global (MLP_QM)	-3.0 (-0.5%)	[-0.8, -0.1]	134.1 (20.4%)	[19.9, 20.9]	0.84
	Local (MLR_PCA_Loc)	3.8 (0.6%)	[0.3, 0.9]	114.2 (17.4%)	[16.9, 17.9]	0.88
	Comb. (MLRComb_BD)	3.5 (0.5%)	[0.3, 0.8]	111.4 (16.9%)	[16.5, 17.4]	0.89
Salta (GHI) Split 36	CAMS	19.8 (4.6%)	[3.6, 5.6]	132.4 (30.7%)	[29.4, 32.0]	0.83
	SOLCAST	12.6 (2.9%)	[2.1, 3.7]	104.5 (24.2%)	[23.1, 25.3]	0.88
	Global (MLP_BD)	-0.9 (-0.2%)	[-1.1, 0.7]	119.6 (27.7%)	[26.5, 29.1]	0.85
	Local (MLR_PCA_Loc)	-0.7 (-0.2%)	[-1.1, 0.7]	119.1 (27.6%)	[26.2, 29.1]	0.85
	Comb. (MLRComb_BD)	-1.0 (-0.2%)	[-1.1, 0.6]	111.9 (25.9%)	[24.5, 27.2]	0.87
Petrolina (GHI) Split 36	CAMS	25.8 (5.0%)	[4.4, 5.7]	116.3 (22.7%)	[22.0, 23.4]	0.86
	SOLCAST	9.7 (1.9%)	[1.2, 2.6]	120.2 (23.5%)	[22.7, 24.3]	0.84
	Global (MLP_BD)	-1.6 (-0.3%)	[-0.9, 0.3]	109.1 (21.3%)	[20.6, 22.0]	0.88
	Local (MLP_SEL_Loc)	7.3 (1.4%)	[0.7, 2.1]	115.6 (22.6%)	[21.8, 23.3]	0.86
	Comb. (MLRComb_BD)	-0.5 (-0.1%)	[-0.7, 0.5]	107.8 (21.1%)	[20.4, 21.7]	0.88
Petrolina (DNI) Split 34	CAMS	84.9 (21.5%)	[20.2, 22.9]	202.1 (51.3%)	[49.9, 52.7]	0.71
	SOLCAST	28.5 (7.2%)	[5.8, 8.7]	198.6 (50.4%)	[48.9, 51.8]	0.68
	Global (MLP_QM)	8.3 (2.1%)	[0.9, 3.4]	171.4 (43.5%)	[42.2, 44.8]	0.74
	Local (MLR_PCA_BD_Loc)	11.8 (3.0%)	[1.7, 4.3]	175.5 (44.5%)	[43.1, 45.8]	0.73
	Comb. (MLRComb_BD)	8.5 (2.1%)	[1.0, 3.3]	162 (41.1%)	[39.8, 42.4]	0.76
Petrolina (DHI) Split 34	CAMS	-27 (-12.8%)	[-13.7, -11.9]	72.7 (34.5%)	[33.4, 35.5]	0.73
	SOLCAST	-2.4 (-1.2%)	[-2.2, -0.2]	74.4 (35.3%)	[34.2, 36.4]	0.72
	Global (MLP)	12.0 (5.7%)	[4.8, 6.6]	63.0 (29.9%)	[29.0, 30.8]	0.78
	Local (MLR_PCA_BD_Loc)	8.2 (3.9%)	[2.7, 5.2]	90.0 (42.7%)	[41.1, 44.4]	0.65
	Comb. (MLRComb)	13.7 (6.5%)	[5.6, 7.3]	62.8 (29.8%)	[28.9, 30.8]	0.79
Gobabeb (GHI) Split 33	CAMS	-19.1 (-3.2%)	[-3.3, -3.0]	51.8 (8.6%)	[8.2, 8.9]	0.98
	SOLCAST	-0.6 (-0.1%)	[-0.2, 0.1]	44.0 (7.3%)	[7.0, 7.6]	0.98
	Global (MLP_BD)	-0.4 (-0.1%)	[-0.2, 0.1]	42.0 (6.9%)	[6.6, 7.3]	0.98
	Local (MLP_SEL_BD_Loc)	-1.0 (-0.2%)	[-0.3, 0.0]	42.9 (7.1%)	[6.7, 7.5]	0.98
	Comb. (MLRComb_BD)	0.2 (0.0%)	[-0.1, 0.2]	40.2 (6.6%)	[6.3, 7.0]	0.98
Gobabeb (DNI) Split 33	CAMS	-52.8 (-7.2%)	[-7.6, -6.9]	131.1 (18.0%)	[17.5, 18.4]	0.84
	SOLCAST	-13.3 (-1.8%)	[-2.2, -1.5]	126.1 (17.3%)	[16.7, 17.8]	0.82
	Global (MLP_QM)	2.4 (0.3%)	[0.0, 0.6]	96.4 (13.2%)	[12.7, 13.7]	0.90
	Local (MLP_PCA_BD_Loc)	1.6 (0.2%)	[0.0, 0.5]	93.3 (12.8%)	[12.2, 13.4]	0.90
	Comb. (MLRComb_BD)	2.2 (0.3%)	[0.1, 0.6]	89.3 (12.2%)	[11.7, 12.7]	0.91
Gobabeb (DHI) Split 33	CAMS	10 (8.6%)	[7.8, 9.4]	46.7 (40.5%)	[39.1, 42.0]	0.70
	SOLCAST	12.1 (10.5%)	[9.6, 11.4]	54.1 (46.9%)	[45.0, 48.8]	0.61
	Global (MLP_BD)	1.1 (1.0%)	[0.3, 1.6]	36.8 (31.9%)	[30.7, 33.1]	0.83
	Local (MLP_PCA_Loc)	2.0 (1.7%)	[1.1, 2.3]	34.8 (30.2%)	[29.0, 31.4]	0.85
	Comb. (MLRComb_BD)	0.7 (0.6%)	[0.0, 1.2]	33.1 (28.7%)	[27.6, 29.8]	0.86

The statistical results presented so far consider a 15-min time resolution. To allow a direct comparison with other studies, additional data from Gobabeb station was used to further test the site adaptation model performance. Fernández-Peruchena et al. (2020) proposed a site-adaptation based on a combination of regression models, applied under clear and cloudy sky conditions, followed by a quantile mapping correction. They report results for Gobabeb station at hourly resolution. Here, the site adaptation model MLRComb_BD, trained using the split 33 (which adopts the randomization split strategy), was applied for the additional years of 2012-2014 and 2021 for the Gobabeb station. Tab. 11 shows that the current procedure reduces the DNI RMSEn in 7% when compared to the CAMS model (hourly resolution), whereas Fernández-Peruchena et al. (2020) reported a 4% reduction.

Tab. 11. Site adaptation results for DNI in Gobabeb (for the unseen years of 2012-2014 and of 2021).

Temporal resolution/Reference	Model	MBEn	RMSEn	ρ
15 min/Current work	CAMS	-9.2%	19.4%	0.91
	MLRComb_BD	-0.2%	13.1%	0.95
Hourly/Current work	CAMS	-9.1%	17.6%	0.92
	MLRComb_BD	-0.1%	10.7%	0.96
Hourly/Fernández-Peruchena et al. (2020)*	CAMS	-11.0%	20.0%	0.91
	SiteAdapted DNI	0.1%	16.0%	0.91

*Values were visually estimated according to the Fig. 5 of the cited work

The accuracy of the Random Forest model on 15-min resolution was 75.8% for the independent years of 2012-2014 and 2021, which is lower than the previous accuracy of 84.6% reported in Fig. 6. The strategy of shuffling individual 15-min timestamps may lead to optimistic Random Forest accuracy estimates because of the temporal autocorrelation of sky conditions (data leakage). However, when evaluating the site-adaptation results, which is the main focus of this work, the final combination model remains robust when applied to independent years from the Gobabeb station, with similar RMSEn values of 12.2% (Tab. 10) and 13.1% (Tab. 11) for DNI. These results indicate that the site-adaptation framework generalizes well to unseen years despite the optimistic accuracy observed in the intermediate Random Forest classification step. The randomization strategy also provides a more balanced representation of seasonal variability and cloud regimes during model training. As the objective of site adaptation is to derive a stable mapping between ground measurements and satellite-based irradiance estimates, assumed to be approximately stationary over time scales of 10–20 years, the random splitting strategy remains suitable for training the correction models, while chronological validation on independent years provides a more rigorous assessment of operational performance.

4.4 Results of site adaptation per cluster

The previous analysis shows that the site-adaptation methods have the largest gains for GHI at El Rosal and DNI at Gobabeb. This section provides further analysis of the role of introducing clusters in the site adaptation. The more accurate global, local, and combination models were selected per cluster following the rank criteria of Tab. 8. Fig. 12 presents the evaluation for each sky condition for GHI and DNI. Normalised metrics were computed using the mean irradiance of the complete time series rather than the mean irradiance within each sky-condition class to avoid artificially inflated relative errors for low-irradiance classes such as overcast DNI. A major reduction of the MBEn and RMSEn values in the GHI of El Rosal station is achieved in all classes when comparing original CAMS and combination models, except for overcast skies conditions. For the GHI on Salta, the global model presents the lowest MBEn in all clusters, except for the clear sky cluster where the Solcast model has the lowest MBEn. Note that the combination model for Salta has positive and negative biases that balance out in the all-time series evaluation, showing the relevance of inspecting the results per cluster.

