Renewable Energy 195 (2022) 92–106

Contents lists available at ScienceDirect

Renewable Energy

journal homepage: www.elsevier.com/locate/renene

Site adaptation of global horizontal irradiance from the Copernicus Atmospheric Monitoring Service for radiation using supervised machine learning techniques

Vasileios Salamalikis *, Panayiotis Tzoumanikas , Athanassios A. Argiriou , Andreas Kazantzidis

Laboratory of Atmospheric Physics, Department of Physics, University of Patras, Patras, 26500, Greece

ARTICLE INFO

Article history:
Received 15 February 2022
Received in revised form
7 June 2022
Accepted 8 June 2022
Available online 11 June 2022

Keywords:
Global horizontal irradiance
CAMS-Rad
Clear sky detection
Site-adaptation
Supervised machine learning algorithms
Goodness-of-fit statistics

ABSTRACT

Satellite and reanalysis-derived solar products have gained great attention due to the inadequate number of radiometric stations worldwide, however, they are associated with considerable uncertainties. This study deals with the ground-based validation of Global Horizontal Irradiance from CAMS radiation service (GHI_{CAMS}) and the application of supervised machine learning algorithms (MLAs) to site-adapt GHI_{CAMS}. The validation of GHI_{CAMS} against measurements shows significant systematic and dispersion errors for all-sky (nMBE = 4.9% and nRMSE = 15.7%) and cloudy conditions (nMBE = 17.6% and nRMSE = 38.8%). Under clear skies, CAMS performs adequately (nMBE <1% and nRMSE <5%). All MLAs lead to reduced errors for the site-adapted irradiances. MBE is improved by more than 50%, accompanied by significant reductions in RMSE for various solar zenith angles and cloud fractions. The best results are revealed for the tree-based MLAs and especially for Random Forests.

1. Introduction

The design of solar energy projects requires long-term, up-to-date, high-quality solar radiation datasets at the finest spatiotemporal resolution. The accurate knowledge of surface solar irradiance (SSI) and its components is crucial for assessing the solar potential in a specific area. The most important advantage of the measured SSI is that it provides high-quality data at an appropriate temporal resolution. However, ground-based observations refer to specific locations and cannot fully characterize the solar potential of the surrounding area, especially over complex terrains. The limited geographical coverage, cost, and difficulties in installing and maintaining radiometric stations enhance the need to develop satellite-derived methods for estimating solar resources. Various satellite solar products with different geographical coverage and temporal resolutions exist [1,2] and several validation studies revealed the discrepancies against observations [3–9]. It is shown that the inaccurate description of aerosols and the spatiotemporal variability of cloudiness may induce significant uncertainties in the modelled products.

The Copernicus Atmospheric Monitoring Service for radiation (CAMS-Rad) has gained significant visibility among the available solar products. The CAMS-Rad service provides solar data at various temporal resolutions, easily retrieved through the Solar Radiation Data (SoDa) website (http://www.soda-pro.com/). Several validation studies examined the performance of CAMS-Rad at global [9,10] or country level [11–15].

The Baseline Surface Radiation Network (BSRN) is the reference database for global-based validations providing high-quality SSI measurements at 1-min temporal resolution. The discrepancies among the CAMS-Rad Global Horizontal Irradiance (GHI_{CAMS}) and the GHI from BSRN depend on the location, the temporal resolution of solar datasets and the periods for comparison. For example [9], based on 21 BSRN stations reported normalized MBE and RMSE (errors divided by the mean observed GHI) ranges of −13.60%–29.66% and 10.99%–44.56%, respectively, with most nMBE values within ±5%. However, lower errors were presented in Ref. [10] using GHI data for the most recent three-year period of fifteen BSRN stations (nMBE: −4.11 – 6.84% and RMSE: 8.23–36.88%).

In Europe, two independent validation studies for GHI_{CAMS} were

* Corresponding author. Laboratory of Atmospheric Physics, Physics Department, University of Patras, Patras, 26500, Greece.
E-mail address: vsalamalik@upatras.gr (V. Salamalikis).

https://doi.org/10.1016/j.renene.2022.06.043
0960-1481/

conducted in the Netherlands [14] and Germany [15]. The performance of GHI_{CAMS} was very good with correlations (R) from 0.94 to 0.97 for GHI and 0.85 to 0.89 for the clearness index (k_t, the ratio of GHI to top-of-atmosphere solar irradiance) in Netherlands. The distance from the shoreline proved to affect nMBE (from −4% to 10%). The underestimation of GHI (or k_t) in coastal locations was explained by the underestimation of cloud-free occurrences. Correlation values between 0.83 and 0.92 were revealed in Germany. In addition, the error magnitude in Germany increased from east to west with over- and underestimation in the southern and northern stations, respectively. nMBE ranged from −11% to 10% with contrasting spatial pattern when compared to the average observed GHI. The standard deviation ranged between 25% and 39%, without any distinct spatial variation. The observed underestimation of GHI from CAMS-Rad was because of the erroneous classification of clear sky conditions and the wrongly assigned cloud optical depth used in the GHI calculations under clear and overcast skies.

Following the discussion above, there are distinguishable errors between modelled and observed irradiances. Such errors become important when the satellite-derived products are considered the primary solar resource in Concentrating Solar Projects (CSP) or photovoltaic (PV) applications. The most critical uncertainty sources were summarized in Ref. [16], including modelling issues under clear sky conditions, uncertainties in input data (cloud and aerosol optical parameters), non-adequate description of terrain effects and ground albedo, as well as the variability of the above factors within the area covered by a single satellite image pixel. For this reason, several “site adaptation” methods have been evaluated for correcting the mismatch between modelled and observed solar irradiances, including Distribution Mapping (DM), Measure Correlate Predict (MCP), Model Output Statistics (MOS) procedures or their combinations [16–29]. The site adaptation techniques can be grouped into two main categories a) quantile mapping and b) regression-based methods [23]. Quantile mapping methods aim to adjust the individual quantiles of a modelled parameter to follow the statistical distribution of the observations. This is achieved by assuming that the target parameter follows a specific statistical distribution (e. g. gaussian). Instead of the parametric approach, the quantile-based correction can be implemented using the empirical quantiles or by applying kernel density functions. The adjustment of individual quantiles in a modelled parameter is translated to minimized deviations compared to observations, also resulting in reduced systematic and dispersion errors. Conversely, the mismatch between modelled and observed parameters can be adjusted by applying regression-based techniques. The target parameter (here, GHI) can be parameterized using supervised machine learning algorithms (MLAs) with auxiliary information (solar zenith angle, simple meteorological information, etc.) and/or modelled values as independent features. On the other side, process-oriented or conditional site adaptation approaches in terms of the state of the sky [20–24] or the aerosol types [30] were also found in the literature. A disaggregation of cloudy and clear instances before applying a site adaptation method could improve the overall performance and accuracy of the modelled irradiance [23]. In addition, individual adaptation of GHI under different aerosol types, e.g., under the dominance of coarse or mixed aerosols, potentially reduced the dispersion error of the adapted GHI [30]. In general, a site adaptation method may correct a target parameter adequately at a specific site, however, it may fail in other locations due to different atmospheric and climatic characteristics. The localization of the site adaptation process implies the difficulty of selecting a unique modelling approach to correct solar irradiances.

Post-processing of modelled irradiances deals with the direct comparison against high quality observations or the demonstration of dependencies with auxiliary information (i. e., modelled solar irradiances, solar zenith angle, clear-sky index, air temperature, time components, atmospheric parameters, etc.) using machine learning (ML) and Artificial Intelligence (AI) techniques. Although the application of ML and AI in solar energy studies has been continuously enhanced, only a few studies use such methods for site adaptation of solar irradiances [17,23,24,29,31]. Lorenz et al. [17] adjusted the biases between forecasts and observed irradiances by applying a fourth-order polynomial model with the clear sky index (ratio of GHI to that under clear sky conditions) and solar zenith angle as independent parameters. In a forward step, the addition of plausible independent parameters such as air temperature, surface pressure and relative humidity, as well as the replacement of the polynomial expression with a kernel conditional density estimator [31], lead to improved accuracy compared to the results of traditional Lorentz's MOS technique.

In order to overcome the non-universality of site adaptation methods, Fernandez-Peruchena et al. [24] proposed a regression scheme to correct GHI and DHI based on a best-subsets generalized linear model for the measured clearness index and the diffuse to global irradiance ratio using satellite-derived solar products. Regarding GHI, the modelling strategy was focused on drawing a multiple linear model using the modelled clearness index, modelled clear sky index, optical air mass, solar elevation angle and their interaction terms as independent predictors. The optimal predictors were determined by a stepwise approach by minimizing Akaike's criterion. The modelling procedure was repeated separately for clear and cloudy conditions, but the site adapted solar products were evaluated only at all sky conditions. The comparison against observed GHI independently of the sky's state showed low relative bias (<2%) and reduced dispersion errors when compared to those of the initially modelled GHI.

Instead of using a single model, Narvaez et al. [29] employed and compared different machine learning models for adapting GHI. More specifically, four ML models (multiple linear regression, neural networks, random forests, and AdaBoost) were developed and compared against the Quantile Mapping (QM) method; a state-of-the-art technique that principally adapts the compared statistical distributions. The ML algorithms estimated the observed GHI using the modelled solar irradiance (GHI, DHI and DNI), solar zenith angle and simple meteorological parameters (air temperature and wind speed) as auxiliary inputs. Random Forest was selected as the best ML model regarding the accuracy and overall performance with all ML models except AdaBoost to outperformed QM.

The site adaptation studies reported in the literature showed acceptable results. However, individual comparisons of the adapted irradiances for specific weather and atmospheric conditions, different seasons, etc. were completely missing even for research works reporting two-stage modelling approaches in terms of specific criteria (e. g. the presence or absence of clouds). The present study investigates the performance of GHI_{CAMS} in a south-eastern coastal Mediterranean location with complex terrain and evaluates various supervised machine learning algorithms (MLA) to adjust the all-sky (clear and cloudy) GHI_{CAMS}. The innovative points/extensions compared to the state-of-the art for site adaptation studies are summarized as follows:

- Hourly GHI_{CAMS} is verified against observed GHI under clear, cloudy, and all-sky conditions determined through a cloud-screening statistical approach, based on a newly proposed method to define the thresholds of clearness index for the classification of the sky state as cloudy, intermediate and clear. This approach is used to: a) examine the necessity of applying a site-adaptation scheme, b) assess the performance and accuracy of modelled irradiances.

- Various machine learning models, created both for clear and cloudy conditions, are used. Moreover, their performance to efficiently adapt GHI_{CAMS} is performed separately for clear and cloudy conditions, at different seasons, solar zenith angles and cloud fractions.

2. Data

2.1. Site description and ground-based measurements

Ground SSI observations come from the radiometric station located on the rooftop of the Laboratory of Atmospheric Physics, Patras, Greece (longitude: 21° 47′ 18.90″, latitude: 38° 17′ 28.91″, altitude: 44.5 m a.s.l.). Patras is a coastal city located in Southern Greece with mild winters and dry-hot summers; the climate type of Patras, according to the Köppen-Geiger climate classification system is Csa [32]. The city, except for its western front, which is coastal, is surrounded by mountains with peaks up to 2 km, leading to the considerably high variability of cloudiness. Regarding the atmospheric conditions, aerosols mainly originate from local emission sources such as traffic and biomass burning for domestic heating or agriculture activities, while the effect of transboundary air pollution from Central/East Europe and Sahara Desert is also distinguishable [33]. In general, the selected site is characterized by considerable spatiotemporal cloudiness and aerosol variability throughout the year.

One-minute averaged Global Horizontal Irradiance (GHI_{OBS}), Diffuse Horizontal Irradiance (DHI_{OBS}) and the corresponding standard deviations are measured by Kipp & Zonen CMP11 pyranometers from January 2014 to December 2020. The standard uncertainty is close to 1.9% when GHI reaches 800 W m^{-2} [34]. The manufacturer calibrated the instrument; periodic comparisons with a similar instrument show differences within the standard uncertainly. All good practices were followed for maintaining the instrument performance and the measurements quality control. This study considers only observations between sunrise and sunset. Hourly averages are calculated for clear sky detection and site-adaptation purposes. Hours with missing 1-min data (less than 90% of available data) have been discarded from the subsequent analysis.

