Random forest regression for improved mapping of solar power resources at high latitudes

Bilal Babar*, Luigi Tommaso Luppino, Tobias Boström and Stian Normann Anfinsen

Department of Physics and Technology, The Arctic University – University of Tromsø, Norway

*corresponding author: bilal.babar@uit.no

Abstract

Datasets from meteorological reanalyses and measurements from polar orbiting satellites are the available sources of large-scale information about solar radiation. However, both the reanalyses and the satellite-based estimates can be severely biased, especially in high latitude regions. In this study, solar radiation estimates from the ECMWF Reanalysis 5 (ERA5) and the Cloud, Albedo, Radiation dataset Edition 2 (CLARA-A2) were used as input to a random forest regression (RFR) model to construct a novel dataset with higher accuracy and precision than the input datasets. For monthly averages of global horizontal irradiance (GHI) at Norwegian sites, CLARA-A2 and ERA5 respectively produced a root mean squared deviation (RMSD) of 9.6 Wm^{-2} and 10.2 Wm^{-2}, a mean absolute deviation (MAD) of 6.3 Wm^{-2} and 7.0 Wm^{-2}, and a bias of -1.6 Wm^{-2} and 3.9 Wm^{-2}. In contrast, the proposed regression model provided an RMSD of 6.6 Wm^{-2}, an MAD of 4.3 Wm^{-2}, and a bias of -0.2 Wm^{-2}. This shows that the RFR model is both accurate and precise, and significantly reduces both dispersion and bias in the new dataset with respect to the constituent sources. The proposed model provided more accurate and precise estimates in a seasonal error analysis as well. A sky stratification analysis was performed to evaluate the accuracy of the datasets under different sky conditions. It was found that the proposed model provides better estimates under all sky conditions with particular improvements in overcast conditions. The proposed regression model was also tested on five Swedish locations and it was found to improve solar radiation estimates to a similar degree as for the Norwegian locations, thus proving its consistency under similar climatic conditions.

Keywords: Solar radiation; High latitudes; ERA5; CLARA; CMSAF; Random forest regression

1. Introduction

The bankability of solar power plants largely depends on the accuracy and precision of the solar radiation measurements or estimates, which are required at all stages of solar energy projects. Time series or temporal averages of solar radiation are obtained initially before a particular system can be simulated and its design criteria and performance are evaluated. In the case of flat plate collectors, such as photovoltaic (PV) and thermal, global horizontal irradiance (GHI) or global tilted irradiance (GTI) are used in the feasibility and planning phases. Additionally, long-term variability in solar radiation is used to quantify the solar resource and project worst-case scenarios of energy production in such systems. During operation, real-time data are typically required to verify the performance of the system and detect problems. In both cases, the required data can be obtained from measurement, modelling, or a combination of both (Sengupta et al., 2017; Urraca et al., 2017b).

High quality solar resource assessments make technology deployment possible by helping the decision makers to reduce the uncertainty in investment decisions. However, the assessments cannot rely exclusively on ground measurements of solar radiation, because these are usually not available at most locations in the world. Even though such measurements exist at some locations, they frequently contain missing or erroneous data that must be filled in by using modelled data or interpolation from nearby measurement stations. Lastly, the cost of maintaining local equipment is larger than operating a model, assuming that satellite data and the output of reanalyses are provided free of charge or at a reasonable cost. Although model data are not as accurate as ground measurements, they can be used as an alternative

(Stoffel et al., 2010). Nevertheless, quality ground measurements remain essential because they have low errors and can be used to validate models (Sengupta et al., 2017).