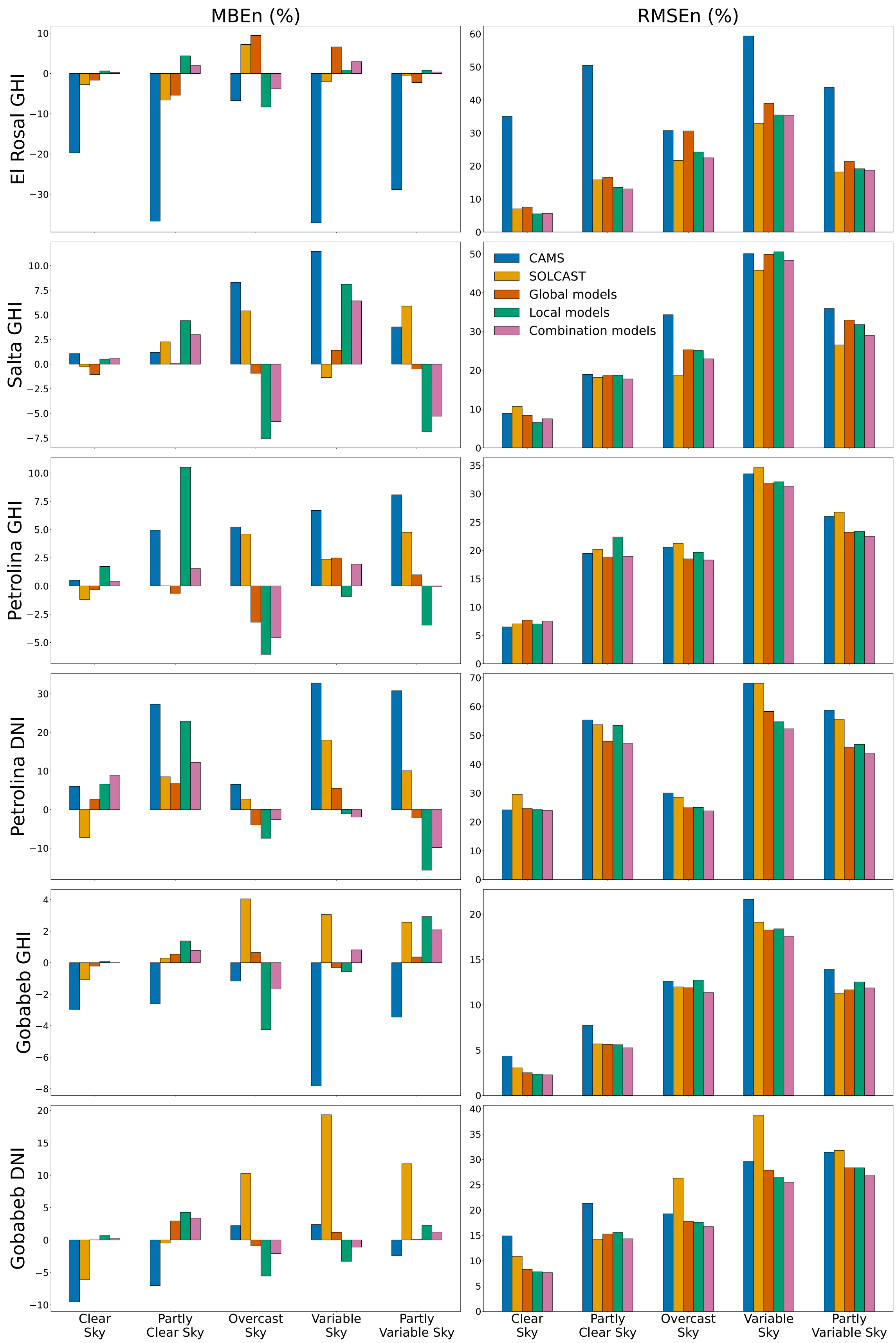


Fig. 12. Results for the best global, local, and combination models compared with CAMS and Solcast models (15-min resolution).

Concerning the RMSEn in Salta, all the models have similar values for the clear and partly clear sky conditions, with the combination and Solcast models having the lowest RMSEn values for the overcast and variable/partly variable sky conditions. For the GHI on Petrolina, global and combination models present the lowest MBEn, except for the partly clear sky and variable conditions. Additionally, global and combination models yield the lowest values for RMSEn in almost all clusters, except for the clear sky cluster, which produces similar results across all models. For the GHI on Gobabeb, the global model shows MBEn closest to zero in all clusters, with similar RMSEn values for the global, local, combination, and Solcast models. For the DNI component in Petrolina and Gobabeb stations, the combination models present the lowest RMSEn values in all clusters. When analysing MBEn, global models have the lowest values in clear conditions for DNI, and the global and combination models show MBEn closer to zero than original CAMS and Solcast for the overcast, variable, and partly variable clusters.

4.5 Monthly evaluation for resource assessment

Long-term surface solar irradiance (SSI) time series are used for solar resource assessment to determine the project bankability. For utility-scale photovoltaic and concentrating solar power (CSP) projects, on-site irradiance measurements of at least one year are commonly used to validate and site-adapt long-term irradiance time series derived from satellite-based or reanalysis models, thereby reducing uncertainty in long-term energy yield assessments (Sengupta et al., 2024). Therefore, the methods presented here aim to obtain an accurate estimation of long-term surface solar irradiance (SSI) time series, which in turn are used to determine solar bankability and financial modeling (P50/P90 estimates). As the BSRN-GOB station used in this study has a data coverage of around 10 years, it can be used to further analyse long-term SSI time series. Fig. 13 presents the monthly time series of GHI, DNI, and DHI on Gobabeb station of accumulated energy values in kWh/m²/month. Ground observations are shown in black, while the CAMS model and the best site adaptation model (MLRComb_BD, split 33, according to Tab. 10) are shown in red and blue respectively. Tab. 12 provided the monthly statistics. CAMS GHI and DNI show an underestimation when compared with the measurements (negative bias), which tend to disappear towards the end of the time series. The same happens with CAMS DHI, but as one overestimation. The site-adapted model largely removes this systematic bias throughout the analysed period for the three SSI components. Comparing the values of Tab. 10 and 12, it is observed that the MBEn for GHI and DNI are similar for both CAMS and site adaptation models. On the other hand, the RMSE reduction is more pronounced on Tab. 12. Monthly aggregation substantially reduces RMSE through temporal averaging, whereas systematic bias is largely preserved. If no site adaptation is applied, the bankability will have a persistent bias which can affect the estimated costs for the solar project. Finally, these results also highlight that the improvements obtained with the randomization split strategy generalize to independent years outside the model-development period (2012–2014 and 2021), supporting that the observed performance gains are not solely a consequence of short-term temporal correlation.

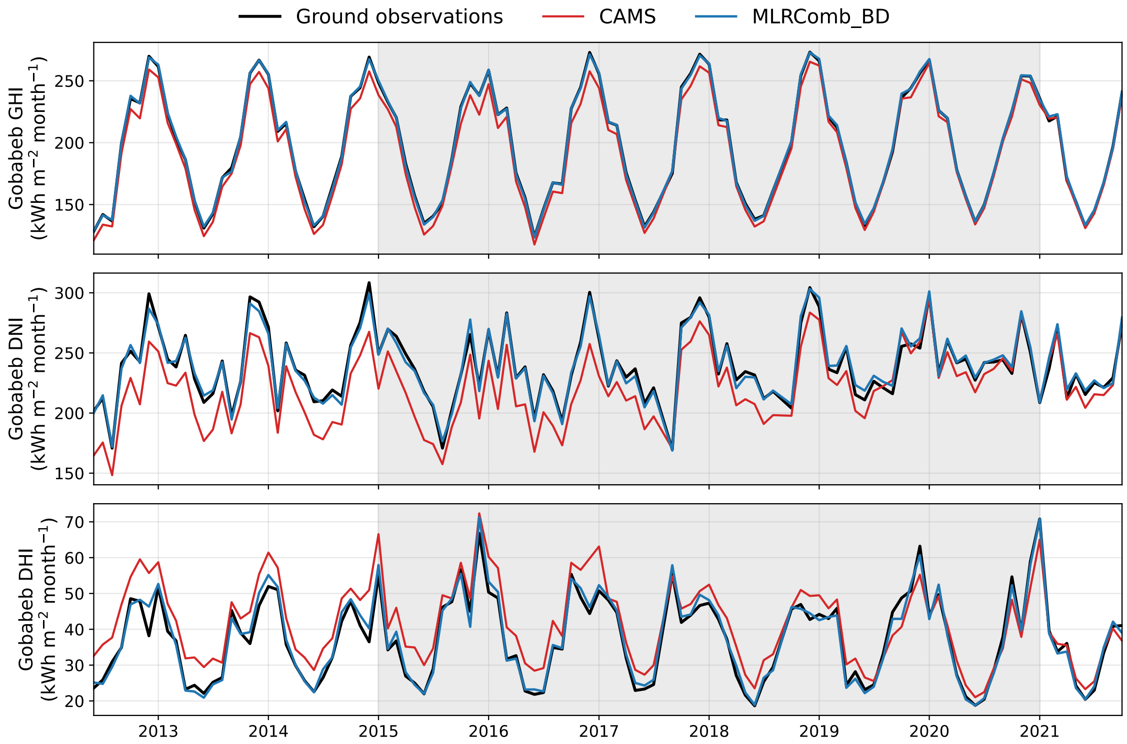


Fig. 13. GHI, DNI, and DHI monthly time-series for Gobabeb station. The grey shade area highlights the years used for site adaptation (2015-2020) in the calibration, combination, and test sets.

Tab. 12. Monthly statistic for GHI, DNI, and DHI in Gobabeb.

Models	MBE in W/m² (MBEn)	RMSE in W/m² (RMSEn)
CAMS GHI	-6.1 (-3.1%)	7.1 (3.6%)
MLRComb_BD GHI	0.3 (0.1%)	1.2 (0.6%)
CAMS DNI	-18.6 (-7.8%)	22.3 (9.3%)
MLRComb_BD DNI	0.3 (0.1%)	4.4 (1.8%)
CAMS DHI	4.5 (11.9%)	6.4 (17.2%)
MLRComb_BD DHI	0.3 (0.9%)	1.9 (5.1%)

5. Conclusions

This work develops site-adaptation techniques based on the introduction of a data classification scheme. This classification is performed using long-term data from CAMS Radiation Service with a statistical procedure trained with the measured data (available short-term data). The site adaptation is tested at four sites in the Southern Hemisphere with different climates, sky conditions, latitudes, and satellite viewing geometry. In particular, two close sites with very different sky conditions and large satellite view angles are tested (El Rosal and Salta). An arid site with mostly clear sky conditions is also included (Gobabeb). The results show that the clusters' introduction allows better modeling of each data cluster's particularity with a global impact on performance. The gain of the site-adaptation procedures depends on the site characteristics, the original quality of the long-term satellite estimates, and the SSI component under consideration. Different types of models applied to the whole time series (global) and to each cluster (local) are developed and tested, as well as their combination. Combination models arise as the best-performing option for most tested metrics in all sites.

The site adaptation of the stations located in Argentina (El Rosal and Salta) shows a great improvement in data quality over the original CAMS GHI estimates. As these stations are located in a region at the edge of the MSG satellite's field of view, the CAMS model may not accurately estimate the radiation attenuation by clouds in this region, especially in the El Rosal Andean site, located at 3355 meters above sea level with a rather high ground albedo. In particular, El Rosal station shows a reduction of 26.2% in the RMSEn when comparing the best performing site-adaptation model (combination) and CAMS models, while for Salta, the reduction in the RMSEn is 4.8%. As El Rosal shows a high occurrence of clear sky conditions (66.2%), and the CAMS model misinterprets the clouds below the station as if they were covering the station due to pixel size and high ground albedo terrain, model errors are increased. The results for GHI at Petrolina and Gobabeb stations also show improvements made by the global, local and combination models, but to a lesser extent. When compared to the Solcast model, the site adaptation models have a slightly better MBEn in all-time series evaluation for GHI, with similar – but better – RMSEn results. This highlights that site adaptation models can correct cloud representation issues caused by the large viewing angle and pixel size of MSG in South America. For DNI and DHI, the MBEn and RMSEn of the global and combination models in all-time series evaluation are similar and more accurate than those of CAMS and Solcast, except for the MBEn in the DHI of Petrolina station with the lowest values presented by Solcast.

The evaluation of the results per sky condition confirms that site-adaptation models substantially reduce biases and errors compared to the CAMS data, with particularly strong improvements for GHI at El Rosal and DNI at Gobabeb. The highest RMSE values are presented in the variable sky condition, highlighting the challenge in modelling solar irradiance under cloud enhancement situations. Global and combination models generally provide the most balanced performance across sites and sky conditions. The combination models consistently achieve the lowest RMSE for DNI, highlighting their robustness under different sky conditions. These results underline the value of cluster-based evaluation, which exposes strengths and limitations that may be hidden in all-sky average assessments.

As the site adaptation of surface solar irradiance is a common practice in the solar industry to improve energy production estimates and support bankability of large solar power plant projects, accurate results are presented here using the freely available CAMS Radiation Service database as input to the machine learning models. Solcast, which uses the best viewing satellite for each region and is a commercial service, is employed as a reference model over South America due to its favorable satellite viewing geometry. The results show that comparable accuracy can be achieved through site adaptation based on CAMS data, offering a cost-free alternative for practitioners.