2.2. CAMS-Rad Service

The Heliosat-4 method is the core module to produce GHI in CAMS-Rad Service, GHI_{CAMS} [35]. It estimates the all-sky GHI through the combination of Meteosat Second Generation (MSG) satellite images and simulations from a radiative transfer model at fast rates. Heliosat-4 consists of a) a clear-sky model, the McClear model, which calculates solar irradiances at clear sky conditions using atmospheric inputs from CAMS reanalysis project [36,37] and b) the McCloud model, which calculates solar irradiances at cloudy atmospheres using the concept of cloud modification factor [38]. The cloud modification factor is a function of cloud attenuation and ground reflection. GHI is derived through multiplying the cloud modification factor with the clear-sky GHI output of McClear. The cloud properties are estimated from MSG satellite images by applying the APOLLO method [35]. The SG2 algorithm of Blanc and Wald [39], provides the sun-position parameters required in solar irradiance simulations.

The CAMS-Rad dataset ranges from 2004-02-01 until two days before the date the query is performed, with various temporal resolutions extending from 1 min to 1 month. The standard database request includes modelled solar irradiances (GHI, DHI, and DNI) at clear skies and all-sky conditions. Apart from the standard solar products, the 'detailed info' expert output mode of CAMS-Rad includes all those atmospheric inputs required for the calculation of the SSI. This detailed description is available only at a 1-min temporal resolution. It includes atmospheric (aerosol optical thickness at 550 nm, total columnar water vapor, ozone) and cloud-related parameters (total cloud fraction, cloud optical thickness, cloud type) [40]. The 'detailed info product' also contains raw and bias-corrected solar irradiances. For the standard products, the CAMS-Rad service evaluates a bias-correction scheme, and the corrected irradiance is delivered to the end-users. Thus, any post-processing regarding local adaptation is not possible in typical requests.

CAMS-Rad data were downloaded using the 'detail info' expert mode from the SoDa Web service from January 2014 to December 2020 to match the time period of measured solar irradiances at the ground station. Since the paper's main objective is the site-adaptation of GHI_{CAMS} at all-sky conditions, raw solar irradiances are used instead of bias-corrected solar products. CAMS-Rad data are upscaled to 1-h resolution and temporally cropped to range from sunrise to sunset.

3. Methodology

3.1. Clear sky detection

The disaggregation of a target dataset into subsets of similar characteristics (e.g., in terms of the sky conditions) and the forthcoming site-adaptation of each subset separately may result in error improvements and better performance of the adapted GHI [22–24]. Cloud information (available in CAMS-Rad at 1-min temporal resolution, Section 2.2), such as cloud fraction and cloud optical thickness, could function as clear sky detection (CSD) tool. However, CSD based on ground-based solar irradiance is preferable due to induced errors in satellite-derived CSD. For example, the Heliosat-4 method, under certain circumstances, suffers from 'false alarm' cases, which erroneously characterize clear sky cases as cloudy [35]. Furthermore, the extraction of cloud information at specific points (coordinates) from satellite images includes considerable uncertainty. Apart from spatial downscaling, the temporal disaggregation to 1-min temporal resolution or the temporal upscaling to hourly or higher temporal scales introduces additional uncertainty. For example, 1-min CAMS-Rad data are retrieved through temporal interpolation of the clearness index between consecutive 15-min intervals (the temporal resolution of MSG satellite images) ignoring the high fluctuating character of cloud cover. In general, CSD applications are evaluated using 1-min GHI observations (GHI_{OBS}). At lower temporal resolutions, CSDs may be designed from scratch or modified appropriately [22,41]. Despite the different algorithmic basis and temporal scales, most CSD methods compare GHI_{OBS} against clear-sky GHI or use the (modified-) clearness index, considering several criteria and thresholds to detect clouds [42].

In this study, a modified version of the CSD methodology, proposed by Ref. [43], discriminates the sky state. Ineichen et al. [43] used the modified clearness index, k_t' [44], to separate the sky type into three main categories, a) cloudy sky ($k_t' \leq 0.3$), b) intermediate sky ($0.3 < k_t' \leq 0.65$), and c) clear sky ($k_t' > 0.65$). The modified clearness index k_t' is defined as,

$$k_t' = k_t / [1.031\exp(-1.4 + 9.4/M) + 0.1] \quad (1)$$

with k_t is the global clearness index, defined as the ratio of GHI to the solar irradiance at the top-of-the atmosphere on a horizontal plane (E_0), $k_t = GHI/E_0$ (2), and M is the optical air mass [45]. However, the thresholds for the three distinct zones were selected arbitrarily and could differ for ground measurements with lower

temporal resolution. Following the three-state sky's division of [43], the specific thresholds are adjusted by applying a Hidden Markov Model (HMM). Considering that a mixture of sky types represents k_t', the HMM aims to identify regimes in k_t' time series, each represented by a specific probability distribution (here the gaussian distribution).

Fig. 1a illustrates an idealized example of the HMM result. The intersection points between the consecutive gaussian distributions could diagnose the three k_t' regimes. The first threshold (Th_1) separates the cloudy from intermediate cloudy conditions while Th_2 separates the intermediate from clear conditions. Fig. 1b shows the frequency histogram of the observed k_t' ($k_t'_{,OBS}$). When the HMM model is applying to $k_t'_{,OBS}$, the threshold Th_2 equals 0.73, i.e., indicating clear sky conditions when the clearness index exceeds 0.73 (shaded blue area in Fig. 1b).

A classification into two (clear and cloudy) rather than three states may also be representative. Using a two-state HMM, the threshold value to classify the sky state to clear or cloudy estimated equal to 0.67. When comparing the two-state against the three-state HMM model using the Bayesian Information Criterion (BIC), the three-state HMM model shows a lower BIC. So, the threshold of 0.73 is finally used to classify the sky conditions. The HMM simulations are performed using the depmixS4 R package [46]. It is worth mentioning that other CSD methodologies could be applied, possibly leading to different results. However, is beyond the scope of this paper to describe the most appropriate CSD method.

3.2. Site adaptation framework

The concept of site adaptation aims to reduce the bias between a modelled and observed parameter. Eq. (3) decomposes the site adapted GHI (GHI_{AD}) as,

$$GHI_{AD} = GHI_{OBS} \Rightarrow GHI_{AD} = GHI_{CAMS} + \Delta GHI \quad (3)$$

where the subscripts OBS, CAMS, and AD correspond to the observed, initially modelled and site adapted GHI. The term ΔGHI ($= GHI_{OBS} - GHI_{CAMS}$) is the bias between the CAMS and the observed solar irradiances. The problem is transformed into a post-processing process focusing on ΔGHI and it can be treated by applying MOS techniques.

Machine learning algorithms (MLAs) may extract possible non-linear patterns and interactions between ΔGHI and independent features. The flowchart of Fig. 2 provides an overview of the methodology to site-adapt the GHI_{CAMS}. GHI_{CAMS} is also used as input along with atmospheric parameters and solar zenith angle to estimate ΔGHI using various MLAs. The incorporation of physics-based information (i.e., GHI_{CAMS}) is the main difference against the traditional residual models. The modelled biases are finally added to the GHI_{CAMS} to obtain the site adapted GHI (GHI_{AD}).

The proposed methodology is applied separately at clear and cloudy conditions with $\Delta GHI = f(GHI_{CAMS}, SZA, AOD_{550}, WV, TCO_3)$ and $\Delta GHI = f(GHI_{CAMS}, SZA, AOD_{550}, WV, TCO_3, COT, CF)$.The independent features are GHI_{CAMS} (raw GHI from CAMS-Rad service),

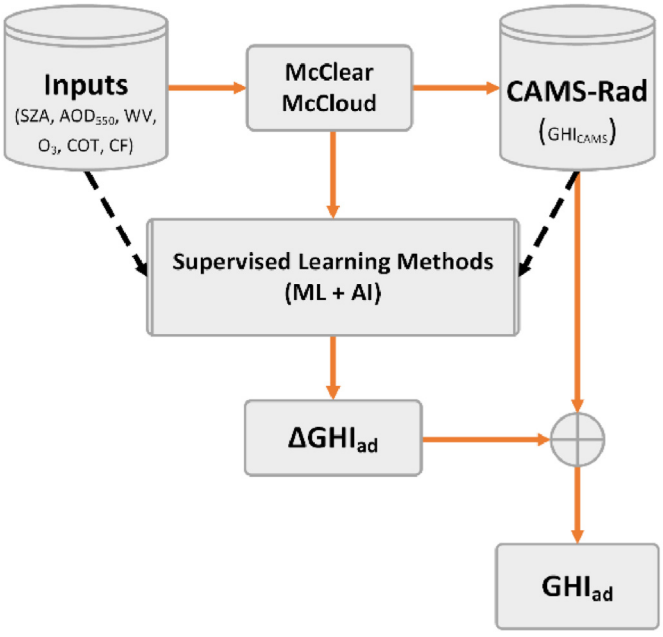

Fig. 2. Flow chart of the site adaptation methodology. ML and AI techniques predict ΔGHI in terms of several atmospheric parameters, solar zenith angle and GHI_{CAMS}.

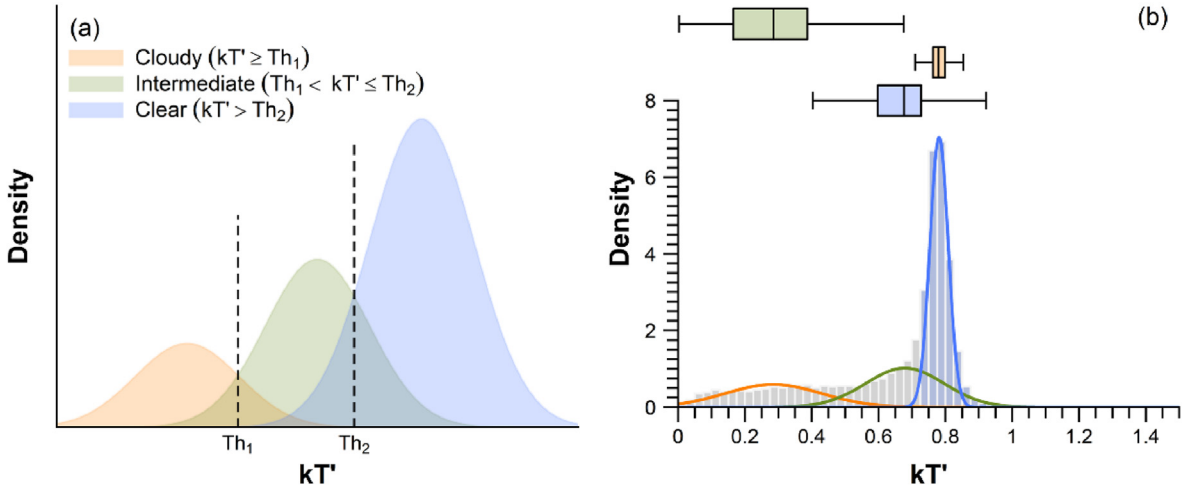

Fig. 1. a) Idealized example of mixture with 3 Gaussian distributions, each one representing different sky conditions. The vertical dashed lines in thresholds 1 (Th_1) and 2 (Th_2) separate the total gaussian mixture area in terms of the sky's state. b) Frequency histogram of the observed ($k_t'_{,OBS}$) The colored lines represent cloudy (orange), intermediate (green), and clear (blue) sky conditions. The Gaussian curves are drawn using a 3-state Hidden Markov Model (HMM). The representative ranges for each sky condition are retrieved through the intersection of consecutive gaussian lines. The shaded blue area indicates the presence of clear skies ($k_t'_{,OBS} > 0.73$). The boxplots summarize the range of $k_t'_{,OBS}$ for the three sky states (outliers are omitted).

SZA (solar zenith angle), AOD_{550} (aerosol optical depth at 550 nm), WV (total column water vapor), TCO_3 (total column ozone), COT (Cloud Optical Thickness), CF (cloud fraction) and $f(\cdot)$ is the selected model. Any physical inconsistencies produced by the site adaptation procedure including a) $GHI_{AD} < 0$ and b) GHI_{AD} exceeding a maximum observation limit ($a_{GHI} = 1.05 \times \max(GHI)$, [24]), are padded with GHI_{CAMS}.

Considering that GHI_{CAMS} is modelled through McClear and McCloud modules in CAMS-Rad, all independent parameters are incorporated into the residual modelling approach to capture any remnant contributions not adequately described by CAMS-Rad. The input parameters are not selected arbitrarily because they represent different physical effects contributing to GHI attenuation. Therefore, feature selection techniques for extracting the most prominent contributors are not implemented at this stage.

Supervised learning methods with various prediction mechanisms (linear, tree-based, distance-based and kernel-based) are implemented for site adapting GHI_{CAMS}, namely.

