This is a postprint version of the following published document:

Aler R., Galván, I.M., Ruiz-Arias, Jose A., Gueymard, Christian A. (2017). Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting. *Solar Energy*, (150), pp. 558-569

DOI: [10.1016/j.solener.2017.05.018](https://doi.org/10.1016/j.solener.2017.05.018)

© 2017 Elsevier Ltd. All rights reserved.

This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](https://creativecommons.org/licenses/by-nc-nd/4.0/).

Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting

Ricardo Aler^a, Inés M. Galván^a, Jose A. Ruiz-Arias^{b,*}, Christian A. Gueymard^c

^a Computer Science Department, Carlos III University of Madrid, Spain

^b Solargis, Bratislava, Slovakia

^c Solar Consulting Services, Colebrook, NH, USA

ABSTRACT

Based on a large and recently developed database of 1 min irradiance and ancillary data observations at 54 world stations, this study uses the gradient boosting Machine Learning (ML) technique to improve the process of components separation, through which the direct and diffuse solar radiation components are estimated from 1 min global horizontal irradiance data. Here, the XGBoost implementation of gradient boosting is used both with ensembles of linear and ensembles of non linear weak prediction models. The predictions of 140 separation models of the literature are combined using XGBoost to overall improve the random errors of the predictions of the individual separation models at any of the validation sites. The minimum prediction error is essentially achieved by a combination of 26 out of the original 140 models, with no meaningful reduction in error by combining more models. Most of these 26 models use at least three inputs in addition to clearness index. In parallel, XGBoost is also used to separate the components directly from the inputs to the separation models. From the 24 possible inputs used in the original 140 separation models, only 14 are found relevant. These 14 inputs could be used with appropriate formalism to subsequently develop a better separation model. It is found that when the training and validation datasets are not collocated, the RMSD of the predictions increases, on average, 2% with respect to the case of collocated datasets. Overall, the present results indicate that a data driven ML approach combining a limited number of existing models can be used to considerably decrease the currently large random errors associated with such models when used separately at high temporal frequency.

Keywords:

Direct-diffuse separation

DNI

Machine learning

Gradient boosting

1. Introduction

Knowledge of the global horizontal irradiance (GHI) at the planetary surface is important in many disciplines, including agriculture, forestry, hydrology, or climatology. GHI is obviously also an essential variable for all solar energy applications. For them, however, additional information is most generally necessary, in the form of the components of GHI: Direct horizontal irradiance and diffuse horizontal irradiance (DIF). Nonetheless, because the direct component is typically measured with a pyrheliometer on a plane normal to the sun rays, direct irradiance is most usually reported in terms of direct normal irradiance (DNI). Knowledge of DNI alone is essential to evaluate the solar resource for any type of solar concentrating thermal or photovoltaic technology. The simultaneous knowledge of DNI and DIF is required to evaluate the irradiance incident on any receiver, such as photovoltaic panels,

thermal collectors, or building fenestration. Typically, the two components are separately necessary to evaluate the global tilted irradiance (GTI) incident on the plane of array (POA) in flat plate photovoltaic (PV) technologies, for instance. In that case, GTI is the magnitude required to evaluate the energy production of such collectors. Since GTI measurements are normally not available at the POA before a solar power plant is designed, GTI has to be modeled from the direct and diffuse components with an appropriate transposition model. Previous investigations (e.g., Cucumo et al., 2007; Gueymard, 2009; Yang et al., 2013) have shown that GTI's accuracy is in large part conditioned by that in the main inputs to the transposition model, i.e., DNI and DIF.

GHI is usually measured with pyranometers at a large number of radiometric stations over the world. In contrast, the measurement of DNI or DIF is more difficult and costly than that of GHI (since it normally requires a sun tracker and additional equipment), making such data scarce. As a result, it is common to use separation models to estimate the direct and diffuse irradiances from GHI observations. Moreover, GHI and its components are

* Corresponding author.

E-mail address: jararias@ujaen.es (J.A. Ruiz-Arias).

often required in remote areas with potential for solar energy exploitation but with complete lack of solar radiation measurements. In that case, solar radiation has to be modeled from ancillary observations, including those obtained from space. Nowadays, the most widespread modeling method uses cloud reflectance imagery from satellite sensors combined with semi-physical modeling approaches (Perez et al., 2013). Unfortunately, these approaches only allow the estimation of GHI, thus making the use of separation models necessary to evaluate its components. This additional step contributes to the overall uncertainty of the method (Cebecauer et al., 2011).

The need to model the components of GHI in basically all solar energy applications has fueled a relentless interest for the development of empirical separation methods, which started with the seminal study of Liu and Jordan (1960). Countless engineering-type relationships (usually referred to as "separation", "decomposition" or even "diffuse fraction" models for a reason that will become clear in Section 2.1) have thus been regularly proposed in the literature. Based on locally measured data from one or many stations in diverse climatic regions of the world, these models target various solar radiation temporal resolutions (e.g., hourly, daily or monthly). However, because of their underlying empirical nature, their application is normally considered "safe" only over those areas where they were initially developed. Spatial extrapolation could lead to significant regional deviations in results. Hence, the search for a "universal" formula i.e., a model providing accurate results at any location must rely on validation studies undertaken at widely different sites. Many such studies have appeared over the years and have typically involved a relatively small number of relationships at a limited number of sites. Among the literature of the last ten years, the reader is referred to, e.g., (Bertrand et al., 2015; Bilbao et al., 2014; Dervishi and Mahdavi, 2012; Engerer, 2015; Gueymard, 2010; Gueymard and Ruiz Arias, 2014; Ineichen, 2008; Magarreiro et al., 2014; Padovan et al., 2014; Tapakis et al., 2016a; Yang et al., 2013). Recently, Gueymard and Ruiz Arias (2016), hereafter GRA16, have expanded this concept of validation to test 140 separation models at 54 research class stations under 4 climate groups over 7 continents. The results of that study by far the most ambitious of its kind showed that none of the single separation model then analyzed could be considered as truly "universal", although a few of them could generate reasonable estimates of DNI at 1 min resolution at most sites, at least in the absence of snow. The study also clearly indicated which models could perform well under which specific climate.

In parallel, the use of Machine Learning (ML) techniques has rapidly developed in recent years, with the potential of providing better empirical relationships than the usual regressive techniques. In particular, the artificial neural network (ANN) approach has been shown to improve the determination of DIF at various time scales (Alam et al., 2009; Elminir et al., 2007; Tapakis et al., 2016b), but its use requires a higher expertise than more traditional approaches, while still not guaranteeing good "universal" results.