The proposed site adaptation framework relies on variables provided by the CAMS Radiation Service (expert mode), which ensures its applicability across all regions covered by CAMS, including Europe, Africa, the Middle East, and parts of South America and the Asia–Oceania region. In addition, since the method is based on general atmospheric and irradiance-related variables (e.g. GHI, DNI, DHI, cloud optical depth, aerosol optical depth, total column of water vapor), it can be extended to other satellite-based radiation products, provided that similar input variables (e.g., cloud and atmospheric properties) are available. The methodology is also designed to be adaptable to different climatic conditions, as the classification and model training are site-specific, provided that 1-min GHI observations are available. The sky condition classification derived from ground measurements enables the site adaptation to account for regional variability in atmospheric conditions and cloud dynamics.

From the point of view of global models, the multilayer perceptron (MLP) neural network shows better results than the multilinear regression (MLR), as it uses a non-linear technique to solve the problem globally. As the local models are applied to specific subsets of the time series divided according to the sky conditions, they achieve better statistical results overall than global models. Since the global and local models are not redundant, combining them on the combination set achieves even better results than the individual models. Furthermore, the best results for all the models evaluated are presented by the combination models. Although the performance gain of the combination model over the global and local models is modest, it is a computationally lightweight method.

The evaluation of the different data splits shows that the data splits with randomization present results with less inter-subset variability and achieve higher accuracy than splits maintaining the chronological sequence of the time-series or splits with intercalation of days. Specifically, the random split strategy of all 15-min yields the lowest error metrics. However, because adjacent sky conditions are temporally correlated, this strategy may produce optimistic estimates of the intermediate Random Forest classification accuracy. An independent chronological evaluation performed for the BSRN-GOB station showed a reduction in classification accuracy, indicating the presence of temporal dependence between neighboring samples. Nevertheless, the final site-adaptation model maintained similar performance when applied to independent years, demonstrating that this sub-hourly data leakage does not substantially affect the long-term generalization capability of the proposed framework for resource assessment. As the tradeoff between accuracy, inter-subset variability, and using chronological sequences (which may be reasonable from an operational point of view) is a matter that has not been previously addressed in specific, full results with different subsets are presented here. A detailed assessment of this issue, including an assessment of temporal data leakage in additional stations, is left for future investigations.

Finally, in practical applications, the long-term GHI and DNI corrected time series are commonly used as inputs for solar resource assessment and bankability studies of photovoltaic (PV) and concentrating solar power (CSP) projects. Estimation of PV energy yield typically requires additional variables such as global tilted irradiance (GTI), module temperature, and wind speed. Commercial software packages generally estimate GTI using well-established transposition models that take GHI, DNI, and DHI as inputs. Therefore, the extent to which improvements in irradiance site adaptation translate into improved GTI estimates and PV energy yield remains an important open research question. Since no measured GTI or PV production data were available in this study, this topic is recommended for future investigation.

Appendix A

As a complement to Fig. 9, all Taylor diagrams for the other stations are presented in Figures A.1 to A.7.

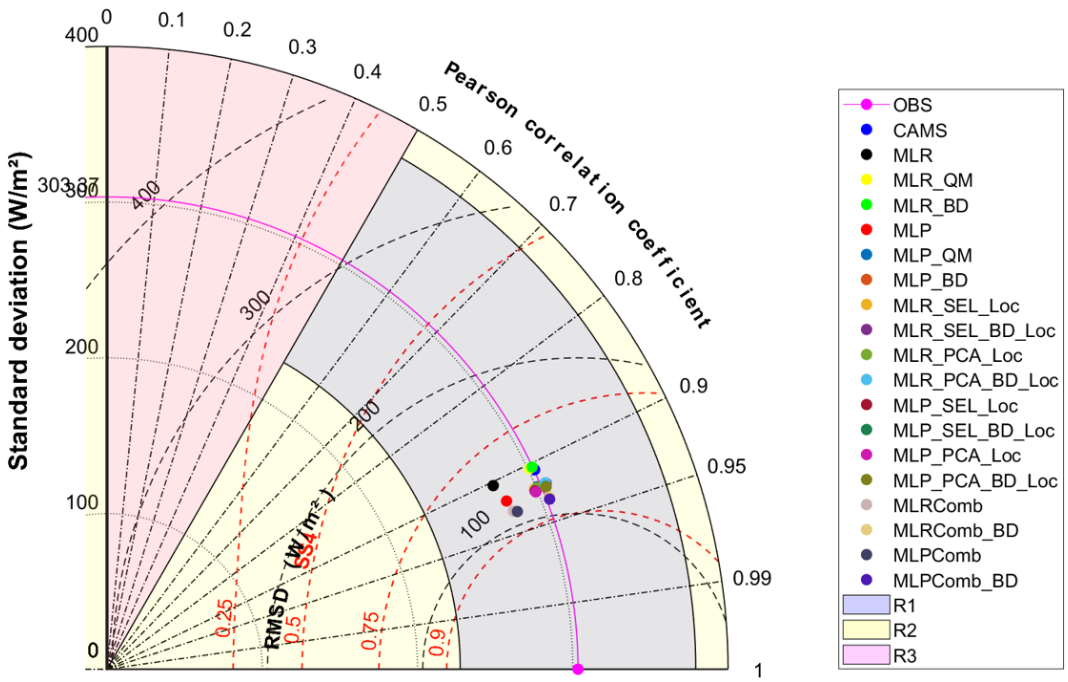


Fig. A.1. Taylor diagram for split 36 in Salta station (GHI, 15-min resolution).

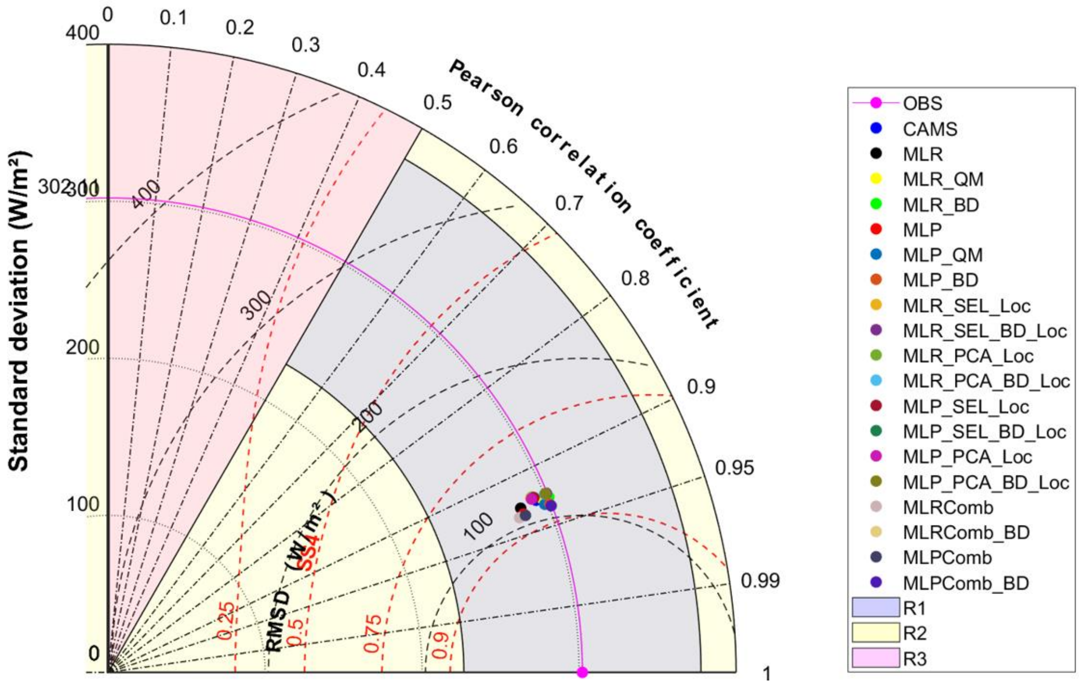


Fig. A.2. Taylor diagram for split 36 in Petrolina station (GHI, 15-min resolution).

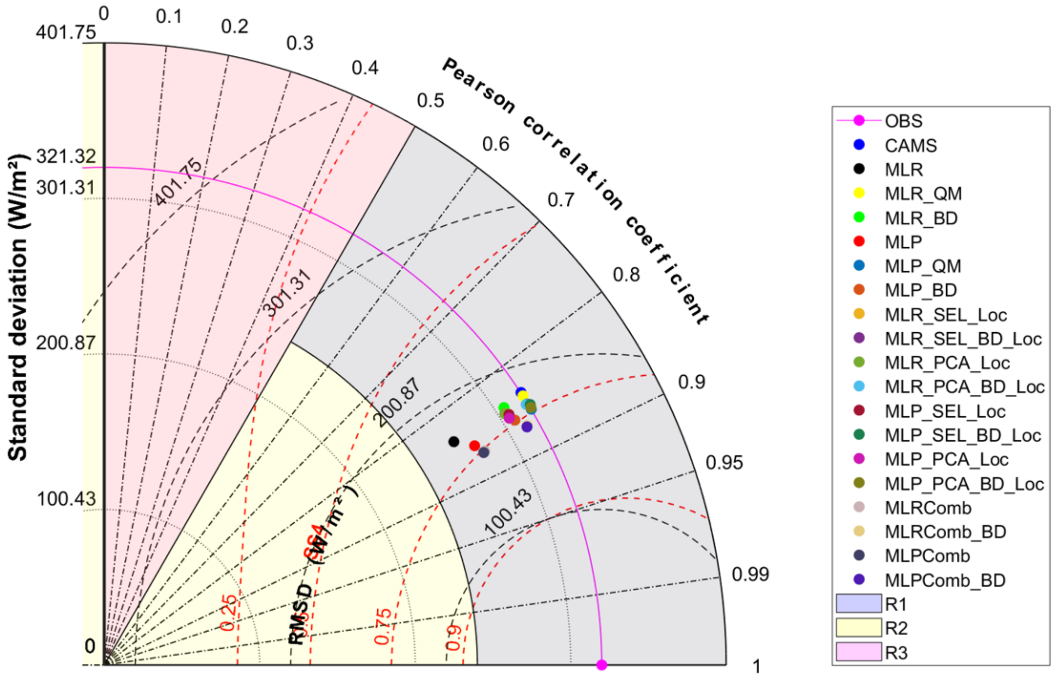


Fig. A.3. Taylor diagram for split 34 in Petrolina station (DNI, 15-min resolution).

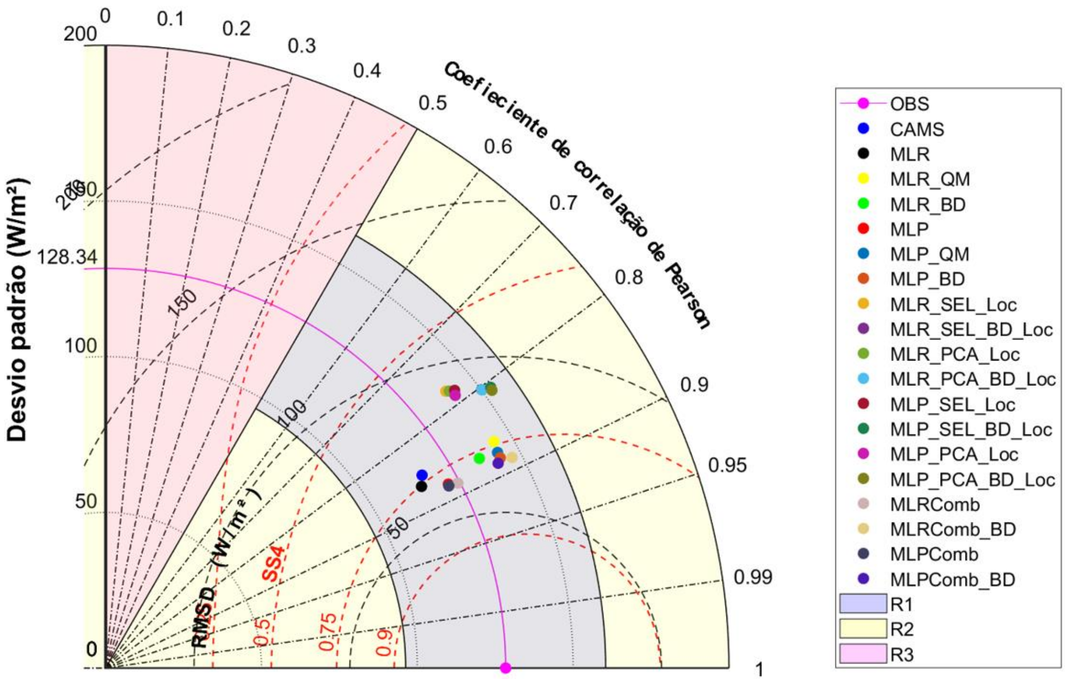


Fig. A.4. Taylor diagram for split 34 in Petrolina station (DHI, 15-min resolution).