Geostationary satellites are widely used for estimating surface solar radiation at low and medium latitudes, where their measurements of top-of-atmosphere upwelling radiances and surface albedos are used to derive GHI at the surface (Cano et al., 1986; Pinker and Laszlo, 1992; Rigollier et al., 2004; Tarpley, 1979). These satellites are positioned over the equator at different longitudes in order to provide a global coverage between -60° and +60° in latitudes. For instance, the Meteosat first and second generation geostationary satellites provide coverage of most of continental Europe (Müller et al., 2015; Pfeifroth et al., 2017; Schmetz et al., 2002; Urraca et al., 2017b). However, estimates above 65°N are prohibited by the slant viewing angle that geostationary satellites experience when they point away from nadir *i.e.*, the vertical direction directly below the satellite (Schulz et al., 2009).

Above the critical latitudes that limit geostationary satellites, polar orbiting satellites can be used to estimate surface solar radiation (Karlsson et al., 2017). Polar orbiting satellites traverse the entire Earth and provide global coverage, but their accuracy decrease at high latitudes because of the large angles between the satellite sensor and the Sun. Another factor that decreases the accuracy at high latitude is the frequent snow cover, which the satellites sensors cannot differentiate from clouds in the visible spectrum. The temporal resolution of solar radiation estimated by polar orbiting satellites is lower than that of geostationary satellites, since the revisit time of the former is higher than the repeat time of image acquisitions used by geostationary satellites. Whereas the latter capture images at least every 15 minutes, the polar orbiting satellites sense a given location twice each day on the equator and about 14 times each day near the poles. The sensing frequency of polar orbiting satellites is best at high latitudes, since swath overlap increases towards the poles, where their orbits converge. The accuracy of solar radiation estimated from satellite data is lower than ground measurements, but the advantages include large spatial and temporal coverage (Noia et al., 1993). In another study it was observed that estimates from polar orbiting satellites provide reasonable accuracy, but estimates obtained over snow-covered surfaces result in high errors because it is difficult to differentiate clouds from snow in the visible spectrum of light (Babar et al., 2018a). For a list of known issues and uncertainty sources, refer to Suri and Cebecauer (2014).

In addition to satellite measurements, meteorological reanalyses also provide surface short-wave incoming radiation estimates (Wild, 2008; Wild et al., 2015). Reanalysis datasets are produced by data assimilation of historical observational data, aiming to obtain the initial state of selected parameters which best fits a numerical weather prediction (NWP) model to the available data (Kennedy et al., 2011). Reanalyses are not as accurate as satellite-based estimates, but they provide global coverage for multi-decadal time range (Babar et al., 2018b; Urraca et al., 2017b; Urraca et al., 2018).

Both the satellite-derived estimates and reanalyses have a certain degree of uncertainty, but proper identification and removal of errors can improve the results. Site adaptation refers to the improvements that can be obtained in satellite-derived or model-based solar irradiance by using short-term ground measurements to reduce the systematic bias in the original dataset. In Polo et al. (2016), the authors have provided a preliminary survey of available site adaptation techniques. Site adaptation can be physically based methods in which the atmospheric input data such as aerosol optical depth and vertically-integrated water column are adjusted to better match the ground based observations (Gueymard, 2012). Other such methods include the use of clear-sky models to adjust the atmospheric aerosol on clear sky days (Cebecauer and Šúri, 2012). The second type of site adaptation is based on statistical adjustment of meteorological observations, such as rain, wind and so forth. The linear statistical methods for bias removal is performed by first fitting a line to the observations and estimations. In the next step an x=y line is subtracted from all observations (Polo et al., 2015). This type of adjustment removes the systematic errors that exist due to the regional inconsistencies or from the radiative models. Moreover, non-parametric regression by using multiple input datasets has been performed by Davy et al. (2016) for Australia. In this study, the authors used generalized additive models with cubic smoothing splines to improve accuracy. By including an NWP model-derived irradiance as input, they reduced the root mean square deviation by a few percent. In the study presented here, an approach similar to the site adaptation technique by Davy et al. (2016) is used.