1. Fully Connected Neural Networks: Neural Networks (NN) [47] attempt to extract relationships between a target parameter and independent features using a series of layers which are consisted of a series of units (neurons). Therefore, NN is formed by three parts; the first includes the input features, the second is the intermediate layer which controls the network's complexity via the number of the hidden layers and units (hidden nodes per layer), and the third layer is the output. Here, all hidden layers are activated through the Rectified Linear Unit (ReLU) function, also including a dropout fraction on its weights to prevent overfitting. The output layer consists of a single linearly activated layer [48]. The AdaDelta (Adam) optimizer performs the stochastic gradient descent on the neural network. The NN architecture depends on various parameters, termed hyperparameters in ML and AI jargon. The internal parameters controlling the NN structure and complexity are estimated by minimizing a regularized empirical loss function. Conversely to the standard NNs, the regularized objective here combines a regression loss function; the mean of squared residuals (MSE), with a structural loss function described by multiplying the L2-norm of the NN weights with a regularization parameter (ℓ_2). The regularized term penalizes the hidden layer's complexity and mitigates overfitting. For optimizing the training procedure, the input parameters are standardized to have zero mean and unit variance before running NNs.
2. Extreme Gradient Boosting machines (XGBoost): XGBoost is a scalable machine learning algorithm that builds an extensive collection of decision trees and ensembles them to make predictions [49]. Each regression tree is created by the entire training set or randomly selected sample portions of the training set using split-based rules of the predictor space. In XGBoost, a tree is grown by splitting into branches following the approximate greedy algorithm. A series of internal parameters estimated by minimizing a regularized loss function control the tree's structure and complexity. In contrast to the unextreme gradient boosting methods, the regularized objective in XGBoost combines a regression loss function such as the sum of squared residuals and a regularization term that penalizes the tree complexity. A shrinkage parameter that controls the effect of each tree in the final prediction process and a feature subsampling process used for training each tree and its levels are also included in XGBoost to prevent overfitting.
3. Random Forests (RF): Random forests is a tree-based method for classification and regression problems [50]. RF method builds a multitude of decorrelated weak learners (decision trees) and then averages them to make the final predictions. The forest of regression trees produces an ensemble of predictions, and the final output is determined through averaging over the ensemble values. RF is based on 'bagging' or 'bootstrap aggregation' to reduce variance and minimize overfitting. Each tree is trained simultaneously by bootstrap samples, including a random set of training values and features at each tree node to ensure low correlation among decision trees. In this case, random forests select a subset of features compared to decision trees that consider all possible feature splits. The complexity and structure of RF are controlled by three main hyperparameters, namely the number of trees, the minimum nodes in each tree and the number of features used for tree splits.
4. Elastic Net regression (GLMNET): The elastic net regression (GLMNET) has similar parametrization to ordinary least squares, but it is equipped with two penalized terms (lasso and ridge regularizers) in the loss function [51,52]. The lasso penalty (ℓ_1) multiplies the L_1-norm of regression coefficients in the loss function and makes a kind of feature selection. On the other side, the ridge penalty (ℓ_2) shrinks the regression coefficients. It multiplies the L_2-norm of regression coefficients in the loss function, and its main objective is to mitigate overfitting. In contrast to the lasso term, the ridge term does not act as a variable selection function.
5. Multivariate Adaptive Regression Splines (MARS): MARS is a non-parametric statistical technique for regression aiming to capture nonlinearities between the dependent and independent variables using piecewise linear relationships [53]. The nonlinearities are incorporated in MARS by assessing knots similar to step functions, also termed linear splines (basis functions). In this way, the target variable in MARS is not directly modelled in terms of the feature space, but it is estimated through an additive form of the individual basis functions. All possible knots are determined across the feature space. The optimal number of terms included in the final model and the position of knots are extracted by minimizing the Generalized Cross Validation (GCV), while the regression coefficients are obtained by the minimization of the residual sum squares like in standard linear regression [52].
6. Support Vector Regression (SVR): Support Vector Machines perform both classification and regression problems. The main objective of this supervised learning method is to extract a hyperplane in the kernel-induced feature space with good generalization performance [52,54]. The non-linear patterns are captured in SVR by applying kernels (here the radial basis function is used) in feature space. The ε-insensitive loss regression is commonly used in SVR problems where the support vectors falling within the margin defined by the ε parameter do not contribute to the minimization of the ε-SVR loss function [54].

Several hyperparameters describe the complexity and the structure of the pre-described supervised learning techniques. Before the application of such methods, hyperparameter tuning is recommended to increase the model's performance [55]. Table A1 describes the tuning hyperparameter space for the statistical techniques used in this study. Generating predictions with the entire cartesian hyperparameter space is computationally expensive and time-consuming. An alternate approach consists of applying a randomized grid search procedure for the extraction of a model converging to "optimal" ones. The tuning scheme for the ML algorithms (XGBoost, RF, GLMNET, MARS and SVR) includes a randomized grid search procedure equipped with an N-fold cross-validation scheme (CV). The randomized searching algorithm runs 100 times with 5-fold CV and RMSE as the fitness function. The total number of simulations is considered capable of estimating the

final hyperparameter set. On the other side, the randomized searching algorithm for neural networks runs 100 times with 5 trials per run. Each NN is trained with 1000 epochs, a batch size of 64 and the mean squared error (MSE) as the fitness function. The tuning procedure for the statistical models is evaluated using keras [56], kerastuner [57], caret [58] and mlr [59] R packages.

The site adaptation scheme of Fig. 2 could be evaluated in every place around the globe with available radiometric measurements, even for short periods. The selected input variables are available from every reanalysis and satellite product, so, they could be easily treated as input parameters. This makes the ML process as simple as possible. However, model tuning is necessary to retrieve the best configurations that optimally fit the site's specific atmospheric and climatic characteristics.

3.3. Statistical evaluation

Numerous statistical indicators exist in the literature for validating a modelled product [60]. provided an extensive review of the most used statistical indices in solar-related applications. Here, the GHI_{CAMS} and GHI_{AD} are statistically verified against GHI_{OBS} with the mean bias error (MBE), the root mean square error (RMSE) and the correlation coefficient (R). Briefly, the MBE and RMSE are dispersion indicators with optimal values equal to zero. The normalized errors (nMBE and nRMSE) are also calculated using the average GHI_{OBS} as reference. The correlation coefficient, R, is a performance metric and measures the linear association between two variables with an optimal absolute value equal to unity. The RMSE can serve as a distribution-scale metric for examining the distribution similitude of the observed and modelled distributions if modified appropriately, e.g., by using as inputs the cumulative distribution functions (CDF),

$$RMSE_{CDF} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(q_{CAMS,i} - q_{OBS,i}\right)^2} \quad (4)$$

with N the number of discrete levels selected for drawing the CDF and q the individual quantiles.

4. Results and discussion

4.1. GHI_{CAMS} against observations in Patras, Greece

This section includes the validation of GHI_{CAMS} against observations from January 2014 to December 2020. First, the sky conditions are classified as to clear or cloudy by the CSD approach (Section 3.1) considering only the measurements with SZA <80° to avoid low-sun and shading issues. The modified global clearness index, $k_t'_{,OBS}$, calculated through GHI_{OBS}, is a proxy for detecting clouds, with clear skies for $k_t'_{,OBS}$ exceeding 0.73. Fig. 3a summarizes the CSD results on a yearly basis. In total, 56.3% of the hourly data (14536 out of 25851) are classified as clear. The annual averages of clear and cloudy hours were 2077 and 1611 respectively. Each year, the number of clear hours is higher than the cloudy ones.

The frequency histograms for the modified global clearness index, k_t', using GHI_{OBS} and GHI_{CAMS}, show unimodal, significantly peaked, and skewed statistical distributions towards higher k_t' (Fig. 3b). Higher frequencies of $k_t'_{,CAMS}$ are revealed, indicating erroneously assigned clear cases under the presence of clouds. CSD outcomes in terms of $k_t'_{,OBS}$ and $k_t'_{,CAMS}$ are further examined using the confusion matrix (subplot in Fig. 3b). The diagonal values represent the number of instances classified correctly as clear (True-Positive, TP) and cloudy (True-Negative, TN). The off-diagonal elements indicate the misclassified cases. More specifically, 1156 False-Negative (FN) and 1936 False-Positive (FP) cases exist, implying that CSD using the GHI_{CAMS} cannot completely reproduce the actual classes. According to the confusion matrix results, the accuracy metric and the F_1-score are equal to 0.88 and 0.90, respectively.

Fig. 4a–crepresent the cross-relationships between GHI_{CAMS} and GHI_{OBS} in Patras, Greece, for the entire measurement period. The CAMS-Rad exhibits high performance at all-sky conditions (Fig. 4a), with a correlation coefficient R = 0.97, indicating the strictly linear relationship between the modelled and measured solar irradiances. CAMS-Rad overestimates the observations with MBE = 22.8 W m^{-2} and a slope for the best fit line of 0.97, while the data points are significantly dispersed along the bisector (dashed line) showing a highly variable pattern (RMSE = 74 W m^{-2}). The systematic (nMBE) and dispersion (nRMSE) errors are significant, accounting for 4.9% and 15.7% of the average GHI_{OBS} (470 W m^{-2}) respectively.

Splitting the entire dataset into clear and cloudy cases enables the precise examination of error propagation concerning the sky state. Clouds are mainly responsible for the error magnitude since the CAMS-Rad performs efficiently at cloudless conditions. The dispersion error accounts for 38.8% of the mean GHI_{OBS} (274.7 W m^{-2}) with an RMSE = 106.5 W m^{-2} and CAMS-Rad significantly overestimates the observations (MBE = 48.2 W m^{-2}, nMBE = 17.6%). A vast number of points (Fig. 4b) exist above the identical line, indicating the CAMS-Rad inability to reproduce GHI_{OBS} under the presence of clouds. Under clear skies (Fig. 4c), the accuracy and performance for GHI_{CAMS} are substantially improved. The systematic and dispersion errors are remarkably lower with nMBE = 0.5% and nRMSE = 4.9% and an equal to unity linear relationship.

The difference between GHI_{OBS} and GHI_{CAMS} statistical distributions is also significant. According to Eq. (4), the CDF-based RMSE ($RMSE_{CDF}$) compares the statistical distribution of two variables by measuring the dispersion between the CDFs. $RMSE_{CDF}$ approaches 55.2 W m^{-2} and 23.8 W m^{-2} for cloudy and all-sky conditions respectively. Under clear skies, GHI_{OBS} and GHI_{CAMS} (Fig. 4d) are distributed similarly with $RMSE_{CDF}$ = 5.5 W m^{-2}.

The discrepancies between cloud-free GHI_{CAMS} and GHI_{OBS} are due to the underestimation of the cloud-free conditions from the CAMS-Rad service. The frequency histograms of Fig. 3b and the confusion matrix's results address this fact. Independently of the magnitude of the k_t', the ideal situation for a clear-sky dataset corresponds to an equal number of modelled and observed k_t', both exceeding the clear sky threshold of 0.73 (True Positive cases – TP). The existence of 1156 False-Negative cases (FN) with $k_t'_{,CAMS} \leq 0.73$ and $k_t'_{,CAMS} > 0.73$, determines the underestimation of the observed cloud-free cases from CAMS-Rad. The GHI underestimation at 'real' clear skies is further addressed by comparing GHI_{obs} and GHI_{CAMS} for for i) $k_t'_{,CAMS} > 0.73$ (TP), and ii) $k_t'_{,CAMS} \leq 0.73$ (FN), respectively (Fig. 5a). The results of Fig. 5a verify that CAMS-Rad calculates GHI assuming the presence of clouds even if the CSD using $k_t'_{,OBS}$ shows the dominance of clear skies. In the case of $k_t'_{,CAMS} \leq 0.73$, the systematic bias reaches −53.9 W m^{-2} (nMBE = −12.1%), and the dispersion error is 83.5 W m^{-2} (nRMSE = 18.7%), significantly higher in absolute values than those calculated for $k_t'_{,CAMS} > 0.73$ (MBE = 8.1 W m^{-2}, nMBE = 1.3%, RMSE = 19.9 W m^{-2}, nRMSE = 3.1%). The slope of the best-fit line drops from 0.98 to 0.91, with all GHI pairs located beneath the 1:1 line. To support further the assumption that the erroneously assigned cloud presence is reflected in GHI_{CAMS} discrepancies, the clear-sky GHI_{CAMS} ($GHICAMS_{,cs}$), provided both in standard and 'detailed info' requests, is compared against observations for TP and FN cases described previously (Fig. 5b). According to Fig. 5b, the clusters of points for both $k_t'_{CAMS}$ cases look similar, indicating that $GHI_{CAMS,cs}$ is closer to observations, and in this case, individual