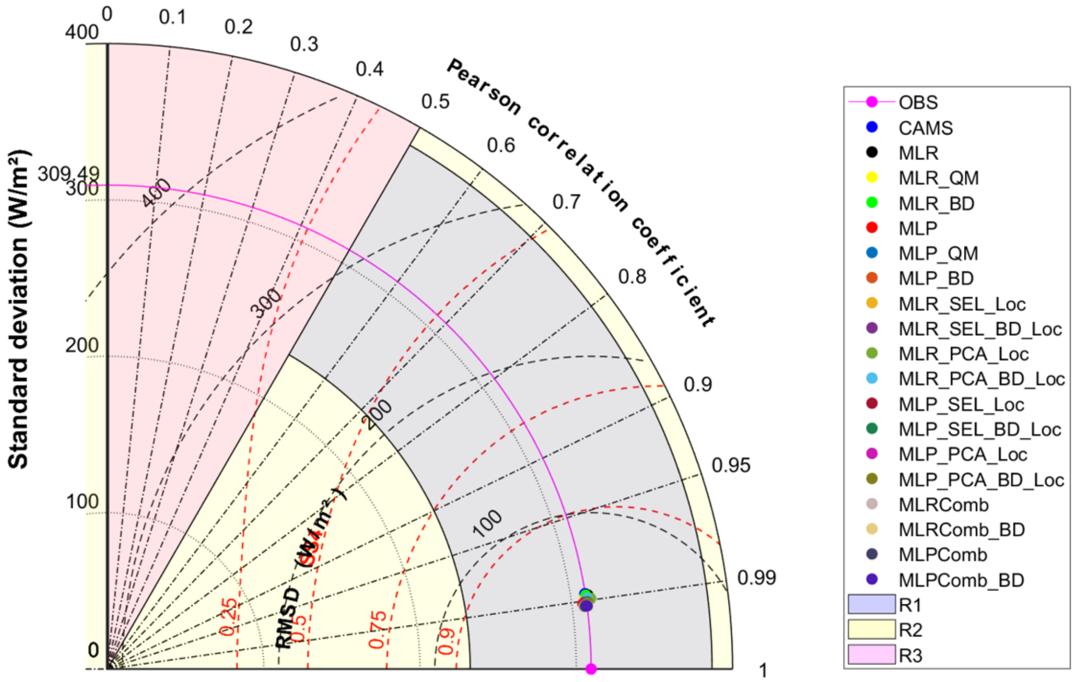


Fig. A.5. Taylor diagram for split 33 in Gobabeb station (GHI, 15-min resolution).

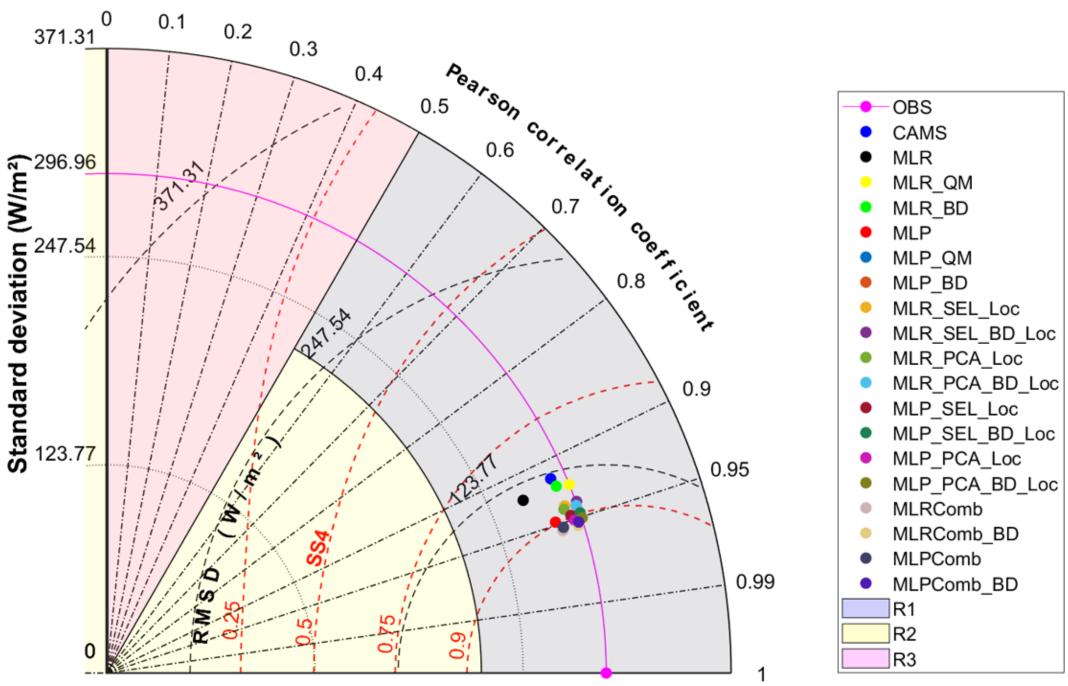


Fig. A.6. Taylor diagram for split 33 in Gobabeb station (DNI, 15-min resolution).

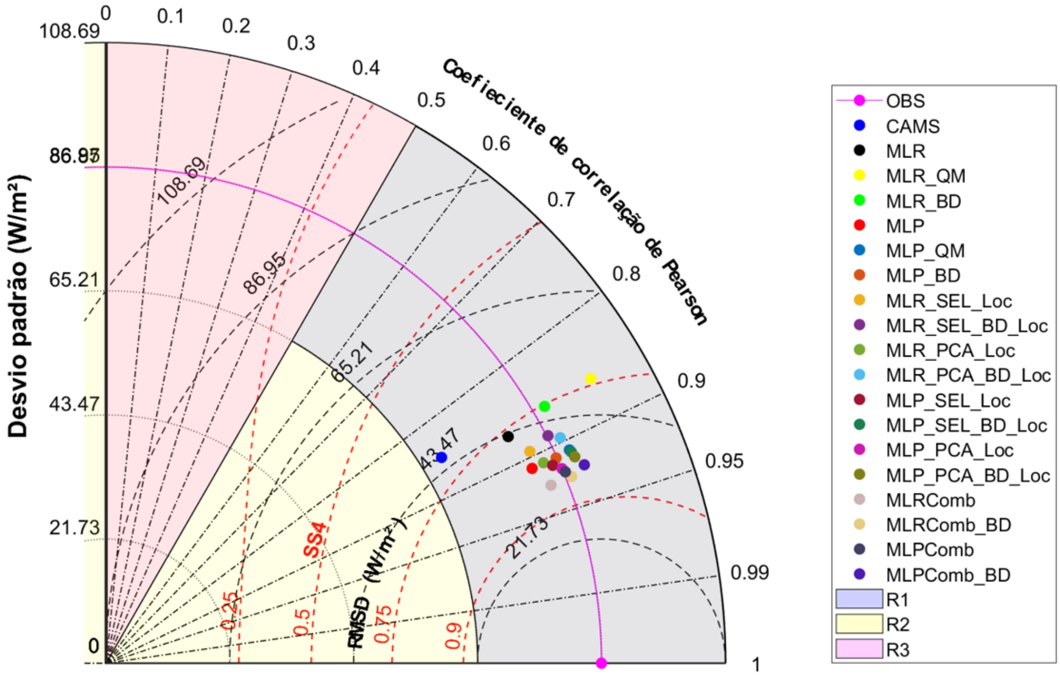


Fig. A.7. Taylor diagram for split 33 in Gobabeb station (DHI, 15-min resolution).

Appendix B

The scatterplots of the statistical results presented in Tab. 10 are shown in Figures B.1 to B8.

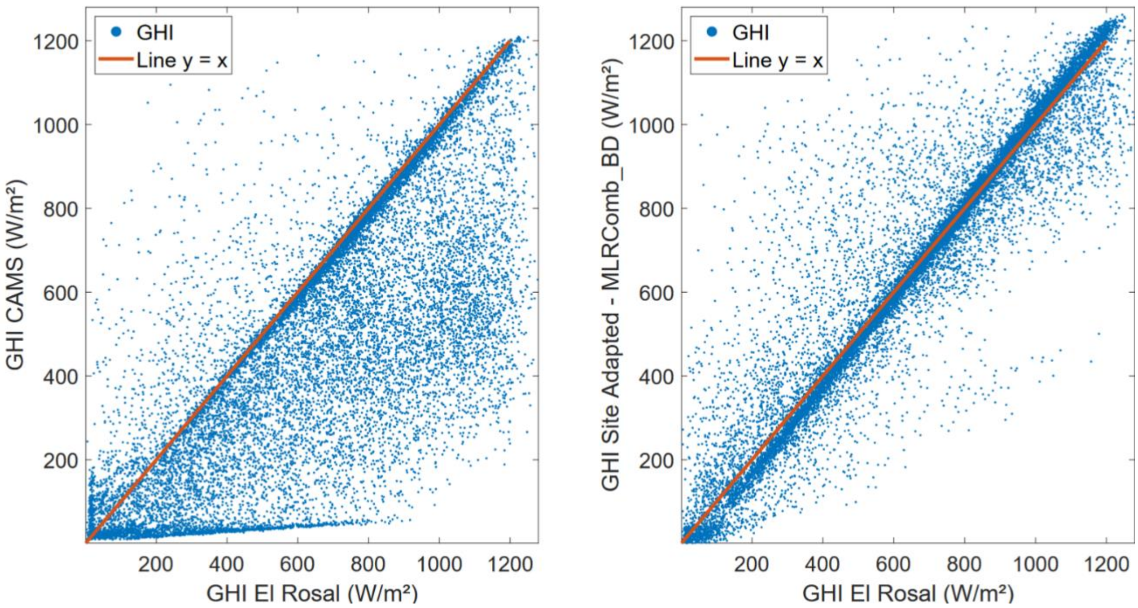


Fig. B.1. Scatterplots for CAMS and site adaptation models compared with GHI measured in El Rosal station (split 40, 15-min resolution).