This study presents a novel dataset that is obtained by using mainly the solar radiation estimates from ECMWF Reanalysis 5 (ERA5) and Cloud, Albedo, Radiation dataset Edition 2 (CLARA-A2), hereafter referred as ERA5 and CLARA. It is observed that reanalyses usually overestimate surface solar radiation and satellite methods usually underestimate it (Babar et al., 2018a; Riihelä et al., 2015; Urraca et al., 2017b). The main motivation behind constructing a new estimate is that we want to overcome the underestimation tendency of satellite methods and the overestimation tendency of reanalyses by combining them into a dataset with lower bias and variance. The input datasets were used together with *in-situ* measurements to develop a novel random forest regression (RFR) model, which can be used to produce accurate and precise estimates of solar radiation at high latitudes.

This paper is formatted as follows: Section 2 describes the datasets, quality control procedures, RFR model and pre-processing used in this study. Section 3 describes the results of the study. Section 4 provides a conclusion of this work.

2. Datasets

CLARA and ERA5 are coarse resolution datasets and provide data on a grid of 0.25° x 0.25° and 0.28° x 0.28°, respectively. Data extraction from these datasets is performed by selecting the four grid points surrounding any location where we have ground measurements, and applying inverse distance weighted interpolation to obtain solar radiation at these coordinates. In case of CLARA, there are missing data points, which implies that at some of the time frames there is data lacking in the surrounding four grid points. When the surrounding points have less than three valid values, the interpolation is replaced by a missing data value, indicating that a valid value could not be extracted for that particular time. The ERA5 dataset does not contain missing values. It will be explained in section 2.6 how the proposed regression model handles missing data values.

2.1 CLARA-A2

This dataset was released in December 2016 and it is the second edition of CLARA (Cloud, Albedo, Radiation dataset) produced by Eumetsat's Satellite Application Facility on Climate Monitoring (CM-SAF) (Karlsson et al., 2017). The dataset covers 1 January 1982 to 31 December 2015, and constitutes an extension of 6 years relative to the previous CLARA-A1 dataset. This dataset has global coverage with a spatial resolution of 0.25° x 0.25° on a regular latitude-longitude grid and it provides daily and monthly averages of surface incoming shortwave (SIS) radiation. To calculate daily averages, at least 20 observations of incoming solar radiation in each grid box are required. Similarly, 20 valid daily averages are required to generate monthly averages (SAF, 2016). Along with SIS, CLARA also provides longwave up- and down-welling surface radiation.

The fundamental method used in calculating surface solar irradiance from satellite observations is that the reflectance measured by the satellite instruments is related to the atmospheric transmittance. The underlying algorithm in CLARA uses Advanced Very High Resolution Radiometer (AVHRR) sensor data to derive the cloud cover, which is used to calculate surface incoming solar radiation (Karlsson et al., 2017). In addition to the cloud cover information, the solar radiation is estimated by using auxiliary data like the solar zenith angle, vertically-integrated water vapour and aerosol optical depth. Finding solar zenith angles is straightforward and can be calculated accurately. In this dataset, all data points with solar zenith angles larger than 80° are set to missing values and solar zenith angles larger than 90° are set to zero. The vertically-integrated water vapour and aerosol optical depth are not available in the AVHRR data and for these external sources are used. For vertically-integrated water vapour, the ERA-Interim Reanalysis (Dee et al., 2011) is used and the vertical ozone column is set to a constant value of 335 DU, as its variability has negligible impact on the estimated solar radiation. Aerosol information for the algorithm is taken from the modified version of the monthly mean aerosol fields from the Global Aerosol Data Set/Optical Properties of Aerosols and Cloud (GADS/OPAC) climatology. In the algorithm, AVHRR data is used to retrieve only the cloud cover information. The first step in estimating surface solar radiation is the classification of the sky condition. Software from Eumetsat's Nowcasting Satellite Application Facility (SAFNWC) is used to derive the information on cloud coverage for each pixel by using the information from the satellite sensor (SAF, 2016). If no cloud is detected (cloud free pixel), surface solar radiation is calculated by using the clear-sky Mesoscale Atmospheric Global Irradiance Code (MAGIC) (Mueller et al., 2009) by using only auxiliary sources. If the pixel is classified as cloudy (cloud contaminated or fully cloudy), visible channels of AVHRR instrument are used to derive broadband reflectance. The reflectance for each pixel is then transferred to broadband fluxes by using a bidirectional reflectance distribution function (BRDF). In the next step, the broadband top-of-the-atmosphere albedo is used to derive transmissivity through a look-up table approach. Finally, the transmissivity is used to calculate the surface solar radiation. However, as a temporally constant surface albedo is used by the algorithm, it does not provide radiation estimates on snow and sea ice coverage areas (Karlsson et al., 2017). For more information on the CLARA dataset and its accuracy, refer to Karlsson et al. (2017).