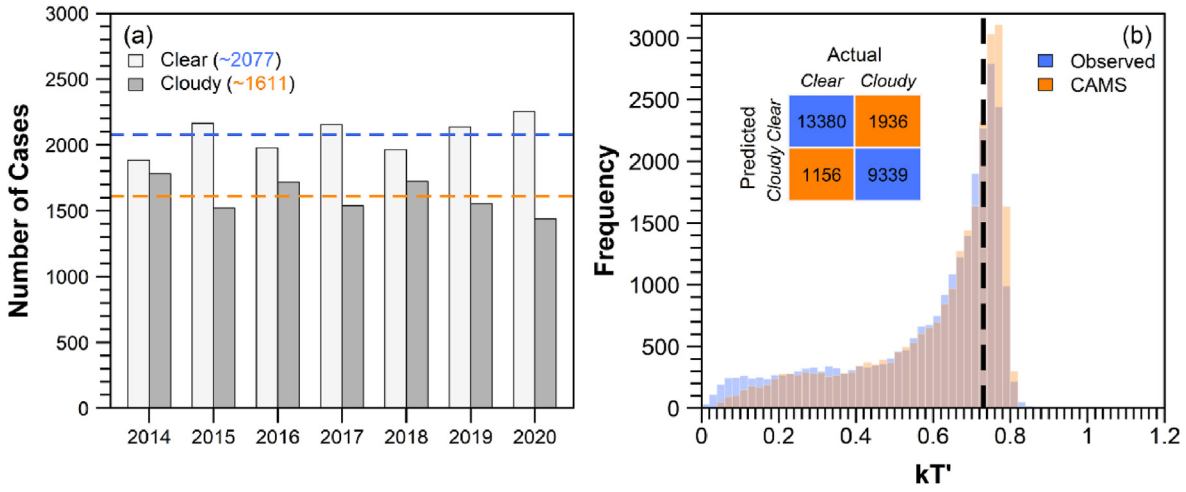

Fig. 3. a) Annual distribution of clear and cloudy hours from January 2014 to December 2020. The dashed horizontal lines (blue for clear and orange for cloudy cases) correspond to average values (Clear: 2080 and Cloudy = 1614) over the whole period. b) Frequency histograms of the modified global clearness index k_t' for GHI_{OBS} (blue) and GHI_{CAMS} (orange). The vertical dashed line for $k_t' = 0.73$ separates clear from cloudy conditions. The subplot corresponds to the confusion matrix of the CSD approach using the $k_t'_{,OBS}$ as reference. CSD is evaluated using $k_{t'OBS}$ for the actual-reference classes and k_t' from GHI_{CAMS} ($k_t'_{,CAMS}$) for the predicted classes.

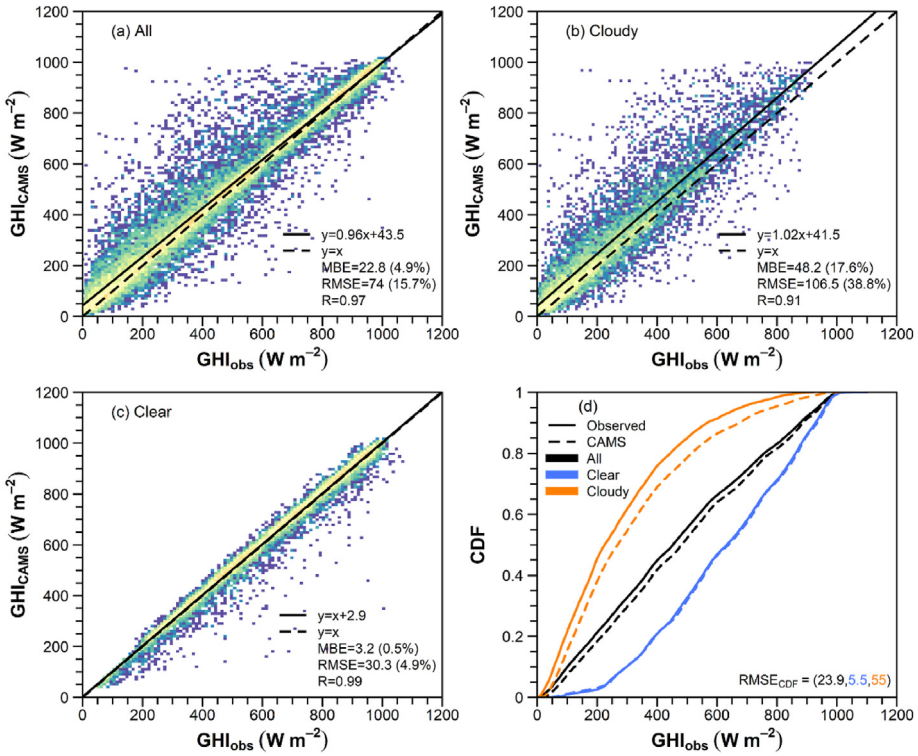

Fig. 4. Scatter density plots between GHIOBS and GHICAMS, Greece, from January 2014 to December 2020 at a) all-sky, b) cloudy (kt',OBS ≤0.73), and c) clear (kt',OBS >0.73) conditions. Warm colours indicate high density. The dashed and solid lines correspond to the 1:1 and best linear fit. The percentage errors in the parenthesis describe the normalized MBE and RMSE (nMBE and nRMSE) with the averaged observed solar irradiance. d) Cumulative Distribution Functions (CDF) for GHIOBS (solid lines) and GHICAMS (dashed lines) at all-sky, cloudy, and clear conditions. The RMSECDF describes the distribution similitude via examining the squared differences of the individual distribution quantiles. The MBE, RMSE, and RMSECDF are express W m−2.

comparisons regarding k_t' regimes are unnecessary. The MBE and RMSE explain only a small portion of the averaged GHI_{OBS} with MBE = 7.5 W m^{-2}, nMBE = 1.2%, RMSE = 22.4 W m^{-2}, and nRMSE = 3.6%.

4.2. Site adaptation of GHI_{CAMS}

4.2.1. Data splitting

A common practice in statistical modelling and intercomparison studies includes splitting the available dataset into reference

(training) and target (testing) subsets. The selection of a reasonable period as a reference for site-adaptation purposes depends clearly on the abundance and quality of ground-based measurements [23]. examined the minimum data requirements for performing site adaptation using ground-based datasets with various temporal intervals ranging from 3 months to 2 years. The sensitivity analysis showed that the error metrics (MBE and RMSE) tend to a minimum value if a complete year of observations is used at least. For shorter periods, the errors increase substantially.

In this study, the last two available years (2019 and 2020) are used for training and the entire period (2014–2020) for site adaptation purposes. In addition, two sensitivity analyses were performed to confirm that every arbitrarily chosen 2-year data period for training leads to similar results. The first uses all possible combinations of a 2-calendar year period between 2014 and 2020 as reference. The second takes as reference a random sample of about 7400 values that correspond to a 2-year period. The random sampling process is repeated 1000 times to obtain relatively robust results. Fig. A1 and A2 display the results for both sensitivity analyses. The bars' height and the vertical segment correspond to each statistical metric's median and interquartile range (IQR = $Q_{75\%}$–$Q_{25\%}$). Small IQR values indicate closeness to the central tendency (in this case, the median), leptokurtic distributions of the statistical indices, and stability in the sensitivity analysis. Comparing the two sensitivity analyses, the medians of the statistical metrics are almost similar, with minimal IQR deviations. Low IQR values in the second sensitivity procedure are due to the number of iterations. Also, the medians are close to the statistical indices obtained for the entire data period (Fig. 4). The above discussion confirms the 'unbiased' selection of 2019–2020 as a reference/training period.

According to Section 4.1, the training and testing datasets contain several erroneously classified clear or cloudy instances compared to the 'real' sky conditions. Since the site-adaptation approach does not include any parametrization for handling misclassified cases, all data are used for the model's design process. At this stage, it is necessary to mention that site adaptation aims to correct the systematic and dispersion errors and not generate irradiances capable of performing CSD. For CSD, highly accurate ground observations or, in the ideal case, all-sky images are necessary for precisely discriminating the sky's state.

4.2.2. Statistical assessment of site adapted GHI

This section describes the site adaptation results of MLAs and further comparisons against state-of-the-art methods. Supervised learning techniques attempt to predict the ΔGHI using several exogenous variables, including GHI_{CAMS} as input. The 'optimal' hyperparameter configurations for the statistical methods at clear and cloudy conditions are retrieved through the randomized-grid search method (Section 3.2). The results of the tuning approach are represented in Table A2.

Table 1 shows the statistical indicators of dispersion (MBE, RMSE, and the relative forms), the overall performance (R), and statistical similitude ($RMSE_{CDF}$) of GHI_{CAMS} and GHI_{AD} against GHI_{OBS} under all-sky, clear, and cloudy conditions as classified by $k_{t,OBS}$'. Two state-of-the-art site adaptation techniques (LIN and EQM) are also considered for comparison purposes, apart from the MLAs. Briefly, Linear regression bias removal (LIN) draws a linear relationship between GHI_{CAMS} and GHI_{OBS} and then projects GHI_{CAMS} across the identical line (1:1). Thus, the systematic bias is eliminated, also reducing the dispersion error. On the other side, the Empirical Quantile Mapping (EQM) method adjusts the distribution of the GHI_{CAMS} to follow that of observations acting non-parametrically in terms of the Empirical Cumulative Distribution Function (ECDF). This ECDF matching improves the systematic and dispersion errors of GHI_{AD} by correcting the individual quantiles of the modelled statistical distribution [19,23]. represented the LIN and EQM approaches, highlighting their potential to correct satellite-derived and reanalysis solar irradiances.

According to Table 1, the statistical metrics for all site adaptation models are better than those for GHI_{CAMS}, with significant improvements under cloudy and all-sky conditions. The sky state is discerned by using $k_{t,OBS}'$ and not $k_{t,CAMS}'$. In this case, site adaptation models are evaluated under the 'real' sky conditions. The all-sky systematic errors (MBE and nMBE) are lower than 8 W m^{-2}, except of SVR, with the lowest MBE for MARS and RF models (5.9 W m^{-2}) and reductions exceeding 50% compared to the raw (CAMS) case (22.8 W m^{-2}). The normalized systematic errors are lower than 2%, in agreement with other site-adaptation studies reported in the literature. All-sky errors reflect the improvement detected in cloudy cases. The tree-based MLAs (RF and XGBoost) report the lower systematic errors (<26 W m^{-2}) compared to the other MLAs and the state-of-the-art methods with an approximate doubling value for the raw case (48.2 W m^{-2}). MLAs outperform

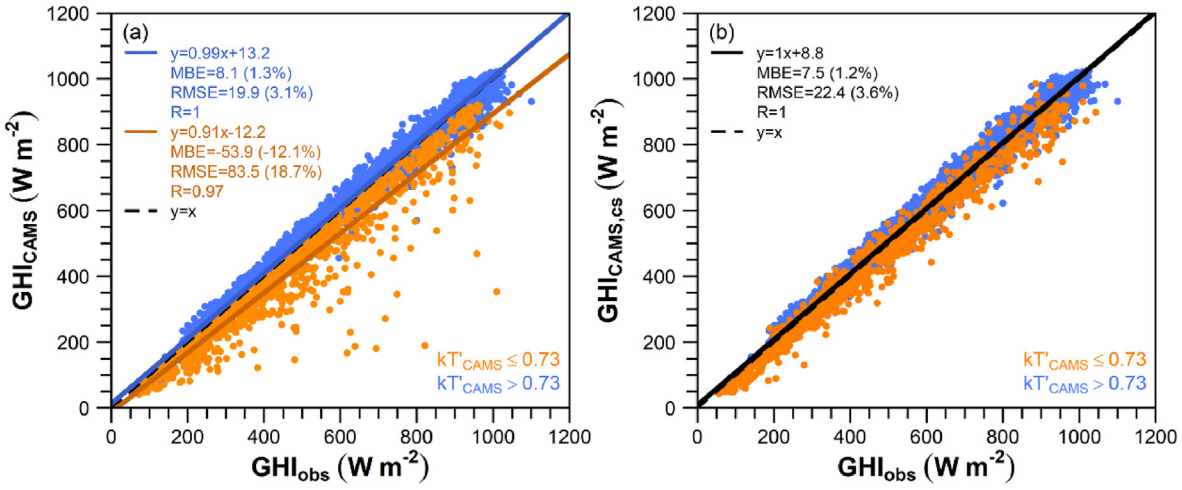

Fig. 5. Scatter plots between GHI_{OBS} and a) GHI_{CAMS}, and b) clear-sky GH_{CAMS} ($GHI_{CAMS,cs}$) from January 2014 to December 2020 at clear sky conditions ($k_{t,OBS}' > 0.73$). The blue and orange dots correspond to GHI pairs with $k_{t,CAMS}' \leq 0.73$ and $k_{t,CAMS}' > 0.73$, respectively. The percentage errors in the parenthesis describe the normalized MBE and RMSE. The MBE and RMSE of the GHI are expressed in W m^{-2}.