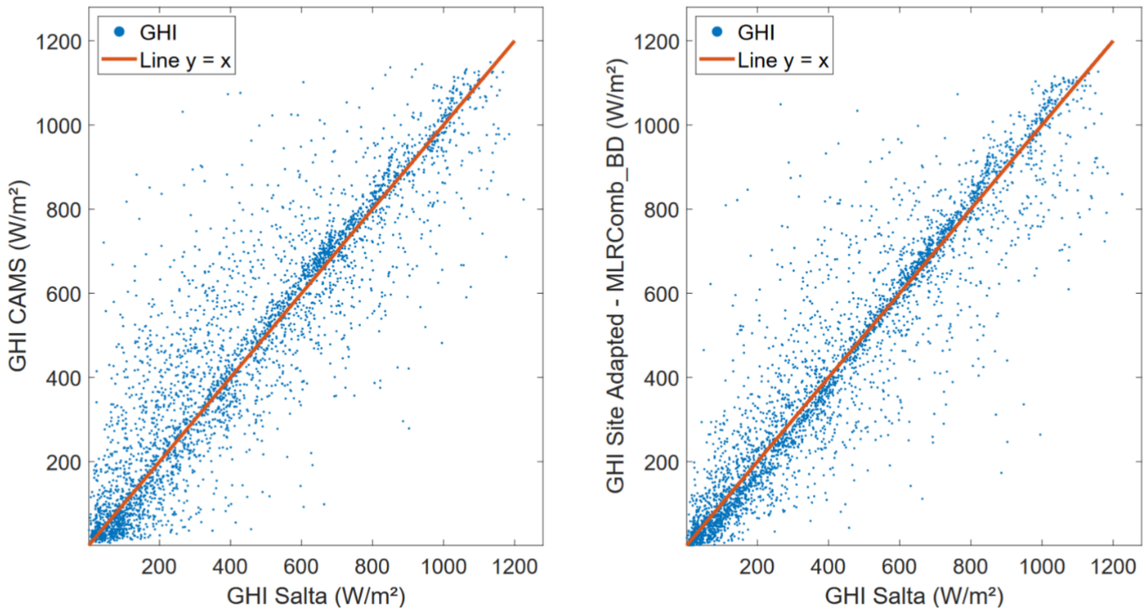


Fig. B.2. Scatterplots for CAMS and site adaptation models compared with GHI measured in Salta station (split 36, 15-min resolution).

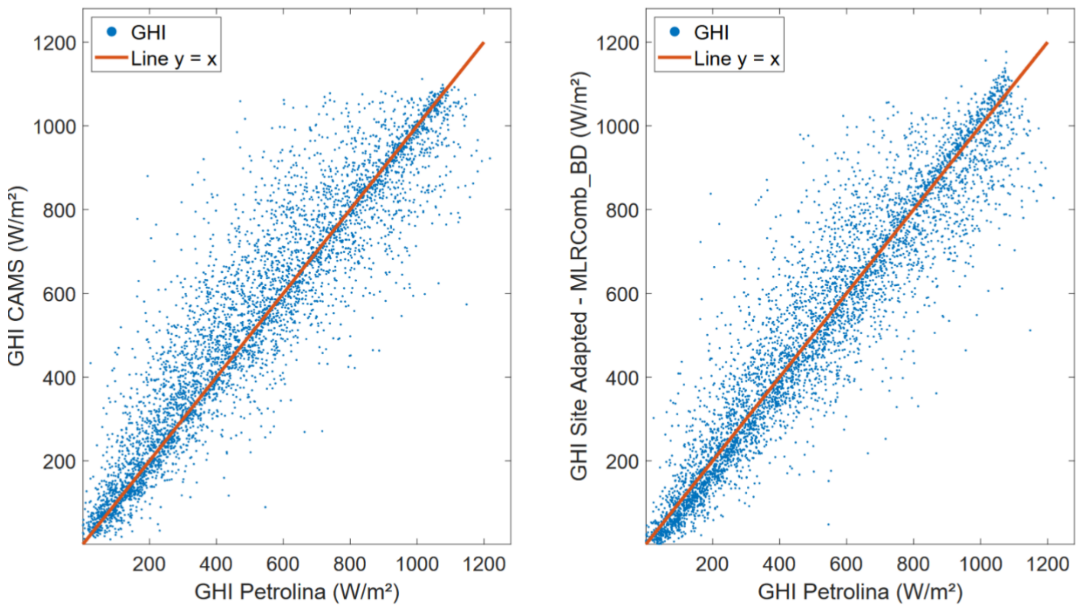


Fig. B.3. Scatterplots for CAMS and site adaptation models compared with GHI measured in Petrolina station (split 36, 15-min resolution).

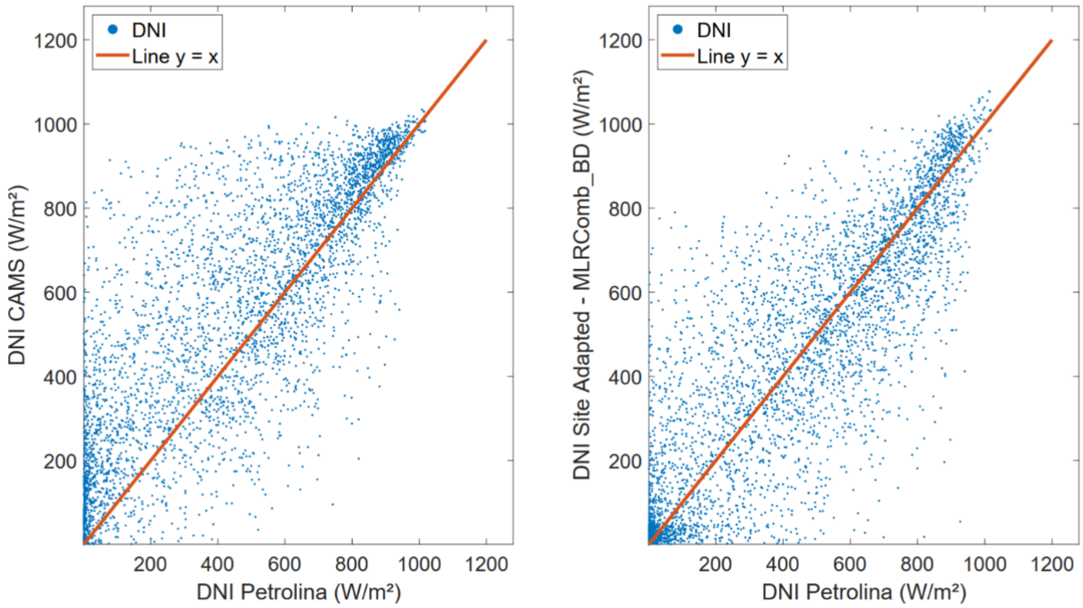


Fig. B.4. Scatterplots for CAMS and site adaptation models compared with DNI measured in Petrolina station (split 34, 15-min resolution).

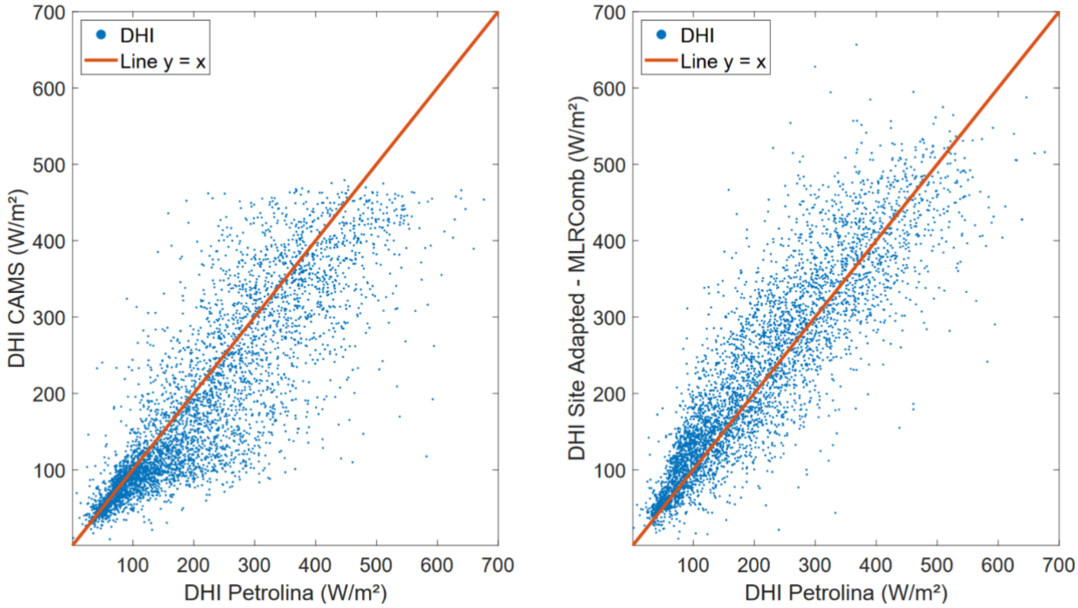


Fig. B.5. Scatterplots for CAMS and site adaptation models compared with DHI measured in Petrolina station (split 34, 15-min resolution).

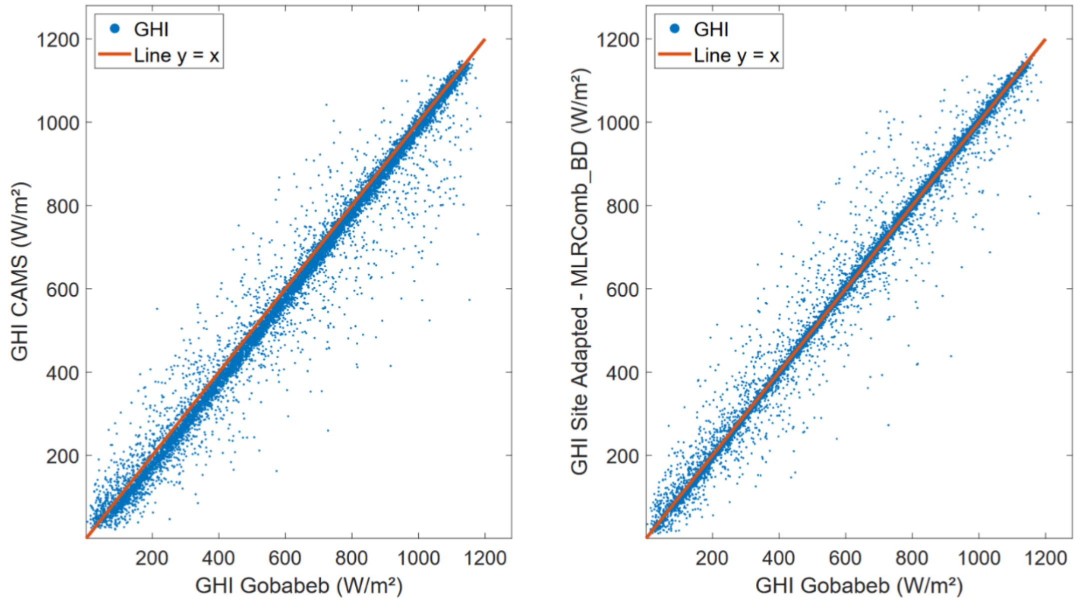


Fig. B.6. Scatterplots for CAMS and site adaptation models compared with GHI measured in Gobabeb station (split 33, 15-min resolution).