Following the study in GRA16, and using the same sets of both observational data and separation models, as briefly described in Section 2, the potential benefits of the use of ML over the traditional methods are investigated in various ways. In particular, it can be construed that an ML-based combination of different separation models could expand their individual applicability from just one climate zone to many, toward the goal of developing a truly "universal" model. This avenue is explored here in detail, using a technique known as gradient boosting (Friedman, 2001, 2002), whose details are provided in Section 3. An approach for the combination of existing separation models is further developed in Sections 4 and 5, with the goal of improving their individual results. In particular, the capability of gradient boosting for ranking the impact of

its input variables is advantageously used to ascertain which separation models are more relevant. Taking advantage of the availability of data from many different climatic regions, another avenue of research is also developed by constructing a new empirical model which, given the same inputs as the existing separation models, is able to predict DNI directly. This direct model construction is developed in Sections 4 and 6. Additionally, in this case, a study about the relevance of the possible input variables is carried out. Finally, Section 7 examines the models' performance at locations different from those where they have been trained. This constitutes an essential step in the process of verifying the desired "universal" quality of such a model.

2. Separation models and observational data

2.1. Separation models

The complete list of separation models, with detailed description, can be found in GRA16 and references therein. The list encompasses models developed since the 1960s, starting with the pioneering work of Liu and Jordan (1960). Most of them have been devised to separate hourly GHI data; however, some were also developed for 1 min, daily, or monthly average values. Note, however, that they have all been tested in GRA16 using data at high temporal resolution (1 min time steps most generally), which is a time frame with increasing interest in the current context of solar applications (Bright et al., 2015; Engerer, 2015; Fernández Peruchena et al., 2015; Gansler et al., 1995; Ngoko et al., 2014; Yang et al., 2015).

The main modeling approach among the existing separation models is the calculation of the diffuse fraction, $K = \text{DIF}/\text{GHI}$, from the clearness index, $K_T = \text{GHI}/E_{0h}$, where E_{0h} is the extraterrestrial horizontal irradiance readily calculated for the same time and location as GHI. Many models also use additional variables such as solar zenith angle, optical air mass, temperature, relative humidity, or custom GHI variability indexes, to name just a few. It is worth mentioning that a few models rather predict the fraction of direct irradiance (i.e., DNI/GHI or $\text{DNI} \cos(Z)/\text{GHI}$, depending on model) instead of K , where Z is the solar zenith angle. From a pure physical standpoint, the two approaches are equivalent by consideration of the closure equation $\text{GHI} = \text{DNI} \cos(Z) + \text{DIF}$.

In this work, the individual 1 min predictions of 140 separation models analyzed in GRA16 at the 54 validation sites described in Section 2.2 are combined together using the AI techniques presented in Section 3.

2.2. Observational data

The observational dataset, which is thoroughly described in Section 3 of GRA16, and to which the reader is referred for further information, consists of independent observations of the three solar radiation components (i.e., GHI, DNI and DIF) at 54 radiometric stations and at 1 min resolution for the vast majority of them. The observations are gathered using thermopile radiometers. DIF is measured with a shading ball attached to the sun tracker also supporting the pyrheliometer measuring DNI. Table 1, adapted from Table 3 in GRA16, summarizes important information about the 54 stations used in the present work. Most of the stations (49) belong to the Baseline Surface Radiation Network (BSRN, <http://bsrn.awi.de/>; Ohmura et al., 1998), a project of the Global Energy and Water Cycle Experiment (GEWEX) under the umbrella of the World Climate Research Programme (WCRP). The rest is obtained at stations operated by the National Renewable Energy Laboratory (NREL) (three stations), the German Aerospace Agency (DLR) (one station), and the Masdar Institute (one station). The

Table 1
Information on the 54 stations used for validation, including: station code, latitude and longitude in degrees, climate zones (acronyms are: AR, HA, TM, and TR for Arid, High albedo, Temperate, and Tropical, respectively), and mean measured values of GHI and DNI in W/m².

Code	Station	Latitude	Longitude	Climate	Mean GHI	Mean DNI
ALE	Alert	82.49	-62.42	HA	234.5	389.0
ASP	Alice Springs	-23.80	133.89	AR	499.5	611.2
BER	Bermuda	32.27	-64.67	TM	548.8	448.0
BIL	Billings	36.61	-97.52	TM	486.5	504.1
BON	Bondville	40.07	-88.37	TM	592.6	533.7
BOU	Boulder	40.05	-105.01	TM	529.8	529.3
BRB	Brasilia	-15.60	-47.71	TR	428.8	361.6
CAR	Carpentras	44.08	5.06	TM	399.1	476.8
CNR	Cener	42.82	-1.60	TM	403.6	421.0
CLH	Chesapeake Light	36.91	-75.71	TM	405.7	417.5
COC	Cocos Island	-12.19	96.84	TR	498.2	410.9
DOM	Concordia Station	-75.10	123.38	HA	399.1	846.0
DAR	Darwin	-12.43	130.89	TR	531.5	469.6
DWN	Darwin Met Office	-12.42	130.89	TR	481.9	450.1
DAA	De Aar	-30.67	23.99	AR	518.7	656.0
DRA	Desert Rock	36.63	-116.02	AR	688.9	794.6
EUR	Eureka	79.99	-85.94	HA	221.3	304.1
FLO	Florianapolis	-27.53	-48.52	TM	446.6	438.6
FPE	Fort Peck	48.32	-105.10	TM	583.1	576.8
FUA	Fukuoka	33.58	130.38	TM	350.7	273.4
GVN	Georg von Neumayer	-70.65	-8.25	HA	308.1	272.4
GOB	Bobabeb	-23.56	15.04	AR	582.2	732.1
GOL	Golden-NREL	39.74	-105.18	TM	458.2	558.0
GCR	Goodwin Creek	34.25	-89.87	TM	564.3	514.6
ILO	Ilorin	8.53	4.57	TR	256.6	72.1
ISH	Ishigakijima	24.34	124.16	TM	367.7	246.2
IZA	Izaña	28.31	-16.50	AR	617.9	791.4
KWA	Kwajalein	8.72	167.73	TR	544.0	433.5
LAU	Lauder	-45.05	169.69	TM	371.7	402.1
LER	Lerwick	60.14	-1.19	TM	204.6	131.0
LIN	Lindenberg	52.21	14.12	TM	279.9	255.8
MAS	Masdar	24.44	54.62	AR	506.0	552.5
MNM	Minamitorishima	24.29	153.98	TM	477.7	451.9
MAN	Momote	-2.06	147.43	TR	504.2	383.7
NAU	Nauru Island	-0.52	166.92	TR	539.5	439.7
NYA	Ny-Alesund	78.93	11.93	HA	189.8	209.9
PAL	Palaiseau	48.71	2.21	TM	304.7	280.2
PAY	Payerne	46.82	6.94	TM	378.0	356.5
PTR	Petrolina	-9.07	-40.32	TR	528.7	485.4
PSA	PSA-DLR	37.09	-2.36	AR	457.0	89.0
REG	Regina	50.21	-104.71	TM	362.9	394.6
PSU	Rock Springs	40.72	-77.93	TM	499.5	420.1
SMS	Sao Martinho da Serra	-29.44	-53.82	TM	435.4	444.2
SAP	Sapporo	43.06	141.33	TM	325.4	267.6
SBO	Sede Boquer	30.86	34.78	AR	607.2	642.9
SXF	Sioux Falls	43.73	-96.62	TM	582.7	602.3
SOV	Solar Village	24.91	46.40	AR	567.3	580.4
SON	Sonnblick	47.05	12.96	HA	367.7	296.7
TAM	Tamanrasset	22.79	5.53	AR	589.1	622.5
TAT	Tateno	36.05	140.13	TM	332.8	277.2
TIK	Tiksi	71.59	128.92	HA	260.2	238.7
TOR	Toravere	58.25	26.46	TM	239.6	126.4
TUC	Tucson	32.23	-110.96	AR	557.3	694.9
XIA	Xianghe	39.75	116.96	TM	420.8	376.0