Table 1
Goodness-of-fit (GOF) statistics for GHI_{CAMS} and GHI_{AD}. MBE, RMSE, and $RMSE_{CDF}$ are expressed in W m^{-2}; nMBE, nRMSE in %.

	MBE	nMBE	RMSE	nRMSE	R	$RMSE_{CDF}$
All						
CAMS[a]	22.8	4.9	74	15.7	0.98	23.8
NN	7.9	1.7	62.8	13.4	0.98	9.6
XGBoost	7.2	1.5	57.1	12.1	0.98	8.8
RF	5.9	1.3	56	11.9	0.98	8
GLMNET	7.2	1.5	65.7	14	0.97	11.5
SVR	9.7	2.1	64.9	13.8	0.97	10.6
MARS	5.9	1.3	64.4	13.7	0.97	8.2
EQM	6.4	1.4	71.7	15.3	0.98	7.6
LIN	7.4	1.6	75.5	16.1	0.98	13.1
Clear						
CAMS	3.2	0.5	30.3	4.9	0.99	5.50
NN	−8.5	−1.4	30.2	4.9	0.99	9.4
XGBoost	−7.2	−1.2	29.2	4.7	0.99	7.8
RF	−8.1	−1.3	29.8	4.8	0.99	8.7
GLMNET	−10.9	−1.8	35.6	5.7	0.99	11.4
SVR	−5.5	−0.9	30.7	4.9	0.99	6.6
MARS	−10	−1.6	36.2	5.8	0.99	11
EQM	−12.9	−2.1	32.9	5.3	0.99	15
LIN	−12	−1.9	32.1	5.2	0.99	16.8
Cloudy						
CAMS	48.2	17.6	106.5	38.8	0.95	55
NN	29	10.5	88.6	32.3	0.96	33.7
XGBoost	25.8	9.4	79.7	29	0.93	29.1
RF	24	8.7	77.7	28.3	0.94	26.4
GLMNET	30.4	11.1	90.9	33.1	0.91	32.7
SVR	29.2	10.6	91.7	33.4	0.92	34.7
MARS	26.4	9.6	88.3	32.2	0.92	29.4
EQM	31.3	11.4	101.9	37.1	0.95	43.7
LIN	32.4	11.8	108.2	39.4	0.95	53.3

[a] CAMS: uncorrected GHI_{CAMS}.

EQM and LIN regarding the systematic bias with MBE differences exceeding 3.4 W m^{-2}. The dispersion error shows similar patterns as the systematic bias. RMSE ranges for MLAs are 56 W m^{-2} (RF) – 64.9 W m^{-2} (SVR), 77.7 W m^{-2} (RF) – 91.7 W m^{-2} (SVR) in all-sky and cloudy conditions, respectively. MLAs efficiently reduce RMSE. The relevant RMSE reductions exceed 8.3 W m^{-2} (GLMNET) and 14.8 W m^{-2} (SVR) compared to the uncorrected case (all: 74 W m^{-2} and cloudy: 106.5 W m^{-2}). Among regression methods, NN is the best model outperforming MARS and GLMNET, indicating the better reproduction of possible nonlinearities among the observed GHI and the auxiliary predictors. As denoted in Ref. [23], it is hard to obtain distinguishable reductions in the error metrics in cases where the quality of modelled data is already high. This fact is observed in clear sky conditions where CAMS performs efficiently. According to Table 1, slight RMSE reductions are calculated for XGBoost and RF. NN provides comparable RMSE to the uncorrected GHI, while GLMNET and MARS cannot reduce the dispersion bias giving high RMSEs (Table 1).

The overall performance assessed through the correlation coefficient (R) does not change significantly for the raw and the site adapted irradiances, remaining at high levels (all: $R > 0.98$, cloudy: $R > 0.95$, and clear: $R \approx 1$). Another critical point is to explore whether site adaptation improves the distribution similitude between the adapted and the observed solar irradiances. All models except LIN provide $RMSE_{CDF}$ lower than 10 W m^{-2} at all-sky conditions. A significant result for the distribution comparisons is that MLAs outperform EQM at clear and cloudy conditions, even if the mathematical 'kernel' of EQM is distribution matching.

The site adapted GHIs are also evaluated on a seasonal basis. Figs. 6–8 represent the systematic and dispersion errors at all-sky, clear and cloudy conditions for the uncorrected and the ML-adapted GHI. It is clearly shown that MLAs perform efficiently at all seasons in all-sky and cloudy cases, with higher MBE and RMSE in springtime. The high variability of cloudiness explains this behavior in spring. Especially under the dominance of broken clouds, the satellite suffers from reproducing accurately the cloud fraction and the cloud optical depth giving erroneous GHI predictions. Such cloud cases are favorable in spring over the selected location. Site adaptation methods equipped with cloud-related exploratory variables try to reduce the error patterns to a certain degree. In contrast, the simple methods of EQM and LIN show high systematic and dispersion errors.

MARS records the lowest MBE in spring under all-sky conditions (6.49 W m^{-2}). However, looking closer at the MBE under clear and cloudy skies, the relevant values are −19 W m^{-2} and 31 W m^{-2}, indicating that strong underestimation for clear skies (Fig. 7a) is balanced by strong overestimation of GHI during the presence of clouds (Fig. 8b), finally giving the low MBE. Thus, it is better to examine MBE separately for clear and cloudy conditions. In spring, the minimum underestimation under clear skies is calculated with SVR (−8.9 W m^{-2}), while the uncorrected GHI_{CAMS} is close to observations with MBE = −1.92 W m^{-2}. Additionally, RF gives the lowest systematic error for cloudy conditions (28.6 W m^{-2}). Generally, under clear skies for all seasons, site adaptation models cannot beat the uncorrected GHI_{CAMS}. When clouds are present, RF is the optimal model, producing the lowest MBEs (Fig. 8a). The seasonal analysis of the systematic bias implies that MLAs with tree-based mechanisms give the best MBEs.

Regarding the dispersion error, MLAs show similar patterns under clear skies with lower RMSE values in autumn and summer. Most MLAs perform similarly with the worst RMSE for GLMNET and MARS. Except springtime, the seasonal reduction of dispersion error is not notable (Fig. 7b). The potential of MLAs to correct GHI is depicted in Fig. 8b. RMSE diminishes substantially with reductions compared to uncorrected GHI_{CAMS} between 9.9 W m^{-2} and 40 W m^{-2}. The lowest RMSEs are calculated in winter. The lowest RMSEs are derived with XGBoost (56 W m^{-2}) and RF (53.6 W m^{-2}), while SVR has the highest RMSE among MLAs (65.1 W m^{-2}). Similar RMSE patterns are also represented in the other seasons.

4.2.3. *Sensitivity analysis of GHI_{AD} to solar zenith angle and cloud fraction*

This section is focused on the sensitivity analysis of the site adapted GHI against various solar zenith angles and cloud fraction cases. The attenuation of solar irradiance depends on the solar zenith angle and the prevailing atmospheric and sky conditions, with clouds, when present, being the most prominent factor. GHI is related to solar zenith angle (SZA) since it controls the attenuation of sunlight traveling through the atmosphere with more favorable phenomena at longer slant paths via the concept of optical air mass (analogous to the factor 1/cos(SZA) in a plane-parallel atmosphere). On the other side, the amount of cloudiness designated by the cloud fraction (CF) is responsible for the absorption and scattering phenomena that occur during sunlight traveling. The effects of SZA and CF on MBE and RMSE for the uncorrected GHI_{CAMS} and GHI_{AD} are represented in Fig. 9. Fig. 9a–d shows the MBE and RMSE disaggregation in terms of SZA and CF, and the MLAs for each SZA and CF case. The errors are calculated using SZA and CF windows of ±5° and ±5%. For example, MBE at SZA = 50° and CF = 20% uses GHIs within [45°–55°) and [15%–25%), respectively.

According to Fig. 9a, the MLAs represent MBEs lower than 18 W m^{-2} at all SZAs with lower values for RF. MBE is substantially reduced with increasing SZA, and for SZA >60°, MBE is lower than 5 W m^{-2}. In general, GHI_{AD} report lower errors than GHI_{CAMS} at all zenith angles, except for the SZA >70°, where the LIN and EQM-adapted GHI fails to follow GHI_{OBS} showing the worst error performance. Especially in the last SZA category (SZA = 80°), a doubling in MBE is calculated for LIN compared to the uncorrected

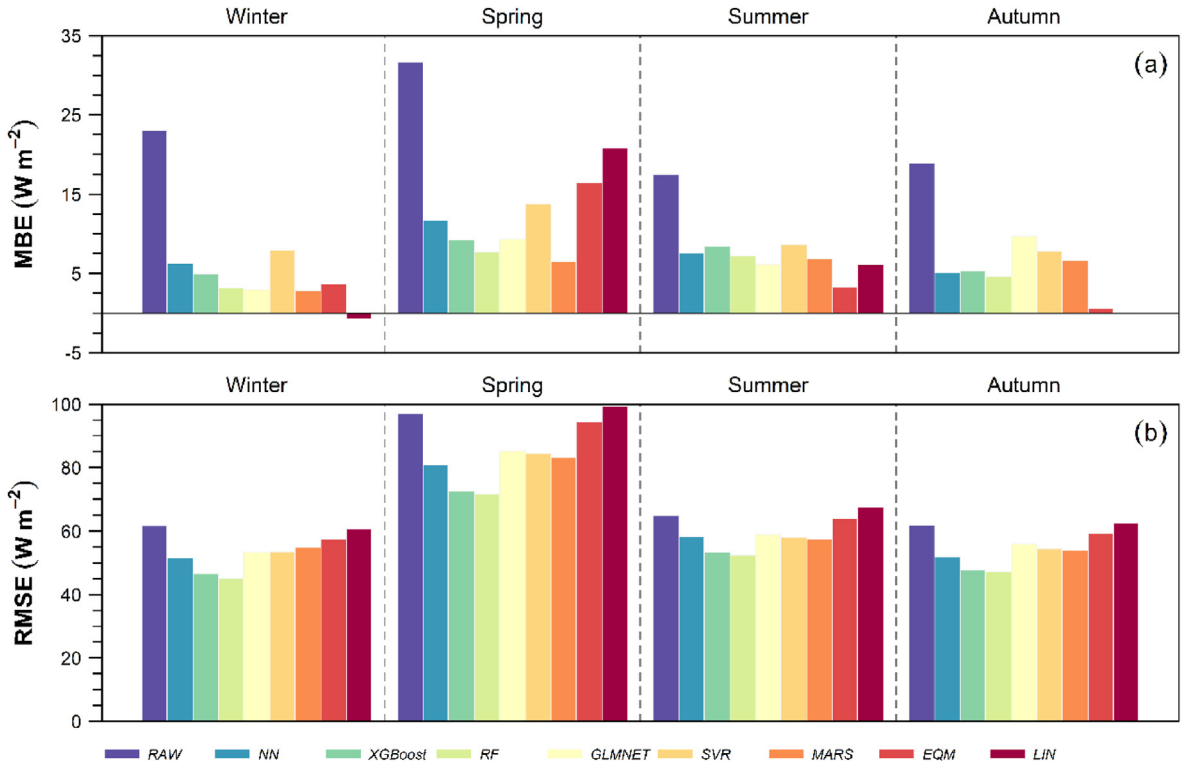

Fig. 6. Seasonal barplots of a) MBE and b) RMSE for the site adapted GHI at all-sky conditions.