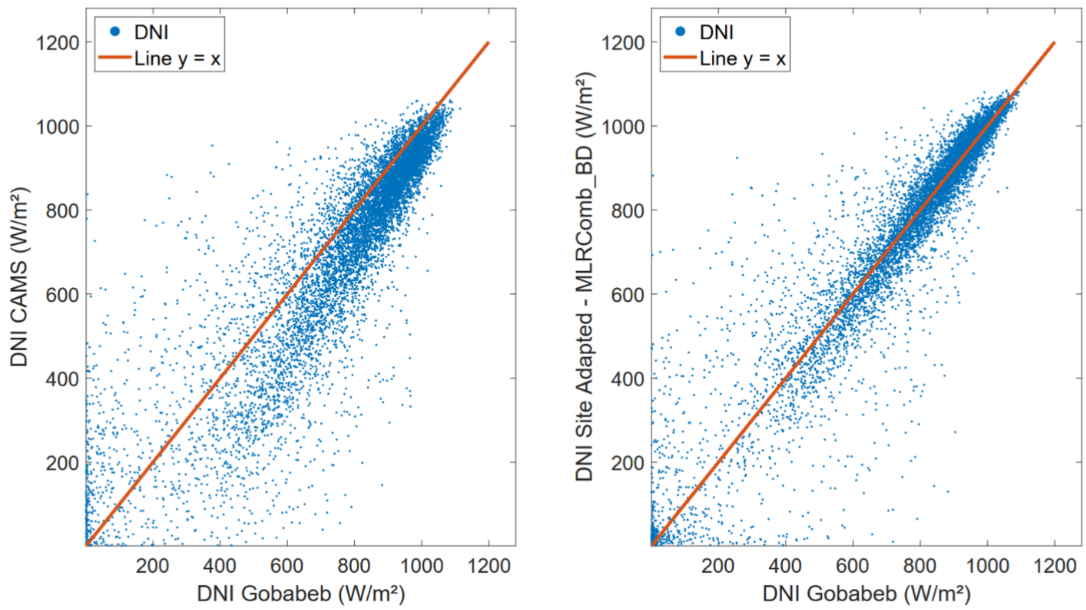


Fig. B.7. Scatterplots for CAMS and site adaptation models compared with DNI measured in Gobabeb station (split 33, 15-min resolution).

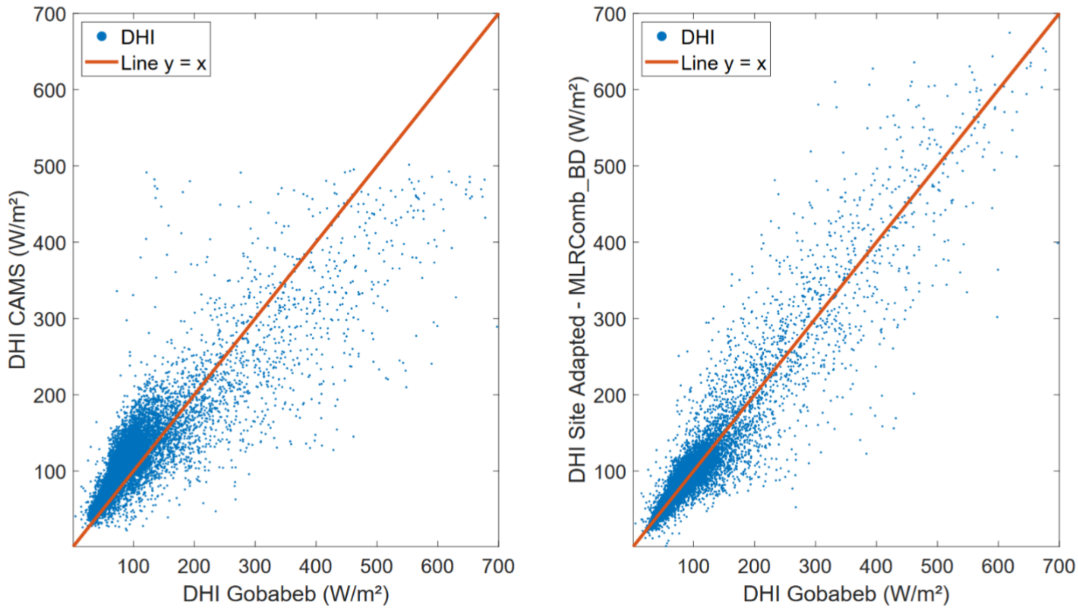


Fig. B.8. Scatterplots for CAMS and site adaptation models compared with DHI measured in Gobabeb station (split 33, 15-min resolution).

Appendix C

This Appendix illustrate the feature selection algorithm (SEL) results in a specific data split and irradiance component. The amount of input variables varies among the clusters according to the thresholds defined on Tab. 5. Tabs. C1 and C2 present the selected features for Gobabeb in split 33 and for Petrolina in split 36, both for GHI. The final selection is station-dependent, as some variables are selected on Petrolina while they are not present in any clusters for Gobabeb (as is the case for skin temperature, wind speed and the fgeo parameter). It is interesting to note that the CAMS-based clearness index (k_{tCAMS}) is presenting in all clusters, which can be expected as the correlation coefficients are obtained with the clearness index of the station (k_t). The example shown on Tabs. C1 and C2 corresponds to the CAMS dataset version available at the time of the experiments, which includes both bias-corrected and non-bias-corrected irradiance variables (NoCorr). Later CAMS versions no longer provide the NoCorr variables separately; however, this difference does not affect the feature-selection methodology or the conclusions regarding the role of feature selection in the proposed site-adaptation framework. BHI corresponds to the Beam Horizontal Irradiance, defined as the DNI multiplied by the cosine of the zenith angle.

Tab. C3 show the feature selection results for the combination models for GHI at the Gobabeb and Petrolina stations (splits 33 and 36, respectively). From the 15 input models (the CAMS model, 6 global models, and 8 local models), 6 variables were selected for Gobabeb and 12 for Petrolina. As the threshold used to remove highly correlated variables was a Pearson correlation coefficient greater than 0.999, a larger number of variables was discarded for Gobabeb than for Petrolina. This suggests that the model outputs are more strongly correlated at Gobabeb, which is consistent with its predominance of clear-sky conditions and lower cloud variability. In contrast, Petrolina exhibits greater variability in cloud conditions, resulting in more heterogeneous model outputs and, consequently, fewer variables exceeding the correlation threshold.

Tab. C1. Feature selection results for the local models in Gobabeb GHI on split 33.

Gobabeb GHI (split 33) - Feature Selection Results for the Local Models				
Clear Sky	**Partly Clear Sky**	**Overcast Sky**	**Variable Sky**	**Partly variable Sky**
k_{tCAMS}	ClearSkyDNI	DHINoCorr	k_{tCAMS}	GHI
DNI	k_{tCAMS}	GHINoCorr	GHINoCorr	BHI
ClearSkyDNI	DNI	GHI	BHI	k_{tCAMS}
TOA	ClearSkyBHI	k_{tCAMS}	DNI	BHINoCorr
ClearSkyDHI	TOA	DHI	BHINoCorr	DNI
DHI	ssrd**	ClearSkyDNI	GHI	ClearSkyBHI
t2m*	BHI	ssrd**	DNINoCorr	ClearSkyDNI
	DNINoCorr	BHI	ClearSkyDNI	TOA
	ClearSkyDHI	ClearSkyBHI	ClearSkyBHI	ssrd**
	DHINoCorr	TOA	ssrd**	DNINoCorr
	DHI	DNI	ClearSkyGHI	ClearSkyDHI
	t2m*	BHINoCorr	TOA	t2m*

*t2m: 2-meter ambient temperature from ERA5-Land
**ssrd: surface solar radiation downwards from ERA5-Land

Tab. C2. Feature selection results for the local models in Petrolina GHI on split 36.

Petrolina GHI (split 36) - Feature Selection Results for the Local Models				
Clear Sky	**Partly Clear Sky**	**Overcast Sky**	**Variable Sky**	**Partly variable Sky**
k_{tCAMS}	ClearskyBNI	k_{tCAMS}	k_{tCAMS}	k_{tCAMS}
ClearskyBNI	TOA	DHInocorr	GHInocorr	GHInocorr
TOA	GHI	GHInocorr	BHI	BHI
ssrd	BHI	GHI	BNI	BHInocorr
skt*	ssrd	ClearskyBNI	BHInocorr	GHI
DHI	BHInocorr	DHI	GHI	BNI
ClearskyDHI	k_{tCAMS}	ws**	BNInocorr	ClearskyBNI
DHInocorr	BNI	ssrd		BNInocorr
t2m	skt*	BHI		ssrd
ws**	BNInocorr	TOA		TOA
fgeo***	ClearskyDHI			skt*
	t2m			
	DHI			
	DHInocorr			

*skt: skin temperature from ERA5-Land
**ws: wind speed from ERA5-Land (obtained with the 10m u-component and 10 m v-component of wind)
***fgeo: MODIS-like Bidirectional Reflectance Distribution Function (BRDF) parameter from CAMS

Tab. C3. Feature selection results for the combination models in Petrolina and Gobabeb for GHI.

Gobabeb GHI (split 33)	Petrolina GHI (split 36)
MLP_SEL_BD_Loc, MLP_BD, MLP_PCA_BD_Loc, MLR_PCA_BD_Loc, MLR_QM, CAMS GHI	MLP, MLP_QM, MLR_BD, CAMS GHI, MLP_PCA_BD_Loc, MLP_SEL_BD_Loc, MLP_PCA_Loc, MLR_SEL_BD_Loc, MLP_SEL_Loc, MLR_PCA_BD_Loc, MLR_SEL_Loc, MLR_PCA_Loc

CRediT authorship contribution statement

Diego Rodrigues de Miranda: Conceptualization, Software, Methodology, Formal analysis, Validation, Visualization, Investigation, Writing – original draft, Writing – review and editing; Olga de Castro Vilela: Supervision, Writing – review & editing, Resources; Germán Ariel Salazar: Data curation, Writing – review & editing; Rodrigo Alonso-Suárez: Writing – review & editing; Alexandre Costa: Resources; Renan Soares Siqueira Costa: Software; Tsang Ing Ren: Conceptualization.

Acknowledgements

Diego Miranda would like to express his gratitude for the support for his MSc thesis by Brazilian research agency CAPES (grant number 88887.492710/2020-00) and the R&D project "Ferramenta operacional para previsão do vento local e geração de usina eólica em escalas horária/diária/mensal baseada em métodos inteligentes para adaptação de parâmetros e reconfiguração de modelos (IBITU.INTELIPREV)", funded by IBITU Energia in the framework of the ANEEL Brazilian R&D Program; Olga Vilela and Alexandre Costa would like to express their gratitude to the São Francisco Hydroelectric Company (CHESF-ELETROBRAS) for their support through the CHESF-CASANOVA project (ANEEL PD-00048-0217) "Sistema inteligente com aerogerador integrado às fontes de energia eólica, solar e storage como plataforma de desenvolvimento visando melhorias contínuas no processo de geração de energia elétrica. Subprojeto P6. Avaliação da qualidade de dados observacionais e previsão operacional a curto e médio prazos da potência de saída da central híbrida eólico-solar". The authors gratefully acknowledge the CAMS Radiation Service for providing the solar radiation modeled time-series within the Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS), available in https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-solar-radiation-timeseries?tab=overview; Solcast for providing the solar radiation time-series for research purposes; and the Baseline Surface Radiation Network (BSRN) for providing the data of Gobabeb station.

6. References

ABREU, F. M.; GUEYMARD, C. A.; CANHOTO, P.; COSTA, M. J. Performance assessment of clear-sky solar irradiance predictions using state-of-the-art radiation models and input atmospheric data from reanalysis or ground measurements. Solar Energy, v. 252, p. 309-321, https://doi.org/10.1016/j.solener.2023.01.051, 2023.

AGGARWAL, C. C. Data Mining: The Textbook. 1ª. ed., Elsevier, 2015.

AGUIAR, L. M.; POLO, J.; VINDEL, J. M.; OLIVER, A. Analysis of Satellite Derived Solar Irradiance in Islands with Site Adaptation Techniques for Improving the Uncertainty. Renewable Energy, 135: 98–107, 2019

ALIMOHAMMADI, S.; HE, D. Multi-Stage Algorithm for Uncertainty Analysis of Solar Power Forecasting. IEEE Power and Energy Society General Meeting, 2016.