dataset spans three years at each station, with the exception of the DLR's PSA station (two years only), and spreads over four broad climatic classes (11 under arid climate, 27 under temperate climate, 9 under tropical climate, and 7 under high albedo climate i.e., sites surrounded by snow most of the time).

The data underwent a quality control process that is based on the BSRN guidelines (Long and Shi, 2008; Roesch et al., 2011), with additional quality tests (GRA16). Among other things, this additional filtering process rejects data points recorded at low solar altitude (<5°), because they are of marginal importance in solar applications and correspond to significantly increased uncertainty in both observations and models.

As commented in Section 2.1, a significant fraction of the pool of 140 separation models uses more inputs than just GHI (or actually K_T). Some of these inputs are readily available deterministic

variables, such as Z or optical air mass, m . Some models also require other atmospheric or surface variables that are not always observed at the same temporal frequency as GHI, or not even observed at all. In particular, many models require temperature and relative humidity observations. These variables are usually observed at lower temporal resolution than the solar irradiance components and are here interpolated to 1 min time step to match the solar radiation time grid. Other variables, such as ground albedo or atmospheric aerosol turbidity, are also required by a few models. Due to lack of local observations, this information is extracted from *model* databases: The ground albedo is obtained from NASA's Modern Era Retrospective Analysis for Research Applications (MERRA) reanalysis, and the Linke turbidity factor from the SoDa service (http://www.soda-is.com/eng/services/climat_free_eng.php#5). Further details are provided in Section 3 of GRA16.

3. Gradient boosting

Gradient Boosting (GB) is a boosting technique for building both classification and regression models (Friedman, 2001, 2002). The term *gradient* refers to the optimization algorithm used during the learning process of the model. As a boosting technique, GB consists of *ensembles* of weak prediction models, hereinafter also referred to as *learners*. In its most common form, the learners are *non linear* models, also known as decision or regression trees, depending on whether the intended application is classification or regression, respectively. In such case, GB is customarily referred to as Gradient Tree Boosting (GTB). As described below, however, the modeling approach followed here also makes use of linear learners (i.e., linear weighted combinations of the inputs plus a constant). For convenience, the regression trees will be hereinafter referred to as *non linear learners*, as opposed to linear learners.

The ensembles of weak prediction models (e.g., regression trees) are built by adding sequentially new learners. Fig. 1 displays an example of a single regression tree being used as a solver for a regression problem. It contains nodes (circles), where conditions on input attributes are checked, and where leaves (squares) yield predictions based on selections in the decision nodes. Individual regression trees can be trained by means of algorithms such as CART (Steinberg and Colla, 2009). Regression trees are weak models in the sense that although they do not have to be very accurate on an individual basis, their joint accuracy can be improved when many of them are put together into an ensemble.

Considering the case of GTB, each new regression tree that is added to the ensemble is trained from the residuals (customarily referred to as *pseudo residuals*) that result from the previous ensemble of regression trees. Therefore, the ensembles are built incrementally by growing partial ensembles with k members, f_k , which are defined as a weighted summation of the k individual learners, h_m , according to:

$$f_k(x) = \sum_{m=1}^k \gamma_m h_m(x) \quad (1)$$

The process is repeated until M regression trees are included in the ensemble. As will be discussed further in this section and in Section 4.3, M is a design parameter of GB machines. The computation of pseudo residuals depends on the actual loss function to be

minimized. The learning process in the new tree h_k attempts to decrease the differences between the partial ensemble $f_{k-1}(x)$ and the actual output y . To that aim, the value in each tree's leaf (Fig. 1) is adjusted and the weight γ_k is computed for the new tree by minimizing the loss function. This process is "greedy", in the sense that each new h_m and its weight γ_k are both optimized solely for the improvement of the previous ensemble f_{k-1} , while keeping f_{k-1} fixed. This process results in a new ensemble $f_k(x)$. The algorithm continues by adding learners until M trees are reached.

In this work the loss function to be minimized is the root mean squared (RMS) error. In this case, the pseudo residuals are $r = y - f_{k-1}(x)$, where y is the actual output and $f_{k-1}(x)$ is the output estimated by the current partial ensemble, constituted of the $k-1$ preceding learners. The initial model h_1 is the numerical value that minimizes the loss function, which for the RMS error loss is simply the average of the dataset outputs. Further details about this algorithm can be found in Friedman (2001, 2002).

In addition, GB can rank the input variables according to their importance for the ensemble. It is based on the criterion used to select the most appropriate variable at the decision nodes. This criterion consists in a measure of the error improvement after consideration of the new variable. The improvement $I(V)$ due to a variable V that splits a parent node P into two children nodes, L and R , is defined as:

$$I(V) = E(P) - qE(L) - (1-q)E(R) \quad (2)$$

where $E(P)$ is the squared error at the parent node (i.e., the error before using V), and $E(L)$ and $E(R)$ are the squared errors at nodes L and R (i.e., the errors after using V). In addition, q is the fraction of instances that go through node L (hence, $1-q$ instances alternatively go through R). For a single tree, the importance of a specific variable is the sum of the improvements from all tree's nodes in which that variable appears (weighted by the corresponding fraction of training data at each node). In the case of an ensemble, the importance of a variable is the average over all trees in the ensemble. Thus, overall, the importance of a variable is relative to the particular ensemble model, and should *not* be taken as a measure of the actual importance of the variable taken individually. This is because the choice of variables at the decision nodes depends on which other variables have been already selected in the tree, or even in previous trees.

In this contribution, the XGBoost implementation of GB is used (Chen and Guestrin, 2016). It is very efficient and scalable. It automatically takes advantage of the available computer processor cores, CPUs, and memory. It is implemented for different languages, such as Python or R, the latter being the package used in this study (Chen and He, 2015). In addition to generating tree based GB models (i.e., non linear learners), XGBoost is also able to perform GB with linear learners, resulting in an ensemble which is itself a linear model.