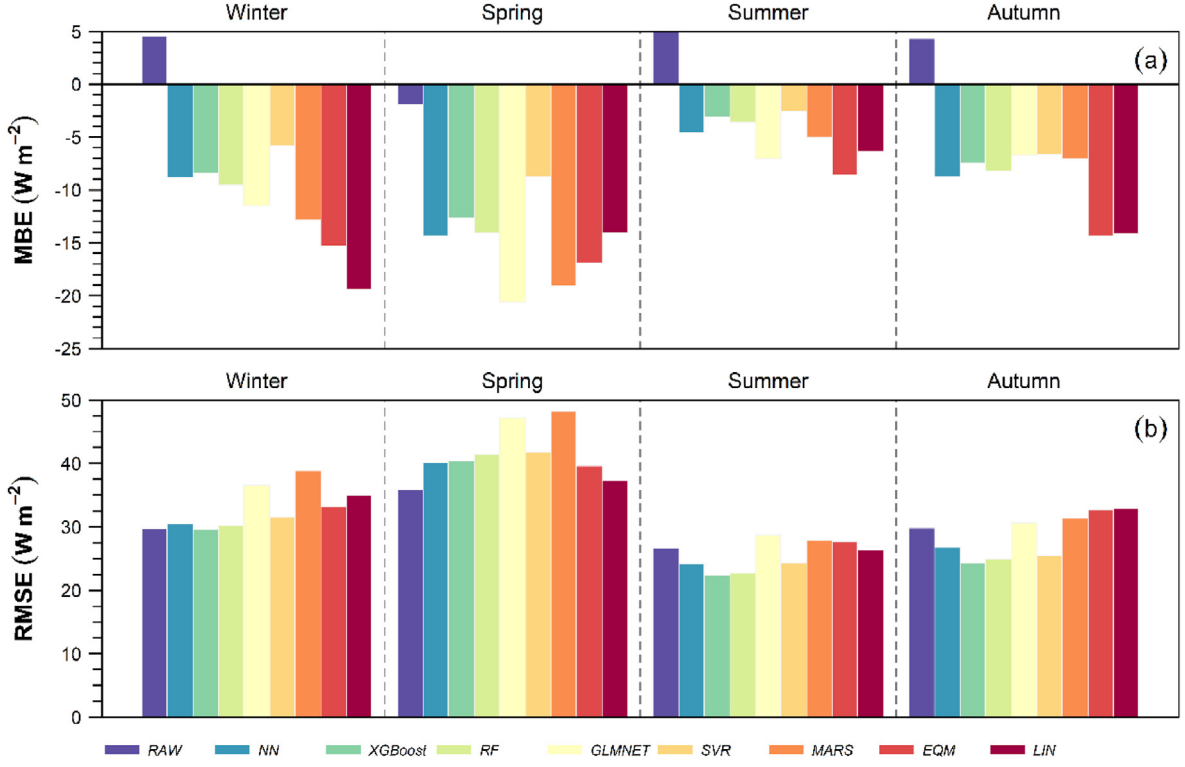

Fig. 7. Seasonal barplots of a) MBE and b) RMSE for the site adapted GHI at clear sky conditions.

GHI_{CAMS}. On the other hand, Fig. 9c represents the systematic biases with respect to cloud fraction as obtained by CAMS-Rad. MBE for MLAs is substantially lower than MBE_{CAMS} for all CF cases (Fig. 9c) with minimal values (<5 W m^{-2}) for CF = 0% and CF = 100% except for GLMNET where MBE for overcast cases exceeds 10 W m^{-2}. MBE_{CAMS} exceeds 25 W m^{-2} for 10% < CF < 90%, while the

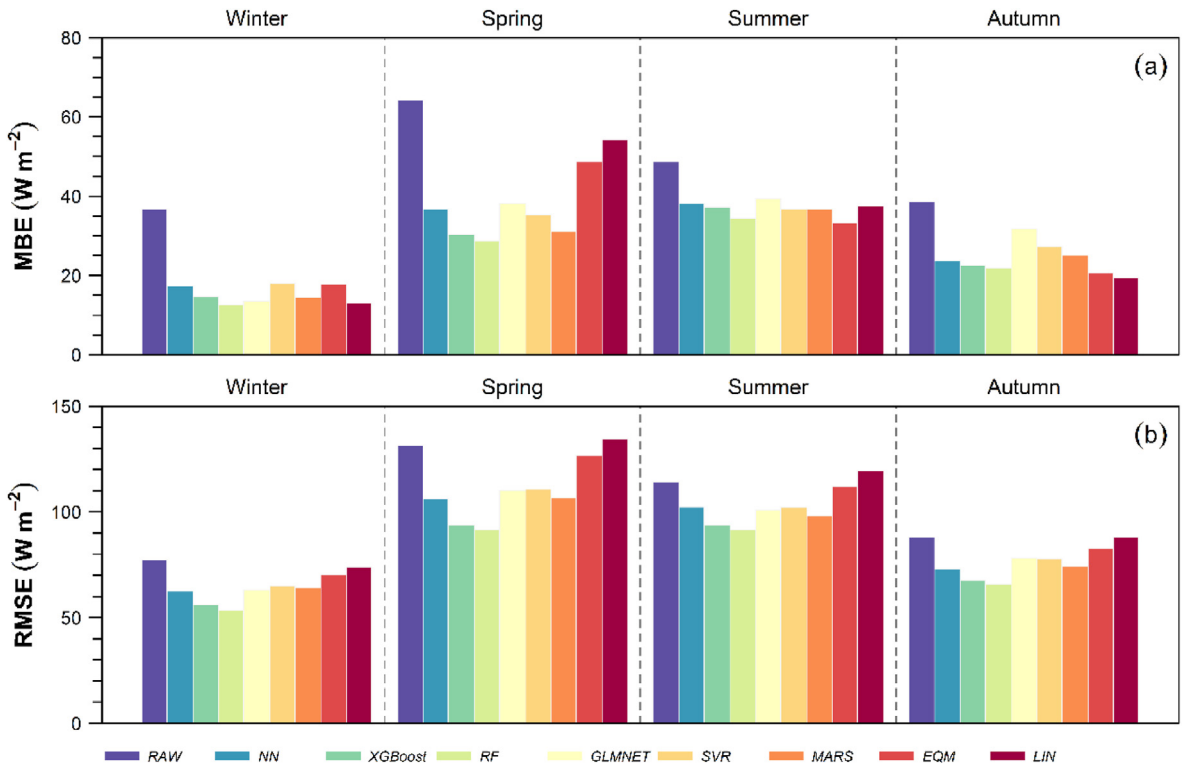

Fig. 8. Seasonal barplots of a) MBE and b) RMSE for the site adapted GHI at cloudy conditions.

corresponding bias is reduced by more than 50% when the ML-based site adapted irradiances are used. The dispersion bias bar plots show the potential of using MLAs for site adaptation (Fig. 9b–d).

RMSE is significantly improved compared to the initial CAMS. The ranges in RMSE difference between MLAs and CAMS ($RMSE_{MLA} - RMSE_{CAMS}$) are NN: −27.5 W m^{-2} to −5.8 W m^{-2}, XGBoost: −36.5 W m^{-2} to −9.3 W m^{-2}, RF: −37.1 W m^{-2} to −10.5 W m^{-2}, GLMNET: −19.6 W m^{-2} to −0.4 W m^{-2}, SVR: −15.5 W m^{-2} to −0.7 W m^{-2}, MARS: −24.4 W m^{-2} to −3.8 W m^{-2}, for the different CF (Fig. 9d) classes. RMSE differences for the selected SZA classes are lower in magnitude. For the traditional site adaptation models (EQM and LIN), high RMSEs are revealed, even comparable or higher than the initial CAMS data. The corresponding RMSE differences, $RMSE_{EQM} - RMSE_{CAMS}$ and $RMSE_{LIN} - RMSE_{CAMS}$, are −3.9 W m^{-2}– 0.1 W m^{-2}, −3.9 W m^{-2}– 0.1 W m^{-2} for SZA and −3.7 W m^{-2} −7.6 W m^{-2}, −5.9 W m^{-2}– −1.1 W m^{-2} for CF.

The collocated effects of SZA and CF to the error performance for the 'best' model is displayed in Fig. 9. The best model is considered as the one with the minimum values of systematic and dispersion metrics. More specifically, the 2-d heatmaps in Fig. 9f and h represent the best model for retrieving the lowest normalized MBE (Fig. 9e) and RMSE (Fig. 9g) for each SZA-CF pair. Therefore, nMBE and nRMSE in each grid cell of Fig. 9e and g are calculated using only those instances falling within the SZA-CF pair. The tree-based MLAs cover a percentage of occurrence equal to 94.8% regarding RMSE, with RF being the model with the highest percentage (58.4%). The heatmap for the optimal model concerning nMBE cannot give a unique result since seven MLAs have 11 to 13 appearances in Fig. 9f nMBE and nRMSE receive the lowest value for CF = 0%, while nRMSE >35% is shown for SZA >70° and a considerable level of cloudiness (CF > 50%). The high errors are probably due to the erroneous extraction of cloud products at high zenith angles as well as modelled uncertainties for high optical paths in the atmosphere. Such errors are induced in the initial calculation of GHI_{CAMS} and cannot be efficiently improved by applying site adaptation.

In general, the state-of-the-art site adaptation models cannot substantially reduce systematic and dispersion errors. Therefore, applying MLAs for site adaptation using numerous exogenous variables is a promising tool and it is highly recommended for local-based correction of solar products. Of course, similar models can be designed to site adapt the other two solar irradiance components (direct and diffuse irradiance) with different ML models configurations and the addition or removal of relevant atmospheric and climate information.

5. Conclusions

Due to the sparsity of radiometric sites worldwide and the increasing energy demand from solar energy sources, time-contiguous and up-to-date solar potential in every place around the globe can be retrieved using reanalysis and satellite-derived solar datasets. The discrepancies between the observed and modelled solar irradiances become critical when satellite-derived solar irradiance is used as the primary solar resource in solar energy applications.

This study presents the validation of GHI_{CAMS} against observations in Patras, Greece and the evaluation of a machine learning framework to adjust GHI_{CAMS} biases for all-sky, clear, and cloudy conditions. A newly proposed method is applied to define the thresholds of clearness index for the classification of the sky state as cloudy, intermediate and clear. GHI_{CAMS} performs efficiently in clear skies because of the high performance of McClear with nMBE <1% and nRMSE <5%. However, significant systematic and dispersion errors exist for all-sky (MBE = 22.8 W m^{-2}, RMSE = 74 W m^{-2}) and cloudy (MBE = 48.2 W m^{-2}, RMSE = 106.5 W m^{-2}) conditions. The calculated discrepancies between GHI_{OBS} and GHI_{CAMS} enable

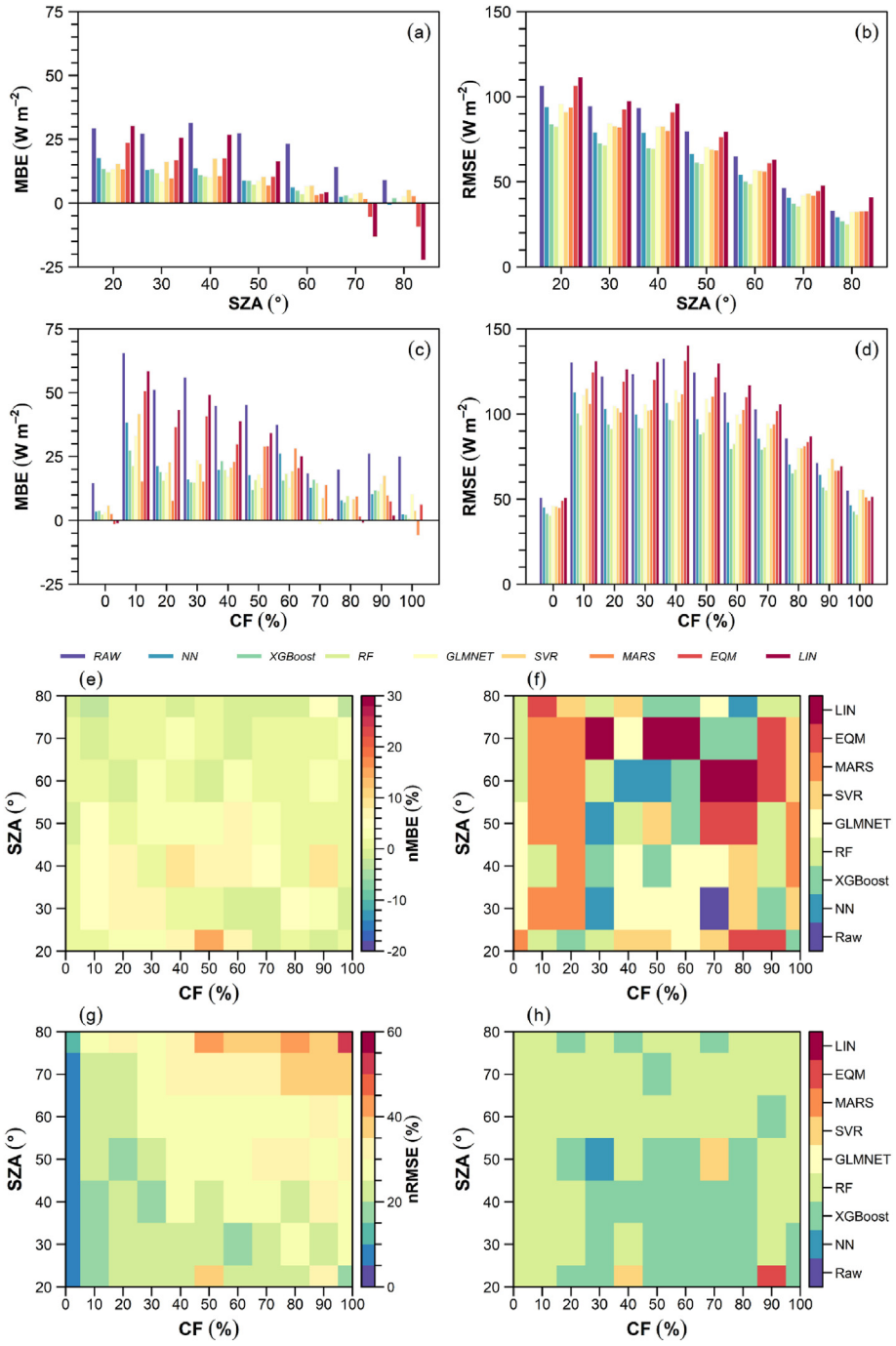

Fig. 9. Bar plots of a) MBE vs. SZA, b) RMSE vs. SZA, c) MBE vs. CF, and d) RMSE vs. CF. Heatmaps of the lowest error metrics e) nMBE and g) nRMSE for different SZA and CF. f) and h) show the best site adaptation model in terms of the lowest MBE and RMSE. nMBE, nRMSE and CF are in % and SZA in degrees.