ASSUNÇÃO, H. F.; ESCOBEDO, J. F.; OLIVEIRA, A.P. A New Algorithm to Estimate Sky Condition Based on 5 Minutes-Averaged Values of Clearness Index and Relative Optical Air Mass. Theoretical and Applied Climatology, 90(3–4): 235–48, 2007.

BABAR, B.; LUPPINO, L. T.; BOSTRÖM, T.; ANFINSEN, S. N. Random Forest regression for improved mapping of solar irradiance at high latitudes. Solar Energy, v. 198, p. 81-92, ISSN 0038-092X, https://doi.org/10.1016/j.solener.2020.01.034, 2020.

BANGARIGADU, K.; HOOKOOM, T.; RAMGOLAM, Y. K.; KUNE, N. F. Analysis of Solar Power and Energy Variability Through Site Adaptation of Satellite Data With Quality Controlled Measured Solar Radiation Data. ASME. J. Sol. Energy Eng., 143(3): 031008, https://doi.org/10.1115/1.4048546, 2021.

BAYASGALAN, O.; ADIYABAT, A.; OTANI, K.; HASHIMOTO, J.; AKISAWA, A. A High-Resolution Satellite-Based Solar Resource Assessment Method Enhanced with Site

Adaptation in Arid and Cold Climate Conditions. Energies, 17, 6433, https://doi.org/10.3390/en17246433, 2024.

BENDER, G.; DAVIDSON, F.; EICHELBERGER, S.; GUEYMARD, C. The road to bankability: Improving assessments for more accurate financial planning. Solar Conference, Raleigh, Estados Unidos, 2011.

BREIMAN, L. Random Forests. *Machine Learning* 45, 5–32, 2001.

CALBÓ, J.; GONZÁLEZ, J. A.; PEGÈS, D. 2001. A Method for Sky-Condition Classification from Ground-Based Solar Radiation Measurements. Journal of Applied Meteorology, 40(12): 2193–99, 2001.

CANNON, A. J.; SOBIE, S. R.; MURDOCK, T. Q. Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes? Journal of Climate, 28(17): 6938–59, 2015.

CEBECAUER, T.; SURI, M. Site-Adaptation of Satellite-Based DNI and GHI Time Series: Overview and SolarGIS Approach. AIP Conference Proceedings, 1734, 2016.

CHICCO, G.; COCINA, V.; SPERTINO, F. Characterization of Solar Irradiance Profiles for Photovoltaic System Studies through Data Rescaling in Time and Amplitude. Proceedings of the Universities Power Engineering Conference, 2014

DAWSON, C. W.; WILBY, R. L. Hydrological modeling using artificial neural networks. Progress in Physical Geography, v. 25, n. 1, p. 80-108, 2001.

DHATA, E. F.; KIM, C. K.; KIM, H.-G.; KIM, B.; OH, M. Site-Adaptation for Correcting Satellite-Derived Solar Irradiance: Performance Comparison between Various Regressive and Distribution Mapping Techniques for Application in Daejeon, South Korea. Energies, 15, 9010. https://doi.org/10.3390/en15239010, 2022.

DHATA, E. F.; KIM, C. K.; KIM, H-G. Probabilistic approach for site-adaptation and economic performance estimation of a photovoltaic project in South Korea. Energy Strategy Reviews, v. 54, 101487, ISSN 2211-467X, https://doi.org/10.1016/j.esr.2024.101487, 2024a.

DHATA, E.F.; KIM, C.K.; OH, M. ET AL; KIM, H-G. Toward Improved Site-Adaptation for Direct Normal Irradiance: Exploiting Sky-Condition Classification for Improved Regression-Based, Quantile-Based, and Neural Network Models. Asia-Pac J Atmos Sci 60, 231–244, https://doi.org/10.1007/s13143-023-00350-4, 2024b.

DRIEMEL, A.; AUGUSTINE, J.; BEHRENS, K.; COLLE, S.; COX, C.; CUEVAS-AGULLÓ, E.; DENN, F.M.; DUPRAT, T.; FUKUDA, M.; GROBE, H.; HAEFFELIN, M.; HODGES, G.; HYETT, N.; IJIMA, O.; KALLIS, A.; KNAP, W.; KUSTOV, V.; LONG, C.N.; LONGENECKER, D.; LUPI, A.; MATURILLI, M.; MIMOUNI, M.; NTSANGWANE, L.; OGIHARA, H.; OLANO, X.; OLEFS, M.; OMORI, M.; PASSAMANI, L.; PEREIRA, E.B.; SCHMITHÜSEN, H.; SCHUMACHER, S.; SIEGER, R.; TAMLYN, J.; VOGT, R.; VUILLEUMIER, L.; XIA, X.; OHMURA, A.; KÖNIG-LANGLO, G. Baseline Surface Radiation Network (BSRN): structure and data description (1992–2017). Earth Syst. Sci. Data 10 (3), 1491–1501, https://doi.org/10.5194/essd-10-1491-2018, 2018.

DUCHON, C. E.; O'MALLEY, M. S. Estimating Cloud Type from Pyranometer Observations. Journal of Applied Meteorology, 38(1): 132–41, 1999.

FERNÁNDEZ-PERUCHENA, C. M.; POLO, J.; MARTÍN, L.; MAZORRA, L. Site-Adaptation of Modeled Solar Radiation Data: The SiteAdapt Procedure. *Remote Sensing* 12(13): 1–17, 2020.

FORTUNA, LUIGI, GIUSEPPE NUNNARI, AND SILVIA NUNNARI. A New Fine-Grained Classification Strategy for Solar Daily Radiation Patterns. *Pattern Recognition Letters* 81: 110–17, 2016.

GASTÓN-ROMEO, M.; LEON, T.; MALLOR, F.; RAMÍREZ-SANTIGOSA, L. 2011. A Morphological Clustering Method for Daily Solar Radiation Curves. *Solar Energy* 85(9): 1824–36, 2011.

GAVIN, H. P. The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems. Department of Civil and Environmental Engineering Duke University, 2020.

GSCHWIND, B.; WALD, L.; BLANC, P.; LEFÈVRE, M.; SCHROEDTER-HOMSCHEIDT, M.; AROLA, A. Improving the McClear model estimating the downwelling solar radiation at ground level in cloud-free conditions – McClear-V3. Meteorologische Zeitschrift, 2019.

GUEYMARD, C. A.; WILCOX, S. M. Spatial and Temporal Variability in the Solar Resource: Assessing the Value of Short-Term Measurements at Potential Solar Power Plant Sites. *38th ASES National Solar Conference 2009, SOLAR 2009* 5(January 2009): 3026–53, 2009.

GUEYMARD, C. A.; GUSTAFSON, W. T.; ETRINGER, A.; STORCK, P. Evaluation of procedures to improve solar resource assessments: optimum use of short-term data from a local weather station to correct bias in long-term satellite derived solar radiation time series. World Renewable Energy Forum, Denver, CO, v. 3, p. 2092–99, 2012.

HABTE, A.; STOFFEL, T.; PEREZ, R.; MYERS, D.; GUEYMARD, C.; BLANC, P.; WILBERT, S. Overview of Solar Radiation Resource Concepts. Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications: Second Edition, NREL, 2.1-2.22, 2017.

HALL, M. A. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. Conference: Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, 2000.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 3ª. ed., Elsevier, 2012.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(6), 417–441, https://doi.org/10.1037/h0071325, 1993.

HUANG, G.; LI, Z.; LI, X.; LIANG, S.; YANG, K.; WANG, D.; ZHANG, Y. Estimating Surface Solar Irradiance from Satellites: Past, Present, and Future Perspectives. Remote Sensing of Environment, 233, 111371, 2019.

HARROUNI, S., A. GUESSOUM, AND A. MAAFI. 2005. Classification of Daily Solar Irradiation by Fractional Analysis of 10-Min-Means of Solar Irradiance. *Theoretical and Applied Climatology* 80(1): 27–36, 2005.

INES, A.V.M.; HANSEN, J. W. Bias Correction of Daily GCM Rainfall for Crop Simulation Studies. Agricultural and Forest Meteorology, 138(1–4): 44–53, 2006.

JADIDI, H; FIROUZI, A.; RASTEGAR, M. A.; ZANDI, M. Bayesian updating of solar resource data for risk mitigation in project finance. Solar Energy, v. 207, p. 1390-1403, ISSN 0038-092X, https://doi.org/10.1016/j.solener.2020.07.096, 2020.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data Clustering: A Review. 31(3), 2000.

JIMÉNEZ-PÉREZ, P. F.; MORA-LÓPEZ, L. Modeling and Forecasting Hourly Global Solar Radiation Using Clustering and Classification Techniques. *Solar Energy* 135: 682–91, 2016.

JOLLIFFE, I. T. Principal Component Analysis. 2ª. ed.: Elsevier, 1986.

KANG, O. K.; TAM, K. S. A New Characterization and Classification Method for Daily Sky Conditions Based on Ground-Based Solar Irradiance Measurement Data. *Solar Energy* 94: 102–18, 2013.

KLÜSER, L.; KILLIUS, N.; GESEL, G. APOLLO_NG – a probabilistic interpretation of the APOLLO legacy for AVHRR heritage channels. Atmos. Meas. Tech., 8, 4155–4170, https://doi.org/10.5194/amt-8-4155-2015, 2015.

LAGUARDA, A.; GIACOSA, G.; ALONSO-SUÁREZ, R.; ABAL, G. Performance of the site-adapted CAMS database and locally adjusted cloud index models for estimating global solar horizontal irradiation over the Pampa Húmeda. Solar Energy 199:295-307, 2020.

LEDESMA, R.; SALAZAR, G. Machine Learning for Site Adaptation of Satellite-Derived Solar Irradiance in Northwestern Argentina. Journal of Computer Science and Technology, v. 25(2), e10, https://doi.org/10.24215/16666038.25.e10, 2025.

LEFÈVRE, M.; OUMBE, A.; BLANC, P.; ESPINAR, B.; GSCHWIND, B.; QU, Z.; WALD, L.; SCHROEDTER-HOMSCHEIDT, M.; HOYER-KLICK, C.; AROLA, A.; BENEDETTI, A.; KAISER, J. W.; MORCRETTE, E. J. J. McClear: A New Model Estimating Downwelling Solar Radiation at Ground Level in Clear-Sky Conditions. Atmospheric Measurement Techniques, 6(9):2403–18, doi: 10.5194/amt-6-2403-2013, 2013.

LEWIS, C.; STRAUSS, J. M.; RIX, A. J. A solar resource classification algorithm for global horizontal irradiance time series based on frequency domain analysis. J. Renewable Sustainable Energy, 13 (3): 036102, https://doi.org/10.1063/5.0045032, 2021.

LI, D. H. W.; LAM, J. C. 2001. An Analysis of Climatic Parameters and Sky Condition Classification. *Building and Environment* 36(4): 435–45, 2001.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. 233(233): 281–97, 1967.

MAYER, B.; KYLLING, A.: Technical note: The libRadtran software package for radiative transfer calculations - description and examples of use. In: Atmos. Chem. Phys. 5 (7), S. 1855–1877. DOI: 10.5194/acp-5-1855-2005, 2005.

MAYER, B.; KYLLING, A.; EMDE, C.; BURAS, R.; HAMANN, U. libRadtran: library for radiative transfer calculations, Edition 1.0 for libRadtran version 1.5-beta. – http://www.libradtran.org, 2010.