At training time, XGBoost requires various design parameters, usually referred to as *hyper parameters*, which set out model features such as topology (e.g., number of trees in the ensemble and maximum number of tree nodes) and complexity (e.g., regularization and learning rate). Details on these and other parameters can be found in (Chen and Guestrin, 2016). Only after the hyper parameters are defined, the model can be trained. The approach used here to select the hyper parameters is described in Section 4.

Although GB has found some application in solar power forecasting among other topics (e.g., Huang and Perry, 2016), the reader should note that, from the solar resource community standpoint, GB and XGBoost are tools that still lag behind ANN or other ML techniques in terms of widespread usage. To these authors' knowledge, only one very recent report (which was actually published after the bulk of this report was already written) uses XGBoost in such a non forecasting context in that specific case to evaluate

Fig. 1. Example of regression tree to predict fuel consumption of cars (in liters per 100 km) from the car's weight, number of cylinders and horsepower. Circles are decision nodes where conditions on inputs are checked. Squares—known as leaves—yield the final tree predictions.

the daily GHI at locations where it is not measured (Urraca et al., 2017).

4. Methodology

Two families of XGBoost models are developed in this work: (i) the combination of *existing* separation models to separate GHI in the direct and diffuse irradiances (hereinafter referred to as *indirect model*); and (ii) the separation of GHI using a *new* model uniquely based on GB (hereinafter referred to as *direct model*). In both cases, model versions using linear and non linear weak learners are evaluated. In addition, two validation approaches are considered: (i) using collocated training and validation data; and (ii) using training and validation data from different sites.

4.1. Models construction

For the first family of models, XGBoost is used to obtain a best performing combination of 140 conventional separation models extracted from the pool of models investigated in GRA16 (hereinafter this approach will be referred to as *indirect model*, as opposed to the direct approach introduced below). The aim of the indirect model is to study whether the performance of individual separation models can be improved by combining them efficiently so as to provide one single prediction of even higher accuracy. Here, this combination is learned separately by means of both linear and non linear learners. Their performance is later on compared in Section 5. The specific value of DNI at time t is approximated by:

$$DNI(t) = F(p_1(t), p_2(t) \dots p_{140}(t)) \quad (3)$$

where $p_1(t), p_2(t), \dots, p_{140}(t)$ are the predictions of the 140 separation models at time t . Recall from Section 1 that, once DNI is known, DIF can be computed by subtracting $DNI \cos(Z)$ from GHI.

In parallel, XGBoost is also used to develop a GB separation model from the variables originally used as predictors for the 140 separation models. Hereinafter, this approach will be referred to as *direct model*. In this case, the aim is to ascertain whether the conventional separation models or even their combination using the indirect model described just above can be outperformed by a new GB model learned directly from the same variables that were originally used as inputs to the 140 separation models. To this end, based on GRA16, a total of 24 variables is considered here. This total is made of those 4 variables already defined in Section 2 (E_{oh} , Z , m and K_T), and of 20 additional variables: observed global horizontal irradiance (E), clear sky global horizontal irradiance (E_c), clear sky direct normal irradiance (E_{bnc}), dry bulb temperature (T), dew point temperature (T_{dp}), relative humidity (RH), atmospheric pressure at the surface (P), ground albedo (ρ), modified clearness index (K_{Tp}), mean daily K_T (K_{Tm}), Linke's turbidity (T_L), and nine solar radiation variability indices ($V_1 \dots V_9$). Detailed information on their origin or calculation can be found in GRA16.

Therefore, if D is the direct model, its prediction of DNI at time t would be:

$$DNI(t) = D(x_1(t), x_2(t) \dots x_{24}(t)) \quad (4)$$

where $x_i(t)$ represents the i th of the 24 input variables (mentioned just above) at time t . Like in the case of the indirect model above, the performances of the linear and non linear learners are compared.

4.2. Model training and evaluation

In the context of ML, separate data sets are used to train and evaluate models. The training data set is used to fit the model,

whereas the test data set serves to evaluate its generalization capability. In this work, every model is first trained and then tested on different data for each station. The method to obtain the training and test partitions is described below. Finally, the test results obtained for each station are averaged over all stations to evaluate the overall model's performance.

Each observational station provides radiation data for 3 years (with the only exception of the DLR's PSA station, which features 2 years). The training samples are obtained by aggregating data from the first two thirds of the dataset (in chronological order). The remaining third is used for testing. The training partition that is necessary to build the model (and in particular, to select the hyper parameters, see Section 4.3) is the result of joining the training data from all stations. The same process is also applied for the test partition. Since the size of the training partition is still too large (approximately 17,647,919 samples for the 54 stations combined), it is subsequently sub sampled to maintain reasonable training times. To this aim, the original training dataset is down sampled by randomly selecting 10% of the 1 min samples. In order to reduce the training dataset size even further, one every four days is selected. After all these operations, the final training dataset is reduced to only 459,908 samples. Note that the test data set is not sub sampled, contrarily to what is done with the training data set, and is made of 8,823,962 samples.

The performance of the different GB models generated here is evaluated using the usual mean bias difference (MBD) and RMS difference (RMSD) statistics, to be consistent with the results in GRA16. The comparative results between the conventional separation models and the XGBoost models developed here are presented in terms of the *RMSD skill score* (or, simply, skill score or SS) defined as:

$$SS = \frac{RMSD_{sep} - RMSD_{XGBoost}}{RMSD_{sep}} \quad (5)$$

where $RMSD_{sep}$ and $RMSD_{XGBoost}$ are the RMSD counterparts for any one of the 140 separation models and the XGBoost model, respectively. Positive (negative) skill scores occur whenever the XGBoost model outperforms (underperforms) the conventional separation model. A perfect XGBoost model would result in unity skill scores. A zero skill score occurs when both the conventional separation model and the XGBoost model perform identically (in terms of RMSD).

4.3. Selection of XGBoost's hyper parameters

Since the XGBoost algorithm is controlled by various hyper parameters, a proper selection of their values is important toward obtaining accurate results. To find the best hyper parameters in each case, an exhaustive search over a grid of values is carried out and the best performing case is retained. Table 2 shows the hyper parameters values explored for each type of learners.

In order to compare the performance of XGBoost for different combinations of hyper parameters and ultimately select the best one, a fraction of the training data set is extracted and used as validation set. Samples corresponding to the time span between the first and the 24th day of each month are used to train the model. The remaining samples are used as the validation subset. The best combination of hyper parameters is defined as the one providing the smallest RMSD for the validation set. Then, specifically, the validation subset is used to obtain an independent evaluation of the hyper parameter selections, whereas the separate test subset is reserved for the final model evaluation.

5. Results for the indirect modeling approach

The aim of this section is to report the results obtained with the indirect modeling approach using XGBoost ensembles of both lin

Table 2

Hyper-parameters for linear and non-linear learners, description, and grid search values.