the site adaptation of GHI_{CAMS} through supervised learning techniques using atmospheric parameters, solar zenith angle and GHI_{CAMS} as inputs. The main findings extracted through the evaluation process are summarized as follows:

- MLAs outperform the state-of-the-art site adaptation methodologies leading to reduced systematic and dispersion errors increasing also the statistical similitude with the observed irradiances.
- MLAs with tree-based prediction mechanism and especially Random Forests can be optimally used for site adaptation of the modelled irradiances. Compared to the other MLAs, those methods provide lower systematic and dispersion error even for the clear sky instances. The seasonal analysis of the error metrics shows similar results with RF being the optimal model.
- Site-adaptation reduces MBE and RMSE at various SZAs and CF cases. The lowest RMSE values are revealed for the tree-based MLAs. The improvement in RMSE extends between -37.1 W m^{-2} and -9.3 W m^{-2}.

Based on the results of this study, new directions can be drawn. The addition of other explanatory variables into the modelling process, the site adaptation of DNI and DHI, and the application of the proposed methodology at sites with different atmospheric conditions

and climate patterns will be future steps for investigation. Moreover, the application of the proposed framework prior to short-term solar forecasting is already in progress for examining whether and how site adaptation could improve solar radiation forecasts.

CRediT authorship contribution statement

Vasileios Salamalikis: Conceptualization, Software, Methodology, Writing – original draft. **Panayiotis Tzoumanikas:** Software, Methodology, Writing – original draft. **Athanassios A.Argiriou:** Data curation, final, Writing – review & editing, part of the, Supervision. **Andreas Kazantzidis:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme « Human Resources Development, Education and Lifelong Learning 2014–2020» in the context of the project "Spatio-temporal variations in the estimation and forecasting of cloudiness and solar radiation using high-resolution satellite and ground-based data (HIRES)" (MIS 81118).

Appendix

Table A1
Cartesian hyperparameter space for supervised learning techniques.

Method	Hyperparameter	Lower	Upper	By	Trans. function
Neural Networks (NN)					
	Hidden layers	1	4	1	–
	Hidden nodes	4	8	1	2^x
	Dropout fraction	0	0.5	0.1	–
	ℓ_2 regularization	1	15	–	–
	learning rate	−4	−1	1	10^x
Extreme Gradient Boosting Machines (XGBoost)					
	nrounds	1	1000	–	–
	eta	−10	0	–	2^x
	subsample	0.1	1	–	–
	max_depth	1	15	–	–
	min_child_weight	0	7	–	2^x
	colsample_per_tree	0	1	–	–
	colsample_by_level	0	1	–	–
	lambda	−10	10	–	2^x
	alpha	−10	10	–	2^x
Random Forests (RF)[1]					
	num.trees	1	1000		
	sample.fraction	0	1		
	mtry	0	1		x*p
	min.node.size	0	1		n^x
Elastic Net regression (GLMNET)					
	alpha	0	1		
	lambda	−10	10		2^x
Support Vector Regression (SVR)					
	cost	−10	10		2^x
	gamma	−10	10		2^x
	degree	1	5		–
Multivariate Adaptive Regression Splines (MARS)					
	degree	1	3		–
	nprune	1	$\binom{p+3}{3}$		–

[1]p – number of predictors and n – number of training samples.

Table A2
Optimal hyperparameters sets for supervised machine learning techniques.

Method	Hyperparameter	Sky State Upper	
		Clear	Cloudy
Neural Networks (NN)			
	Hidden layers	3	3
	Hidden nodes	(64, 32, 64)	(128, 32, 64)
	Dropout fraction	(0.1, 0, 0.2)	(0.2, 0.3, 0.3)
	ℓ_2 regularization	(0.01, 0.001, 0.001)	(0.001,0.001,0.001)
	learning rate	0.01	0.001
Extreme Gradient Boosting Machines (XGBoost)			
	nrounds	854	402
	eta	0.00413	0.0306
	subsample	0.971	0.955
	max_depth	8	11
	min_child_weight	17.9	5.82
	colsample_per_tree	0.844	0.891
	colsample_by_level	0.723	0.115
	lambda	26.9	4.53
	alpha	0.211	17.9
Random Forests (RF)[1]			
	num.trees	234	872
	sample.fraction	0.911	0.885
	mtry	2	6
	min.node.size	2	2
Elastic Net regression (GLMNET)			
	alpha	0.179	0.0355
	lambda	0.00291	0.147
Support Vector Regression (SVR)			
	cost	15.1	25.5
	gamma	0.133	0.136
	degree	0.0543	0.633
Multivariate Adaptive Regression Splines (MARS)			
	degree	1	2
	nprune	11	13

[1]p – number of predictors and n – number of training samples.

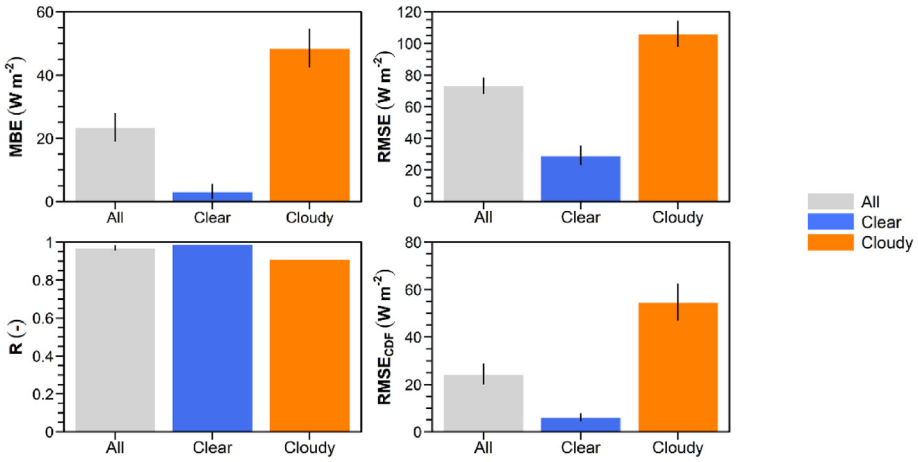

Fig. A1. Barplots of the statistical indicators for the hourly GHI using all possible combinations of 2 years of data from January 2014 to December 2020. The height and the vertical segments in bars correspond to the median and the interquartile range (IQR = $Q_{75\%}$–$Q_{25\%}$).

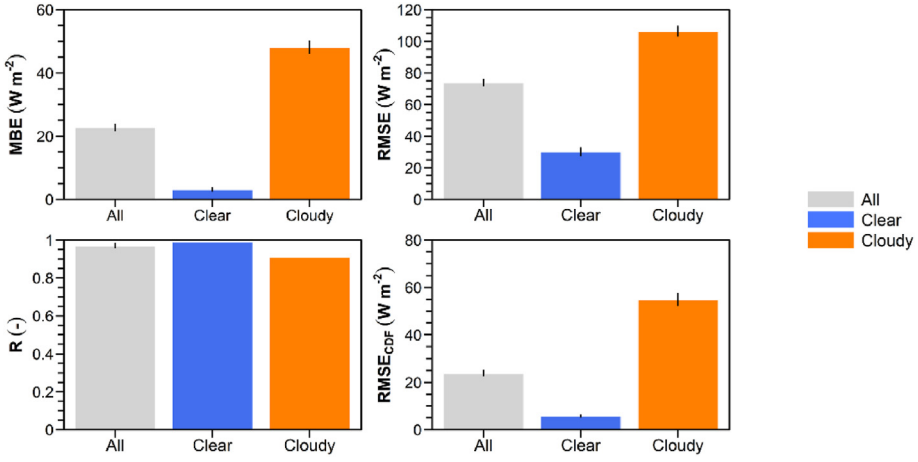

Fig. A2. Barplots of the statistical indicators for the hourly GHI using an iterative procedure with period lengths of approximately two years of data from January 2014 to December 2020. The selected periods cover approximately 7400 hourly values. The height and the vertical segments in bars correspond to the median and the interquartile range (IQR = $Q_{75\%}$–$Q_{25\%}$).

References

[1] J. Polo, R. Perez, Solar radiation modeling from satellite imagery, in: J. Polo, L. Martín-Pomares, A. Sanfilippo (Eds.), Solar Resources Mapping. Green Energy and Technology, Springer, Cham, 2019, pp. 183–197, https://doi.org/10.1007/978-3-319-97484-2_6.

[2] P.A. Jiménez, J.A. Lee, B. Kosovic, S.E. Haupt, Solar resource evaluation with Numerical weather prediction models, in: J. Polo, L. Martín-Pomares, A. Sanfilippo (Eds.), Solar Resources Mapping. Green Energy and Technology, Springer, Cham, 2019, pp. 199–219, https://doi.org/10.1007/978-3-319-97484-2_7.

[3] B. Jia, Z. Xie, A. Dai, C. Shi, F. Chen, Evaluation of satellite and reanalysis products of downward surface solar radiation over East Asia: spatial and seasonal variations, J. Geophys. Res. Atmos. 118 (2013) 3431–3446, https://doi.org/10.1002/jgrd.50353.

[4] A. Boilley, L. Wald, Comparison between meteorological reanalyses from ERA-Interim and MERRA and measurements of daily solar irradiation at surface, Renew. Energy 75 (2015) 135–143, https://doi.org/10.1016/j.renene.2014.09.042.

[5] J. Ruiz-Arias, C. Arbizu-Barrena, F. Santos-Alamillos, J. Tovar-Pescador, D. Pozo-Vazquez, Assessing the surface solar radiation budget in the WRF model: a spatiotemporal analysis of the bias and its causes, Mon. Weather Rev. 144 (2015) 703–711, https://doi.org/10.1175/MWR-D-15-0262.1.

[6] X. Zhang, S. Liang, G. Wang, Y. Yao, B. Jiang, J. Cheng, Evaluation of the reanalysis surface incident shortwave radiation products from NCEP, ECMWF, GSFC, and JMA using satellite and surface observations, Rem. Sens. 8 (225) (2016), https://doi.org/10.3390/rs8030225.

[7] P.D. Jones, C. Harpham, A. Troccoli, B. Gschwind, T. Ranchin, L. Wald, C.M. Goodess, S. Dorling, Using ERA-Interim reanalysis for creating datasets of energy-relevant climate variables, Earth Syst. Sci. Data 9 (2017) 471–495, https://doi.org/10.5194/essd-9-471-2017.

[8] R. Urraca, T. Huld, A. Gracia-Amillo, F.-J. Martinez-de-Pisona, F. Kaspar, A. Sanz-Garcia, Evaluation of global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalyses using ground and satellite based data, Sol. Energy 164 (2018) 339–354, https://doi.org/10.1016/j.solener.2018.02.059.

[9] D. Yang, J.M. Bright, Worldwide validation of 8 satellite-derived and reanalysis solar radiation products: a preliminary evaluation and overall metrics for hourly data over 27 years, Sol. Energy 210 (2020) 3–19, https://doi.org/10.1016/j.solener.2020.04.016.

[10] G.-M. Yagli, D. Yang, O. Gandhi, D. Srinivasan, Can we justify producing univariate machine-learning forecasts with satellite-derived solar irradiance? Appl. Energy 259 (2019), 114122 https://doi.org/10.1016/j.apenergy.2019.114122.

[11] C. Thomas, E. Wey, P. Blanc, L. Wald, Validation of three satellite-derived databases of surface solar radiation using measurements performed at 42 stations in Brazil, Adv. Sci. Res. 13 (2016) 81–86, https://doi.org/10.5194/asr-13-81-2016.