2.2 ERA5

ECMWF Reanalysis 5 (ERA5) is the fifth generation atmospheric reanalysis of the global climate from the European Centre for Medium-Range Weather Forecasts (ECMWF). It spans a period from 1950 to near present time (Hersbach and Dee, 2016). At the time of this study, data from 2000 to 2017 is available. Further data back in time will be released in 2019-20, and the dataset will continue to update forward in near real-time. In ERA5, the solar radiation variable has a spatial resolution of 31km (0.28125° x 0.28125°) and an hourly temporal frequency. ERA5 uses Integrated Forecasting System (IFS) cycle 41r2 with a state-of-the-art four-dimensional variational analysis (4DVAR) assimilation system. ERA5 has a higher number of pressure levels than ERA-Interim (the previous edition of ECMWF reanalysis) and provides more parameters, including hourly estimates of atmospheric, land and oceanic climate variables. For more information on ERA5 refer to ECMWF (2018).

In this study, shortwave surface downward radiation and shortwave surface downward radiation clear-sky are used from this dataset. In ERA5, the incoming shortwave irradiance is obtained from a Radiative Transfer Model (RTM). This model simulates the attenuation in solar radiation caused by the atmosphere. Therefore, the quality of the radiation estimates depends on the RTM used. Reanalyses generally do not assimilate aerosol, clouds or water vapour data, which increases the uncertainty in the estimated surface irradiance (You et al., 2013; Zhao et al., 2013).

2.3 Ground data

The ground-measured data used in this study for regression and validation is obtained from the Norwegian Institute of Bioeconomy Research (NIBIO) for Norwegian locations and from the Swedish Meteorological and Hydrological Institute (SMHI) for Swedish locations. NIBIO and SMHI collect, maintain, and provide data from their respective networks of meteorological measurement stations in Norway and Sweden, including ground-measured solar radiation. NIBIO and SMHI register hourly-average GHI by using Kipp and Zonen CMP11 or CMP13 pyranometers. The data is quality controlled and the equipment is maintained regularly on a daily or weekly basis (NIBIO, 2018; Persson, 2000). The coordinates of the locations, their altitudes and land type are indicated in Appendix A, Tables A1-A2 and an overview of the site locations is shown in Figure 1. The Swedish locations were only used in the testing of regression model, so as to prove its robustness.

For the analysis, the Norwegian sites were divided into inland and coastal regions by observing the proximity to the shoreline. Regions within 30 km of the shoreline were considered as coastal. From the 31 Norwegian locations studied here, 14 sites were classified as coastal and 17 sites as inland. The locations were also divided into two other groups, where locations lying above 65°N were grouped together and locations lying below 65°N were put in another group. In this latitude-based grouping, four sites were in the above 65°N group and 27 sites belonged to below 65°N group. For details on this classification, refer to appendix A, Table A1.