Parameter	Description	Grid values
<i>Linear learners</i>		
M	Number of models in the ensemble	100, 200, 300, 400, 500
Lambda	L2 regularization term on weights	0, 0.25, 0.5, 0.75, 1
Alpha	L1 regularization term on weights	0, 0.25, 0.5, 0.75, 1
Lambda bias	L2 regularization term on bias	0, 0.25, 0.5, 0.75, 1
<i>Non-linear learners</i>		
M	Number of trees in the ensemble	100, 200, 300, 400, 500
Eta	Step-size shrinkage or learning rate	0.05, 0.1, 0.3, 0.5
Max depth	Maximum depth of the tree	1, 2, 3, 4, 5, 6, 7, 8, 9

Table 3

Average and standard deviation (within brackets) of the station-wise RMSD and MBD scores over all stations, expressed in percent after normalization by the mean observed DNI. Results are shown for the linear and non-linear learners in the indirect model and the combination of best separation models at each station.

Boosting method	RMSD	MBD
Linear learners	24.1% (12.8%)	-0.2% (6.6%)
Non-linear learners	20.0% (10.8%)	-0.8% (5.2%)
Combined best separation models	27.6% (16.1%)	2.7% (9.4%)

ear and non linear learners. In addition, the relative importance of each separation model in the indirect model's performance is ranked using specific XGBoost capabilities.

Following the empirical approach described in Section 4.3, the best selection of hyper parameters, when using linear learners, is found to be $M = 500$, $\text{Lambda} = 1$, $\text{Alpha} = 0.5$, and $\text{Lambda bias} = 0$, although other combinations provide similar results. Similarly, for the non linear learners, the best selection is $M = 500$, $\text{Max depth} = 9$, and $\text{Eta} = 0.1$. Again, similar results are obtained with other combinations. Thus, both the linear and non linear learners do not show a strong sensitivity to the selection of hyper parameters. Using the configurations just mentioned, XGBoost is trained and evaluated using the methodology described in Section 4.2.

Table 3 shows the overall averages and standard deviations of station wise RMSD and MBD values. The "combined best separation model" is computed by selecting for each station the separation model that features the lowest RMSD out of the 140 separation models tested here. Thus, it is a best virtual case in the sense that it is made up of the best performance model at each station, which may vary station wise. Overall, it is found that the indirect model using non linear learners is much better than either the one using linear learners or the combination of best separation models. The latter is also outperformed by the indirect model using linear learners. These results hold for both RMSD and MBD. Note also that, as discussed in GRA16, the separation models with the lowest RMSD usually have 5 or 6 input variables. Despite the reduced noise compared to simpler separation models, the results in GRA16 also showed that some of the more complex separation models have a tendency to overpredict DNI, particularly in arid, temperate and high albedo climates. This is consistent with the high MBD value shown in Table 3 for the combined best separation models.

In order to analyze whether XGBoost performs better than any individual separation model, the test RMSD obtained by XGBoost at each station is compared to the test RMSD of the best separation model at that same station using the SS parameter defined in Eq. (5). Figs. 2 and 3 show the SS value for these two RMSD values at each station for the indirect model with both linear and non linear learners, respectively. Fig. 2 reveals that the indirect model with linear learners performs better than the best existing separa-

tion model at most of the stations. However, there are some stations where the best separation model is not improved: At 3 of them (Sede Boquer, Lauder and Brasilia) the results are very similar, whereas at 5 stations (Alice Springs, Toravere, Alert, Concordia Station and Eureka) the indirect model with linear learners is even worse. (Note that the last three are in a high albedo climate.) In contrast, Fig. 3 shows that an indirect GB model based on non linear learners performs always better than the best separation model of the literature at any station. In most cases, this improvement is remarkably larger (on average, twice as much or even more in the case of high albedo sites) than that of the indirect model using linear learners. Interestingly, note that, on average, the model improvement is highest at temperate sites, followed by the arid and tropical sites.

Still using the indirect modeling approach, the third and final aspect evaluated here consists in determining which separation models are contributing the most to the performance of the GB model alternative using non linear learners. The goal here is to determine whether a similar modeling error can be achieved with a smaller subset of separation models. This is equivalent to asking which ones of the 140 separation models are the most relevant for the overall performance of the indirect model. To this aim, 140 different versions of the indirect model with non linear learners are trained adding separation models successively from 1 to 140, ordered according to the ranking provided by XGBoost.

Using the same training and test datasets as before, the training and test RMSDs for each of the 140 indirect model versions are shown in Fig. 4. It can be observed that the same RMSD can be achieved by using only 26 separation models rather than all the 140 original separation models – a remarkable reduction of inputs and complexity by a factor of 5.3. (The training errors with both 26 and 140 separation models are equal to 15.8%, whereas the test errors are 20.1% and 20.0%, respectively.) Hence, a combination of these 26 separation models using non linear learners can be used to reach a similar performance than using the entire set of 140 separation models. The top 26 separation models (following the naming convention of GRA16) are: BUGLER, ENGERER1, KUO1, MAGARREIRO, SKARTVEIT3, BOLAND5, PEREZ1, ENGERER2, POSADILLO5, HAY, SPENCER, PEREZ2, KUO4, REINDI, TAMURA, CHENDO3, CJCUMO, HOLLANDS2, RIDLEY2, BOLAND2, OUMBE, TURNER, TAPAKIS2, STALTER, SKARTVEIT and SUEHRKE. Details on each model can be found in Tables 1 and 2 of GRA16. As underlined in tion 3, the ordering obtained here is purely relative to the XGBoost model, and thus should not be interpreted as a measure of the importance of the separation models taken individually. Thus, a separation model might appear here with a much higher ranking compared to some other, even though the latter may be known of equal or better irradiance separation performance, based e.g. on conventional validation tests of the literature. This explains why this ranking is apparently quite different from that in GRA16, for instance. Note that a much faster indirect GB model could be constructed by using only the 26 "best" models listed above with, virtually, no loss of accuracy. Out of the first 10 separation models, seven have at least 4 inputs. In contrast, only four of the remaining ones have 4 or more inputs. All the 26 top models listed above use more inputs than just K_T . Hence, in general, separation models with many inputs (thus, more complex) have higher impact on the overall performance than those that depend only on K_T , which is consistent with most results reported in the literature.

6. Results of the direct modeling approach

As in the previous section, an exhaustive computational search is carried out, first to establish the best combination of hyper parameters and then to train the XGBoost models. In this case, the best values obtained for the hyper parameters of the model

Fig. 2. RMSD skill score of the indirect model using linear learners compared to the best separation model at each station. The sites are grouped by prevailing climate. The horizontal thick line marks the average skill score for each climatic group. A color version of this figure is available in the on-line version of this paper.

Fig. 3. Same as Fig. 2, but for the indirect model using non-linear learners. A color version of this figure is available in the on-line version of this paper.

using linear learners are: $M = 500$, $\text{Lambda} = 0.25$, $\text{Alpha} = 0$ and $\text{Lambda bias} = 0$. Similarly, for the model using non linear learners, the best values are: $M = 500$, $\text{Max depth} = 9$ and $\text{Eta} = 0.05$. Like with the indirect model, other hyper parameter combinations appear to provide similar results. The training of the direct model with linear and non linear learners using the best combination of hyper parameters is conducted as described in Section 4.2.