MIRANDA, D. R.; PETRIBÚ, L.; GALDINO, J. J. B.; FURTADO, J. V.; BARBOZA, L., VILELA, O. C.; COSTA, A. C. A.; GOMES, E.; PEREIRA, A.; JATOBÁ, E.; NETO, A. C.; FILHO, J. B. Quality Assurance Procedure for Solar Radiation at Minute Resolution. EuroSun 2022: 14th International Conference on Solar Energy for Buildings and Industry, 25-29 Sep. 2022, Kassel, Germany, doi: 10.18086/eurosun.2022.15.04, 2022.

MIRANDA, D. R.; AZAM, F.; LEZACA, J.; SAINT-DRENAN, Y-M; SCHROEDTER-HOMSCHEIDT, M. Cloud type and aerosol aware solar irradiance variability classification for solar resource assessment. Renewable Energy, v. 263, 125387, ISSN 0960-1481. DOI: https://doi.org/10.1016/j.renene.2026.1253, 2026.

MORENO-TEJERA, S.; SILVA, M. A; SANTIGOSA, L.; LILLO, I. Classification of days according to DNI profiles using clustering techniques. Solar Energy, v. 146, p. 319-333, 2017.

MUSELLI, M.; POGGI, P.; NOTTON, G.; LOUCHE, A. Classification of typical meteorological days from global irradiation records and comparison between two Mediterranean coastal sites in Corsica Island. *Energy Conversion and Management*, *41*(10), 1043–1063, 2000.

NARVAEZ, G.; GIRALDO, L. F.; BRESSAN, M; PANTOJA, A. Machine Learning for Site-Adaptation and Solar Radiation Forecasting. *Renewable Energy* 167(December): 333–42, 2021.

PAGÈS, D.; CALBÓ, J.; GONZÁLEZ, J. A. Using Routine Meteorological Data to Derive Sky Conditions. Annales Geophysicae 21(3): 649–54, 2003.

PEEL, M.C.; FINLAYSON, B.L.; MCMAHON, T.A. Updated world map of the Köppen-Geiger climate classification. Hydrology and Earth System Sciences Discussions, 11:1633-1644, 2007.

PÉREZ-ORTIZ, M.; JIMÉNEZ-FERNÁNDEZ, S.; GUTIÉRREZ, P. A.; ALEXANDRE, H.; HERVÁS-MARTÍNEZ, C.; SALCEDO-SANZ, S. A Review of Classification Problems and Algorithms in Renewable Energy Applications. *Energies* 9(8): 1–27, 2016.

PINTO, G. X. D. A.; NASPOLINI, H. F.; BRAGA, M.; GOMES, A. M. F.; CAMPOS, R. A.; and RÜTHER, R. Enhanced Site Adaptation Method for Reducing Uncertainties in Long-Term Photovoltaic Plant Performance Assessment. ASME. J. Sol. Energy Eng., 148(4): 041003, https://doi.org/10.1115/1.4071572, 2026.

POLO, J.; MARTÍN, L.; VINDEL, J. M. Correcting Satellite Derived DNI with Systematic and Seasonal Deviations: Application to India. *Renewable Energy* 80: 238–43, 2015.

POLO, J.; WILBERT, S.; RUIZ-ARIAS, J.A.; MEYER, R.; GUEYMARD, C.; SÚRI, M.; MARTÍN, L.; MIESLINGER, T.; BLANC, P.; GRANTI, I.; BOLAND, J.; INEICHEN, P.; REMUND, J.; ESCOBAR, R.; TROCCOLI, A.; SENGUPTA, M.; NIELSEN, K.P.; RENNE, D.; GEUDERR, N.; CEBECAUERF, T. Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets. Solar Energy, 132:25–37, 2016.

POLO, J.; FERNÁNDEZ-PERUCHENA, C.; SALAMALIKIS, V.; AGUIAR, L. M.; TURPIN, M.; MARTÍN-POMARES, L.; KAZANTZIDIS, A.; BLANC, P.; REMUND, J. Benchmarking on Improvement and Site-Adaptation Techniques for Modeled Solar Radiation Datasets. Solar Energy 201(March):469–79, 2020.

QU, Z.; OUMBE, A.; BLANC, P.; ESPINAR, B.; GESELL, G.; GSCHWIND, B.; KLÜSER, L.; LEFÈVRE, M.; SABORET, L.; SCHROEDTER-HOMSCHEIDT, M.; WALD, L. Fast Radiative Transfer Parameterisation for Assessing the Surface Solar Irradiance: The Heliosat-4 Method. Meteorol. Z., 26, 33-57, doi: 10.1127/metz/2016/0781, 2017.

RAMGOLAM, Y. K.; BANGARIGADU, K.; AND HOOKOOM, T. A Robust Methodology for Assessing the Effectiveness of Site Adaptation Techniques for Calibration of Solar Radiation Data. ASME. J. Sol. Energy Eng. June 2021; 143(3): 031009, https://doi.org/10.1115/1.4048547, 2020.

RENO, M. E.; HANSEN, C. Identification of Periods of Clear Sky Irradiance in Time Series of GHI Measurements. Renewable Energy, v. 90, p. 520-531, 2016.

RUIZ-ARIAS, J. A.; GUEYMARD, C. A. CAELUS: Classification of sky conditions from 1-min time series of global solar irradiance using variability indices and dynamic thresholds. Solar Energy, v. 263, 111895, https://doi.org/10.1016/j.solener.2023.111895, 2023.

RUIZ-MUNOZ, J. F.; HOYOS-GÓMEZ, L. S. Accurate solar radiation site adaptation: Harnessing satellite data and in situ measurements. J. Renewable Sustainable Energy 1 September 2024, 16 (5): 053703, https://doi.org/10.1063/5.0226782, 2024.

SALAMALIKIS, V.; TZOUMANIKAS, P.; ARGIRIOU, A. A.; KAZANTZIDIS, A. Site adaptation of global horizontal irradiance from the Copernicus Atmosphere Monitoring Service for radiation using supervised machine learning techniques. Renewable Energy 195: 92-106, 2022.

SALAZAR, G; GUEYMARD, C.; GALDINO, J.; VILELA, O. C.; FRAIDENRAICH, N. Solar irradiance time series derived from high-quality measurements, satellite-based models, and reanalyses at a near-equatorial site in Brazil. Renewable and Sustainable Energy Reviews, 117, 109478, 2020.

SARAZOLA, I.; LAGUARDA, A.; CEBALLOS, J.C.; ALONSO-SUÁREZ, R. Benchmarking of modeled solar irradiation data in Uruguay at a daily time scale. IEEE Latin America Transactions 21(9):1040-1048, Special Issue on Sustainable Energy Sources for an Energy Transition, 2023.

SCHROEDTER-HOMSCHEIDT, M.; KOSMALE, M.; SAINT-DRENAN, Y.-M. Classifying direct normal irradiance 1-minute temporal variability from spatial characteristics of geostationary satellite-based cloud observations. Meteorol. Z. 27, 2, 160–179. DOI:10.1127/metz/2018/0875, 2018.

SCHROEDTER-HOMSCHEIDT, M.; AZAM, F.; BETCKE, J.; HANRIEDER, N.; LEFÈVRE, M.; SABORET, L.; SAINT-DRENAN, Y. M. Surface solar irradiation retrieval from MSG/SEVIRI based on APOLLO Next Generation and HELIOSAT 4 methods. Contrib. Atm. Sci./Meteorol. Z., v. 31, no. 6, p. 455 – 476, doi:10.1127/metz/2022/1132, 2022.

SCHROEDTER-HOMSCHEIDT, M.; AZAM, F.; LEZACA, J., HOYER-KLICK, C.; LEFÈVRE, M.; SABORET, L.; PRASHANTH, V. User Guide to the CAMS Radiation Service (CRS) Status December 2024. CAMS2_73_2021SC2_D3.2.1_2024_UserGuide_v1, 2024.

SCHUMANN, K.; BEYER, H. G.; CHHATBAR, K.; MEYER, R. Improving Satellite-Derived Solar Resource Analysis with Parallel Ground-Based Measurements. Proceedings from ISES Solar World Congress, Kasel, Germany, 2011.

SENGUPTA, M.; HABTE, A.; WILBERT, S.; GUEYMARD, C.; REMUND, J.; LORENZ, E.; VAN SARK, W.; JENSEN, A. R. Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications: Fourth Edition. Golden, CO: National Renewable Energy Laboratory, NREL/TP-5D00-88300, https://doi.org/10.2172/2448063, 2024.

SOLCAST, 2024. Global solar irradiance data and PV system power output data. URL https://solcast.com/

SOUBDHAN, T.; EMILION, R.; CALIF, R. Classification of Daily Solar Radiation Distributions Using a Mixture of Dirichlet Distributions. *Solar Energy* 83(7): 1056–63, 2009.

STEIN, J. S.; HANSEN, C. W.; RENO, M. J. The Variability Index: A New and Novel Metric for Quantifying Irradiance and Pv Output Variability. *World Renewable Energy Forum, WREF 2012, Including World Renewable Energy Congress XII and Colorado Renewable Energy Society (CRES) Annual Conferen* 4(May): 2764–70, 2012.

TAHIR, Z. R; AZHAR, M.; BLAN, P.; ASIM, M.; IMRAN, S.; HAYAT, N.; SHAHID, H.; ALI, H. The Evaluation of Reanalysis and Analysis Products of Solar Radiation for Sindh Province, Pakistan. *Renewable Energy* 145: 347–62, 2020.

TAYLOR, K.E. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research, v. 106, p. 7183-7192, 2011.

THEMEßL, M. J.; GOBIET, A.; HEINRICH, G. Empirical-Statistical Downscaling and Error Correction of Regional Climate Models and Its Impact on the Climate Change Signal. Climatic Change 112(2):449–68, 2012.

TIBA, C.; SÁ, M. H; LACERDA; L. F. Methods for Site-Adaptation of Satellite-Based DNI Time Series: Application to Brazilian Northeast. *Proceedings of the ISES Solar World Congress 2019 and IEA SHC International Conference on Solar Heating and Cooling for Buildings and Industry 2019* (Ml): 2275–82. 2019

TRUEBLOOD, C.; COLEY, S.; KEY, T.; ROGERS, L.; ELLIS, A.; HANSEN, C.; PLILPOT, E. PV Measures Up for Fleet Duty. *IEEE Power & Energy Magazine* (March/April): 33–44, 2016.

VERNAY, C.; BLANC, P.; PITAVAL, S. Characterizing measurements campaigns for an innovative calibration approach of the global horizontal irradiation estimated by HelioClim-3. Renewable Energy, v. 57, p. 339–347, 2013.

WILKS, D. S. Statistical methods in the atmospheric sciences. 2ª. ed. [S.l.]: Elsevier, 2013.

YANG, D.; BRIGHT, J.M. Worldwide validation of 8 satellite-derived and reanalysis solar radiation products: A preliminary evaluation and overall metrics for hourly data over 27 years. Solar Energy, 210:3–19, 2020.

YANG, D.; GUEYMARD, C.A. Probabilistic Post-Processing of Gridded Atmospheric Variables and Its Application to Site Adaptation of Shortwave Solar Radiation. Solar Energy, v. 225, p. 427–443, https://doi.org/10.1016/j.solener.2021.05.050, 2021.

YU, L.; LIU, H. Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research 5, p. 1205–1224, 2004.

ZAINALI, S.; YANG, D.; LANDELIUS, T.; CAMPANA, P. E. Site adaptation with machine learning for a Northern Europe gridded global solar irradiance product. Energy and AI, v. 15, 100331, https://doi.org/10.1016/j.egyai.2023.100331, 2024.