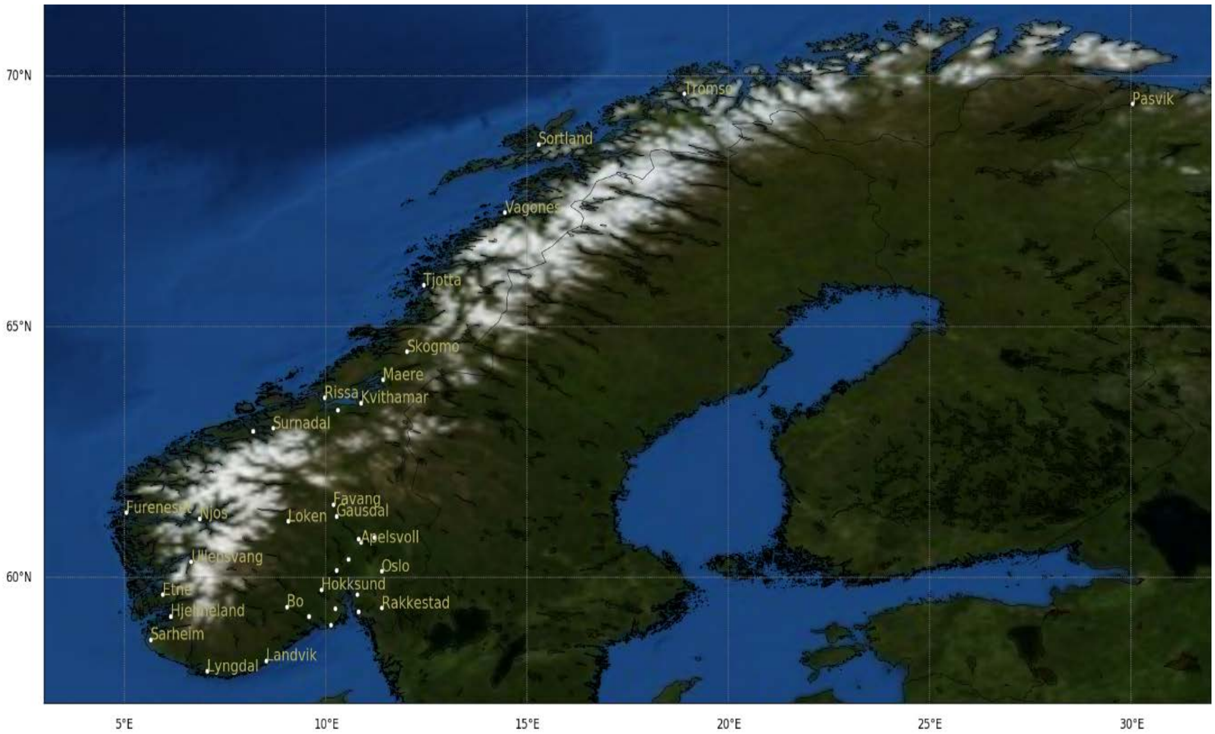


Figure 1: Locations of the Norwegian sites included in the study. To avoid overlapping of names some locations are shown with only white dots.

2.4 Quality Control

Although the data provided by NIBIO are quality controlled, Urraca et al. (2017a) observed that operational and equipment errors exist in NIBIO stations. The first check performed in this study is to look at the percentage of missing data. Any year having more than 5% of missing values was discarded from the analysis. The second check was performed by using the BSRN Global Network recommended quality control (QC) tests, version 2.0 (Long and Dutton, 2010). The BSRN QC test highlights values that are extremely rare and physically impossible. Based on this test, years having more than 1% of flagged values were removed from the ground data. The third quality control procedure was applied by using the QC technique of Urraca et al. (2017a). In this test, CLARA and ERA5 datasets are used to check the quality of ground measurements by constructing confidence intervals to detect the operational and equipment errors. Following Urraca et al. (2017a), the locations in Norway were divided into two sections by grouping locations above 65°N and locations below 65°N. Separate confidence intervals were constructed for both groups. After constructing these confidence intervals, the ground data was passed through an algorithm to check the data with errors, which appear in the form of flags. Following Urraca et al. (2017a) two checks were performed, one to see the operational errors and the other to see the equipment errors. After these checks, the years having large number of flags were visually inspected and removed from the analysis. For example, Pasvik, Mære, Njøs and Ullensvang were found to have a large number of flags from the third QC test, hence these locations were discarded. For more information on this quality control procedure, refer to Urraca et al. (2017a). A number of Norwegian locations were found to have large percentage of missing data points in years 2006 and 2007, hence these years were rejected from all Norwegian locations. See Appendix B, Table B1 for details of the years not included in this study.