Table 4 shows the overall average and standard deviation of the station wise RMSD and MBD for the GB model versions using both the linear and non linear learners, as well as those for the combined best separation models (same separation models as in Table 3), following the methodology described in Section 5. It is observed that the direct model using non linear boosters performs much better than the one using linear boosters in terms of RMSD. However, their MBDs are similar, although the standard deviation using the non linear boosters is smaller.

Compared to the indirect modeling approach evaluated in the previous section (Table 3), the direct model using linear learners now performs remarkably worse than the direct model using non linear learners. In terms of RMSD, the direct model using linear learners is even worse than the combination of best predictors, contrarily to what happens with the indirect model, in which the linear learners yield a lower RMSD than the combination of best predictors. This result had actually to be expected, because the indirect model has available inputs (the 140 separation models) that are already non linear transformations of the original variables (x_1 to x_{24}), whereas the linear learners in the direct modeling approach can only build linear combinations of x_i . In contrast, the direct model using non linear learners performs similarly as its indirect model counterpart, since it internally features non linear transformations of the original variables.

Fig. 4. Training and test RMSD values for the indirect model using non-linear learners trained with the n best separation models, out of the 140 ranked ones. A color version of this figure is available in the on-line version of this paper.

Table 4
Same as Table 3, but for the direct model.

Boosting method	RMSD	MBD
Linear learners	33.9% (18.9%)	0.5% (12.8%)
Non-linear learners	19.2% (11.3%)	-0.6% (5.8%)
Combined best separation models	27.6% (16.1%)	2.7% (9.4%)

In order to analyze the performance of the direct models at individual stations (both for the linear and non linear learners), their predictions are compared with those of the separation model with lowest RMSD at each station using the SS parameter as in Section 5. Figs. 5 and 6 display the SS at each station for the direct model using linear and non linear learners, respectively, compared to the best separation model at that station. An impressive overall difference between the two figures is in their reversed skill score signs, with only a very few exceptions. It is clear that the direct model using linear learners is always worse than the best separation model, as expected from the discussion above. The only exception occurs at Ilorin, whose singular results were pointed out in

GRA16. (Note also that both models behave very similarly at Regina). In contrast, all results obtained using the non linear learners are systematically better than existing models, except at Florianopolis. Additionally, although the direct model is only slightly better than the indirect one in terms of RMSD (19.2% vs. 20.0%; see Tables 4 and 3, respectively) when using non linear learners, the indirect model is better than the best separation model at all stations (see Fig. 3), whereas, as just mentioned, the direct model is worse at Florianopolis (see Fig. 6). The unexpected behavior of the latter station would require further investigation considering local onsite information.

Similarly to Section 5, a study of the relevance of the original input variables (x_1 to x_{24}) for the direct model using non linear learners is carried out here. In the present case, 24 different direct model versions are trained by adding variables successively from 1 to 24, using the ranking provided by XGBoost. The training and test RMSDs are shown in Fig. 7. It is observed that RMSD does not decrease appreciably beyond about 14 input variables. Indeed, with 14 variables, the training and test RMSDs are 14.3% and 19.2%, respectively, compared to 14.4% and 19.2% for the complete pool of 24 variables. The 14 “significant” variables are: K_T , V_6 , K_{Tm} , V_7 , V_6 , RH , V_1 , T , T_L , E_{bmc} , K_{Tp} , P , V_9 , and E . See Section 4.1 and GRA16 for further details. As already underlined in Section 5, this ranking is purely relative to the XGBoost model (see tion 3), and should not be interpreted as a measure of the importance of the variables taken individually. Moreover, as before, note that a much faster direct GB model could be constructed by using only the 10 “best” inputs just mentioned with, virtually, no loss of accuracy. Interestingly, 5 out of the 14 significant variables describe temporal variability. This confirms the general distinctive importance of this factor, but also suggests that a more advanced variability index, which would even have more skill and could thus advantageously replace all existing ones, remains to be defined.

7. Model performance

7.1. Validation at independent sites

The previous sections evaluated the GB models’ capabilities to make predictions at locations where historical data sets are available for model training. In most situations, however, it is necessary to use models at locations for which historical data sets are not

Fig. 5. Same as Fig. 2, but for the direct model using linear learners. A color version of this figure is available in the on-line version of this paper.

Fig. 6. Same as Fig. 5, but for the direct model using non-linear learners. A color version of this figure is available in the on-line version of this paper.

Fig. 7. Same as Fig. 4, but for the direct model using non-linear learners. A color version of this figure is available in the on-line version of this paper.

Table 5

Average and standard deviation (within brackets) of the station-wise RMSD and MBD scores over all stations, expressed in percent after normalization by the mean observed DNI. Results are shown for the indirect and direct models using non-linear learners.

GB method	RMSD	MBD
Indirect model	22.4% (12.6%)	1.7% (8.4%)
Direct model	21.7% (11.8%)	-1.3% (7.6%)

respectively; see Tables 3 and 4). The 2% RMSD increase that occurs when the training and validation data sets are not collocated represents the penalty resulting from the underlying spatial extrapolation process. Fig. 8 displays the RMSD results for the indirect model at each validation station. The RMSD consistently follows the same trend as the error reported in Section 5, with only two major exceptions (Darwin Met Office and Goodwin Creek). Additional results, this time for the direct model, are shown in Fig. 9. In this case, the trend is also similar, except for the stations at Sede Boquer and, again, Goodwin Creek. One more exception is Ilorin, where the trend is unexpectedly reversed, i.e., doing both training and testing at Ilorin results in higher RMSD than training using data from independent locations.

7.2. Overall comparisons of GB and conventional models

According to GRA16, the PEREZ2 (Perez et al., 2002) and ENGERER2 (Engerer, 2015) models are generally the best performing conventional separation models at most of the 54 locations studied. Hence, for the sake of comparison, Fig. 10 compares the density scatterplots of ground DNI observations against DNI modeled values obtained with the PEREZ2 and ENGERER2 (upper two rows) models, as well as with the indirect and direct XGBoost models with nonlinear learners (bottom two rows). The test dataset is stratified by climate region, one per column (see Table 1). Overall, the XGBoost models clearly outperform the two conventional separation models. The improvement is especially remarkable for the high albedo sites. With all models, however, there are situations for which the modeled DNI is zero whereas the DNI observations are often quite large. As discussed in GRA16, this results from cloud enhancement effects, which are significant at 1 min resolution. At the HA (high albedo) sites, the additional albedo enhancement

available. For such locations, the aim of this subsection is to study the performance of the two modeling approaches evaluated in this work. To this goal, a different evaluation process is devised, in which only one of the stations is used for validation, whereas data from the remaining 53 stations is used for training all GB models. This follows the leave one out cross validation concept (Arlot and Celisse, 2010), where each station is successively left out and used for testing. Thus, the validation location is independently tested by using no training data from that station. In order to compare the results thus obtained to the ones shown so far, only the last chronological third of data at the validation site is used for testing, as in previous sections. The training data is also subsampled here as described in Section 4.2. The training and validation sequence is repeated at each validation site, and the MBD and RMSD values thus obtained at each site are subsequently aggregated to provide their overall mean and standard deviation like in previous sections.