[12] B. Ameen, H. Baltzer, C. Jarvis, E. Wey, C. Thomas, M. Marchand, Validation of hourly global horizontal irradiance for two satellite-derived datasets in Northeast Iraq, Rem. Sens. 10 (2018) 1651, https://doi.org/10.3390/rs10101651.

[13] M. Marchand, A. Ghennioui, E. Wey, E.L. Wald, Comparison of several satellite-derived databases of surface solar radiation against ground measurement in Morocco, Adv. Sci. Res. 15 (2018) 21–29, https://doi.org/10.5194/asr-15-21-2018.

[14] M. Marchand, M. Lefèvre, L. Saboret, E. Wey, L. Wald, Verifying the spatial consistency of the CAMS Radiation Service and HelioClim-3 satellite-derived databases of solar radiation using a dense network of measuring stations: the case of The Netherlands, Adv. Sci. Res. 16 (2019) 103–111, https://doi.org/

10.5194/asr-16-103-2019.
[15] M. Marchand, Y.-M. Saint-Drenan, L. Saboret, E. Wey, L. Wald, Performance of CAMS Radiation Service and HelioClim-3 databases of solar radiation at surface: evaluating the spatial variation in Germany, Adv. Sci. Res. 17 (2020) 143–152, https://doi.org/10.5194/asr-17-143-2020.
[16] T. Cebecauer, M. Súri M, Site-adaptation of satellite-based DNI and GHI time series: overview and SolarGIS approach, AIP Conf. Proc. 1734 (2016), 150002, https://doi.org/10.1063/1.4949234.
[17] E. Lorenz, J. Hurka, D. Heinemann, H. Beyer, Irradiance forecasting for the power prediction of grid-connected photovoltaic systems, IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 2 (1) (2009) 2–10, https://doi.org/10.1109/JSTARS.2009.2020300.
[18] C. Vernay, P. Blanc, S. Pitaval S, Characterizing measurements campaigns for an innovative calibration approach of the global horizontal irradiation estimated by HelioClim-3, Renew. Energy 57 (2013) 339–347, https://doi.org/10.1016/j.renene.2013.01.049.
[19] J. Polo, S. Wilbert, J.A. Ruiz-Arias, R. Meyer, C. Gueymard, M. Súri, L. Martín, T. Mieslinger, P. Blanc, I. Grant, J. Boland, P. Ineichen, J. Remund, R. Escobar, A. Troccoli, M. Sengupta, K.P. Nielsen, D. Renne, N. Geuder, T. Cebecauer T, Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets, Sol. Energy 132 (2016) 25–37, https://doi.org/10.1016/j.solener.2016.03.001.
[20] C.W. Frank, S. Wahl, J.D. Keller, B. Pospichal, A. Hense, S. Crewell, Bias correction of a novel European reanalysis data set for solar energy applications, Sol. Energy 164 (2018) 12–24, https://doi.org/10.1016/j.solener.2018.02.012.
[21] A. Rincón, O. Jorba, M. Frutos, L. Alvarez, F.P. Barrios, J.A. González, Bias correction of global irradiance modelled with weather and research forecasting model over Paraguay, Sol. Energy 170 (2019) 201–211, https://doi.org/10.1016/j.solener.2018.05.061.
[22] L. Mazorra Aguiar, J. Polo, J.M. Vindel, A. Oliver A, Analysis of satellite derived solar irradiance Islands with site adaptation techniques for improving uncertainty, Renew. Energy 135 (2019) 98–107, https://doi.org/10.1016/j.renene.2018.11.099.
[23] J. Polo, C.M. Fernández-Peruchena, V. Salamalikis, L. Mazorra-Aguiar, M. Turpin, L. Martín-Pomares, A. Kazantzidis, P. Blanc, J. Remund, Benchmarking on improvement and site-adaptation techniques for modeled solar radiation datasets, Sol. Energy 201 (2020) 469–479, https://doi.org/10.1016/j.solener.2020.03.040.
[24] C.M. Fernández-Peruchena, J. Polo, M. Martín, L. Mazorra, Site-adaptation of modeled solar radiation data: the SiteAdapt procedure, Rem. Sens. 12 (2020) 2127, https://doi.org/10.3390/rs12132127.
[25] A. Laguarda, G. Giacosa, R. Alonso-Suárez, G. Abala G, Performance of the site-adapted CAMS database and locally adjusted cloud index models for estimating global solar horizontal irradiation over the Pampa Húmeda, Sol. Energy 199 (2020) 295–307, https://doi.org/10.1016/j.solener.2020.02.005.
[26] I. Vamvakas, V. Salamalikis, D. Benitez, A. Al-Salaymeh, S. Bouaichaoui, N. Yassaa, A. Guizani, A. Kazantzidis, Estimation of global horizontal irradiance using satellite-derived data across Middle East-North Africa: the role of aerosol optical properties and site-adaptation methodologies, Renew. Energy 157 (2020) 312–331, https://doi.org/10.1016/j.renene.2020.05.004.
[27] D. Yang, Ensemble model output statistics as a probabilistic site-adaptation tool for satellite-derived and reanalysis solar irradiance, J. Renew. Sustain. Energy 12 (2020a), 016102, https://doi.org/10.1063/1.5134731.
[28] D. Yang, Ensemble model output statistics as a probabilistic site-adaptation tool for solar irradiance: a revisit, J. Renew. Sustain. Energy 105 (2020b) 487–498, https://doi.org/10.1063/5.0010003.
[29] G. Narvaez, L.-F. Giraldo, M. Bressan, A. Pantoja, Machine learning for site-adaptation and solar radiation forecasting, Renew. Energy 167 (2021) 333–342, https://doi.org/10.2016/j.renene.2020.11.089.
[30] A. Kazantzidis, V. Salamalikis, S.-A. Logothetis, I. Vamvakas, Aerosol classification and bias-adjustment of global horizontal irradiance for middle East-North Africa region, AIP Conf. Proc. 2303 (2020), 180002, https://doi.org/10.1063/5.0028544.
[31] D. Yang, Post-processing of NWP forecasts using ground or satellite-derived data through kernel conditional density estimation, J. Renew. Sustain. Energy 11 (2019), 026101, https://doi.org/10.1063/1.5088721.
[32] M. Kottek, J. Grieser, C. Beck, B. Rudolf, F. Rubel, World Map of the Köppen-Geiger climate classification updated, Meteorol. Z. 15 (3) (2006) 259–263, https://doi.org/10.1127/0941-2948/2006/0130.
[33] M.G. Kratzenberg, H.G. Beyer, S. Colle, A. Albertazzi, Uncertainty calculations in pyranometer measurements and application, Sol. Energy ASME (2006) 689–698, https://doi.org/10.1115/ISEC2006-99168.
[34] G. Kosmopoulos, V. Salamalikis, S.N. Pandis, P. Yannopoulos, A.A. Bloutsos, A. Kazantzidis, Low-cost sensors for measuring airborne particulate matter: field evaluation and calibration at a South-Eastern European site, Sci. Total Environ. 748 (2020), 141396, https://doi.org/10.1016/j.scitotenv.2020.141396.
[35] Z. Qu, A. Oumbe, P. Blanc, B. Espinar, G. Gessel, B. Gschwind, L. Klüser, M. Lefèvre, L. Saboret, M. Schroedter-Homscheidt, L. Wald, Fast radiative transfer parameterisation for assessing the surface solar irradiance: the Heliosat-4 method, Meteorol. Z. 26 (1) (2017) 33–57, https://doi.org/10.1127/metz/2016/0781.
[36] M. Lefèvre, A. Oumbe, P. Blanc, B. Espinar, B. Gschwind, Z. Qu, L. Wald, M. Schroedter-Homscheidt, C. Hoyer-Klick, A. Arola, A. Benedetti, J.W. Kaiser, J.-J. Morcrette, McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions, Atmos. Meas. Tech. 6 (2013) 2403–2418, https://doi.org/10.5194/amt-6-2403-2013.
[37] B. Gschwind, L. Wald, P. Blanc, M. Lefèvre, M. Schroedter-Homscheidt, A. Arola, Improving the McClear model estimating the downwelling solar radiation at ground level in cloud-free conditions – McClear-v3, Meteorol. Z. 28 (2) (2019) 147–163, https://doi.org/10.1127/metz/2019/0946.
[38] A. Oumbe, Z. Qu, P. Blanc, M. Lefèvre, L. Wald, S. Cros, Decoupling the effects of clear atmosphere and clouds to simplify calculations of the broadband solar irradiance at ground level, Geosci. Model Dev. (GMD) 7 (2014) 1661–1669, https://doi.org/10.5194/gmd-7-1661-2014.
[39] P. Blanc, L. Wald, The SG2 algorithm for a fast and accurate computation of the position of the Sun for multi-decadal time period, Sol. Energy 86 (10) (2012) 3072–3083, https://doi.org/10.1016/j.solener.2012.07.018.
[40] M. Schroedter-Homscheidt, C. Hoyer-Klick, N. Killius, J. Betcke, M. Lefèvre, L. Wald, E. Wey, L. Saboret, User Guide to the CAMS Radiation Service (CRS), Status December 2020, 2020.
[41] P. Ineichen, Validation of models that estimate the clear sky global and beam solar irradiance, Sol. Energy 132 (2016) 332–344, https://doi.org/10.1016/j.solener.2016.03.017.
[42] J.M. Bright, X. Sun, C.A. Gueymard, B. Acord, P. Wang, N.A. Engerer, BRIGHT-SUN: a globally applicable 1-min irradiance clear sky detection model, Renew. Sustain. Energy Rev. 121 (2020), 109706, https://doi.org/10.1016/j.rser.2020.109706.
[43] P. Ineichen, C.S. Barroso, B. Geiger, R. Hollmann, A. Marsouin, R. Mueller, Satellite application facilities irradiance products: hourly time step comparison and validation over Europe, Int. J. Rem. Sens. 30 (21) (2009) 332–344, https://doi.org/10.1080/01431160802680560.
[44] R. Perez, P. Ineichen, R. Seals, A. Zelenka, Making full use of the clearness index for parametrizing hourly insolation conditions, Sol. Energy 45 (2) (1990) 111–114, https://doi.org/10.1016/0038-092X(90)90036-C.
[45] F. Kasten, A.T. Young, Revised optical air mass tables and approximation formula, Appl. Opt. 28 (1989) 4735–4738, https://doi.org/10.1364/AO.28.004735.
[46] I. Visser, M. Speekenbrink, depmixS4: an R package for hidden Markov models, J. Stat. Software 36 (7) (2010) 1–21, https://doi.org/10.18637/jss.v036.i07.
[47] B.D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, 2007.
[48] Courville I. J, Y. Goodfellow, A. Bengio, Deep Learning, MIT Press, 2016.
[49] T.Q. Chen, C. Guestrin, XGBoost: a Scalable Tree Boosting System, 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, https://doi.org/10.48550/arXiv.1603.02754.
[50] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[51] J. Friedman T, R.Tibshirani Hastie, Regularization paths for generalized linear models via coordinate descent, J. Stat. Software 33 (2010) 1–22.
[53] Friedman T. Hastie R, J.H. Tibshirani, The Elements Os Statistical Learning: Data Mining, Inference , and Prediction, second ed., Springer, New York, 2009.
[52] J.H. Friedman, Multivariate adaptive regression splines, Ann. Stat. 19 (1) (1991) 1–67.
[54] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.
[55] P. Probst, A.-L. Boulesteix, B. Bischl, Tunability: importance of hyperparameters of machine learning algorithms, J. Mach. Learn. Res. 20 (2019) 1–32, https://doi.org/10.48550/arXiv.1802.09596.
[56] F. Chollet, keras (2015). https://github.com/fchollet/keras.
[57] T. O'Malley, E. Bursztein, J. Long, F. Chollet, KerasTuner, 2019. https://github.com/keras-team/keras-tuner.
[58] M. Kuhn, Building predictive models in R using the caret package, J. Stat. Softw. Art. 28 (5) (2008) 1–26, https://doi.org/10.18637/jss.v028.i05. https://www.jstatsoft.org/v028/i05.
[59] B. Bischl M, L. Lang, J. Kotthoff, J. Schiffner, E. Richter, G. Studerus, Z. Casalicchio, J. Jones, Mlr: machine learning in R, J. Mach. Learn. Res. 17 (170) (2016) 1–5.
[60] C.A. Gueymard, Clear-sky irradiance predictions for solar resource mapping and large-scale applications: improved validation methodology and detailed performance analysis of 18 broadband radiative models, Sol. Energy 86 (8) (2014) 2145–2169, https://doi.org/10.1016/j.solener.2011.11.011.