Table 5 displays the average RMSD and MBD at the validation locations with the indirect and direct models using non linear learners. As could be expected, the average error is slightly higher than the error when the training and validation data sets are collocated (20.0% and 19.2% RMSD for the indirect and direct models,

Fig. 8. RMSD score for each station using the indirect model with non-linear learners. For each station, the bars in the background show the model performance when the training and validation data sets are collocated. The foreground bars show the model performance when training and validation data sets are not collocated. The stations are grouped by climatic region. A color version of this figure is available in the on-line version of this paper.

Fig. 9. Same as Fig. 8, but for the direct modeling approach. A color version of this figure is available in the on-line version of this paper.

effect amplifies this issue even further. The indirect XGBoost model is affected (since all the underlying separation models it depends on are inaccurate in such cases), but still provides noticeable improvement under such difficult conditions, compared to PEREZ2 or ENGERER2. The direct XGBoost model's scatterplots are similar to those of its indirect counterpart, overall. There is some additional scatter, however, around DNI observations of $\approx 600 \text{ W/m}^2$ at temperate sites. Conversely, the scatter is reduced at high albedo sites. As in the case of the indirect GB model, the impact of cloud and albedo enhancements remains. This suggests that a more appropriate input variable than the 10 "most efficient" ones indicated in Section 6 would be needed to detect these effects in the GHI data, and take them into account more effectively.

At HA sites, it can be hypothesized that the scatter could be further reduced if the surface albedo had a more important role and became part of the list of significant inputs. It has apparently not

been perceived as a significant input by the GB selection process described in Section 6 because of the relatively small number of HA stations and observations compared to the whole dataset. Considering that each of the four climatic regions considered here has a specific signature in the scatterplots of Fig. 10, a direct GB model dedicated to each region separately would have the potential to reduce errors and scatter further.

8. Discussion and conclusions

The accurate separation of the direct and diffuse irradiance components has been an elusive research topic for the last six decades. Recent results in GRA16 have shown that none of the 140 separation models reviewed could be considered "universal", and that only limited progress could be noticed since Liu & Jordan's simple diffuse fraction correlation of 1960. This justifies the devel

Fig. 10. Density scatterplots of observations against DNI modeled values with the Perez2 model (top), the Engerer2 model (middle), and the indirect and direct XGBoost models with nonlinear learners (bottom). The validation is conducted over the test dataset and the results are stratified according to the four prevailing climatic conditions. A color version of this figure is available in the on-line version of this paper.

opment of alternate approaches, including the opportunities offered by Machine Learning (ML) techniques. The present study builds on 140 separation models at all 54 world stations examined in GRA16. These stations are scattered over the world and are representative of 4 main different climates.

In this contribution, the XGBoost gradient boosting (GB) machine learning algorithm has been applied to address two main goals: the creation of an indirect GB model through the combination of conventional separation models, and the construction of a direct GB model that is driven by a subset of the same inputs as the conventional separation models. In each case, both linear and non linear weak prediction models (also referred to as learners) have been used.

The combination of separation models (indirect approach) using both linear and non linear learners performs generally better than the best existing conventional separation model at each station. However, only the GB model version that uses non linear learners is able to outperform the best separation model at all stations. In the case of the direct modeling approach, the linear learners perform worse than the best conventional separation model at almost all stations. This is explained by the non linearity of the relation between the input variables and DNI. Conversely, the GB model version using non linear learners outperform the best

conventional separation model at almost all stations. Both the indirect and direct models perform similarly in terms of average RMSD, although the former is better at all stations than the best conventional separation model at each station.

Taking into account the conclusions above, the performance of the new GB models (indirect and direct) has been evaluated at each location, successively by excluding it from the training data set, thus resulting in spatial extrapolation and strict independent testing. In this approach, one of the stations is used for validation (i.e. the independent location) and data from all the 53 remaining stations is used for training. The results suggest that both the indirect and direct approaches can be used for making predictions at independent locations with only an increase of 2% in RMSD on average, with respect to models developed from collocated training and validation data sets.

As part of the learning process, XGBoost has been used to rank the predictors with regard to their relative importance in the final outcome. This feature proved useful to determine the most important conventional separation models in the indirect modeling approach, and the most important input variables in the direct approach. It is found that the original list of 140 conventional separation models used in the indirect modeling approach can be reduced down to only 26 such models without any performance

reduction. Most of the top models in this subset have at least 4 predictors, indicating that conventional separation models with fewer predictors have less generalization potential. Analogously, a similar performance of the direct model could be achieved using only 14 particular inputs out of the 24 possible variables included in the initial data set.

It is emphasized that the indirect and direct GB models developed here are dynamic and data driven. Thus, they offer a parallel pathway to exploit the current rapid increase in solar resource data volume. Unfortunately, this data driven ML approach means they can be hardly formulated as, e.g., a deterministic multi linear combination of existing models, which would be very convenient for most engineering applications. Despite this limitation, the present results can help develop better practical models in the future, at least by delineating the best functional forms (presumably those used in the 26 "best" models) and the inputs with the best prediction potential (presumably the 14 "essential" variables).

Overall, the present results suggest that advanced ML techniques, such as gradient boosting, can also significantly improve the performance of single conventional separation models by considering ensembles, and can considerably expand the original geographical or climatological area of application of any single separation model. Although the models developed here aim at universality, their development is still empirical. It can be anticipated that different combinations of separation models would have been obtained with different observational data sets. Further studies should therefore be conducted to compare the performance of the best models developed here to that of alternate models based on data from stations in other regions or under completely different climates.

Acknowledgements

The authors express their gratitude to the dedicated personnel who maintain the 54 radiometric stations considered here, particularly those from the Baseline Solar Radiation Network, whose high quality measurements are central to this study. The PSA DLR and Masdar data were kindly provided by Stefan Wilbert and Peter Armstrong, respectively, whose collaboration is highly appreciated. The first two authors have been funded by the Spanish Ministry of Science under project ENE2014 56126 C2 2 R (AOPRIN SOL project). Jose Antonio Ruiz Arias was supported by the Spanish Ministry of Economy and Competitiveness under the project ENE2014 56126 C2 1 R.

References

Alam, S., Kaushik, S.C., Garg, S.N., 2009. Assessment of diffuse solar energy under general sky condition using artificial neural network. *Appl. Energy* 86, 554–564.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. *Stat. Surv.* 4, 40–79.

Bertrand, C., Vanderveken, G., Journée, M., 2015. Evaluation of decomposition models of various complexity to estimate the direct solar irradiance over Belgium. *Renew. Energy* 74, 618–626.

Bilbao, J., Román, R., de Miguel, A., 2014. Measurements and model evaluations of direct normal irradiance in Central Spain. In: *Proc. Conf. EuroSun 2014, Aix-Les-Bains, France, International Solar Energy Soc.*

Bright, J.M., Smith, C.J., Taylor, P.G., Crook, R., 2015. Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data. *Sol. Energy* 115, 229–242.

Cebecauer, T., Sári, M., Gueymard, C.A., 2011. Uncertainty sources in satellite-derived direct normal irradiance: How can prediction accuracy be improved globally? In: *SolarPACES Conf., Granada, Spain (2011)*.

Chen, T., He, T., 2015. Xgboost: eXtreme Gradient Boosting. R package version 0.4-2.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*.

Cucumo, M., De Rosa, A., Ferraro, V., Kalliatatos, D., Marinelli, V., 2007. Experimental testing of models for the estimation of hourly solar radiation on vertical surfaces at Arcavacata di Rende. *Sol. Energy* 81, 692–695.

Dervishi, S., Mahdavi, A., 2012. Computing diffuse fraction of global horizontal solar radiation: a model comparison. *Sol. Energy* 86, 1796–1802.

Elminir, H.K., Azzam, Y.A., Younes, F.I., 2007. Prediction of hourly and daily diffuse fraction using neural network, as compared to linear regression models. *Energy* 32, 1513–1523.

Engerer, N.A., 2015. Minute resolution estimates of the diffuse fraction of global irradiance for southeastern Australia. *Sol. Energy* 116, 215–237.

Fernández-Peruchena, C.M., Bernados, A., 2015. A comparison of one-minute probability density distributions of global horizontal solar irradiance conditioned to the optical air mass and hourly averages in different climate zones. *Sol. Energy* 112, 425–436.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 1189–1232.

Friedman, J.H., 2002. Stochastic gradient boosting. *Comput. Stat. Data Anal.* 38 (4), 367–378.

Gansler, R.A., Klein, S.A., Beckman, W.A., 1995. Investigation of minute solar radiation data. *Sol. Energy* 55 (1), 21–27.

Gueymard, C.A., 2009. Direct and indirect uncertainties in the prediction of tilted irradiance for solar engineering applications. *Sol. Energy* 83, 432–444.

Gueymard, C.A., 2010. Progress in direct irradiance modeling and validation. In: *Proc. Solar 2010 Conf., Phoenix, AZ, American Solar Energy Soc.*

Gueymard, C.A., Ruiz-Arias, J.A., 2014. Performance of separation models to predict direct irradiance at high frequency: validation over arid areas. In: *Proc. Conf. EuroSun 2014, Aix-Les-Bains, France, International Solar Energy Soc.*

Gueymard, C.A., Ruiz-Arias, J.A., 2016. Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance. *Sol. Energy* 128, 1–30.

Huang, J., Perry, M., 2016. A semi-empirical approach using gradient boosting and k-nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting. *Int. J. Forecast.* 32, 1081–1086.

Ineichen, P., 2008. Comparison and validation of three global-to-beam irradiance models against ground measurements. *Sol. Energy* 82, 501–512.

Liu, B.Y.H., Jordan, R.C., 1960. The interrelationship and characteristic distribution of direct, diffuse and total solar radiation. *Sol. Energy* 4, 1–19.

Long, C.N., Shi, Y., 2008. An automated quality assessment and control algorithm for surface radiation measurements. *Open Atmos. Sci. J.* 2, 23–37.

Magalhães, C., Brito, M.C., Soares, P.M.M., 2014. Assessment of diffuse radiation models for cloudy atmospheric conditions in the Azores region. *Sol. Energy* 108, 538–547.

Ngoko, B.O., Sugihara, H., Funaki, T., 2014. Synthetic generation of high temporal resolution solar radiation data using Markov models. *Sol. Energy* 103, 160–170. <http://dx.doi.org/10.1016/j.solener.2014.02.026>.

Ohmura, A., Dutton, E.G., Forgan, B., et al., 1998. Baseline surface radiation network (BSRN/WCRP): new precision radiometry for climate research. *Bull. Am. Meteorol. Soc.* 79, 2115–2136.

Padovan, A., Del Col, D., Sabatelli, V., Marano, D., 2014. DNI estimation procedures for the assessment of solar radiation availability in concentrating systems. *Energy Proc.* 57, 1140–1149.

Perez, R., Ineichen, P., Moore, K., Kmiecik, M., Chain, C., George, F., Vignola, F., 2002. A new operational model for satellite-derived irradiances: description and validation. *Sol. Energy* 73 (5), 307–317. [http://dx.doi.org/10.1016/S0038-092X\(02\)00122-6](http://dx.doi.org/10.1016/S0038-092X(02)00122-6).

Perez, R., Cebecauer, T., Sári, M., 2013. Semi-empirical satellite models. In: Kleissl, J. (Ed.), *Solar Energy Forecasting and Resource Assessment*. Elsevier.

Roesch, A., Wild, M., Ohmura, A., Dutton, E.G., Long, C.N., Zhang, T., 2011. Assessment of BSRN radiation records for the computation of monthly means. *Atmos. Meas. Tech.* 4, 339–354. Corrigendum, www.atmos-meas-tech.net/4/973/2011/.

Steinberg, D., Colla, P., 2009. CART: classification and regression trees. *Top Ten Algorithms Data Min.* 9, 179.

Tapakis, R., Michaelides, S., Charalambides, A.G., 2016a. Computations of diffuse fraction of global irradiance: Part 1 – Analytical modeling. *Sol. Energy* 139, 711–722.

Tapakis, R., Michaelides, S., Charalambides, A.G., 2016b. Computations of diffuse fraction of global irradiance: Part 2 – Neural Networks. *Sol. Energy* 139, 723–732.

Urraca, R., Antonanzas, J., Antonanzas-Torres, F., Martínez-de-Pison, F.J., 2017. Estimation of daily global horizontal irradiation using extreme gradient boosting machines. *Proc. International Joint Conference SOCO'16-CISIS'16-ICEUTE'16, Advances in Intelligent Systems and Computing*, vol. 528. Springer, p. 11. <http://dx.doi.org/10.1007/978-3-319-47364-2>.

Yang, D., Dong, Z., Nobre, A., Khoo, Y.S., Jirutitijaroen, P., Walsh, W.M., 2013. Evaluation of transposition and decomposition models for converting global solar irradiance from tilted surface to horizontal in tropical regions. *Sol. Energy* 97, 369–387.

Yang, D., Ye, Z., Hong, L., Lim, I., Dong Zibo, 2015. Very short term irradiance forecasting using the lasso. *Sol. Energy* 114, 314–326. <http://dx.doi.org/10.1016/j.solener.2015.01.016>.