2.5 Random forest regression

The motivation for adopting a regression model from the recent machine learning literature came from the hypothesis that our regression analysis might benefit from using an algorithm, which applies different regression functions for different subsets of the predictor data space. Conventional regression methods apply the same regression function, parametric or nonparametric, to the whole dataset and include all independent variables (predictors) as arguments to this function. More advanced methods can, on the other hand, allow more flexibility by judiciously selecting subsets of predictors or tailoring the regression function for subsets of the data in a manner that improves the overall performance in the regression analysis.

An example of such an approach is stratified regression analysis (Anderson et al., 1980; Tso and Yau, 2007), where separate regression models are set up for stratified samples of the independent variables, that are observed or hypothesized to exhibit different relations to the dependent variable. The strata can often be identified directly from the independent variables as natural groupings of the data. This idea is further developed in so-called clusterwise regression or regression clustering (Bagirov et al., 2017; Hsu, 2015; Späth, 1979), where clusters in the independent data are identified during the adaption of the regression model. Both the cluster-specific regression functions and the optimal clustering of the independent variable space are learnt iteratively from a training dataset containing paired independent and dependent samples. Input data points (vectors of predictor data) may be assigned to a unique cluster, or they may be given a fuzzy membership in multiple clusters. These membership values may then be used as weights in an ensemble approach where the dependent variable is predicted as a weighted average of the clusterwise regression functions. Another approach is the use of regression trees (Tso and Yau, 2007; Yu et al., 2010), where the predictor data space is recursively partitioned into finer regions using a tree structure, hoping that stronger relationships between independent and dependent variables can be formulated in these fine regions or branches of the tree. This may capture relations that are difficult to perceive in an explanatory data analysis if structures in the data are not visually apparent.

RFR is a regression tree method that has become very popular in recent years due to its strong performance, ease of implementation and low computational cost. It is an ensemble learning technique developed by Leo Breiman (Breiman, 2001), which is based on the construction of a multitude of decision trees. Branches of the trees represent particular paths that the input data can traverse, determined by threshold tests at the bisections. Leaves represent the output values stored at the end points of branching. In RFR, a particular tree is grown in accordance with the realization of a random vector in order to introduce variation. The final prediction is based on aggregation over the ensemble of trees, referred to as the forest (Segal, 2004). On each of the trees, branches or nodes are made which are based on comparing a randomly selected feature to a random threshold. The randomness introduced in both variable selection and threshold determination has been shown to result in attractive properties such as a controlled variance, resistance to overtraining, and robustness to outliers as well as irrelevant variables. Moreover, RFR inherently provides estimates of generalization error and measures of variable importance (Bylander, 2002; Siroky, 2009). The process of dividing the input training data over branches are repeated until one or a pre-set number of data points are contained in each branch. This final node of the tree is referred to as a leaf, and it represents the outcome of that particular regression in the whole model. The structure of the forest and hence the RFR behaviour can be controlled by three parameters: the number of trees, the number of variables considered in each node (set to $m=P/3$, where P is the total number of predictor variables), and the number of data points that can reside in a leaf (our default value is 10). Having a very high number of leaves in the model can cause overfitting, which can be overcome by pruning, i.e. limiting the number of data points in each leaf. Increasing the number of trees in the forest has two main effects: The computation load will increase. An initial increase in the accuracy of the regression will also be observed, before reaching a saturation point (Luppino et al., 2018), after which improvements are limited by a strong correlation between the trees (Breiman, 2001).

2.6 Pre-processing and input data for the model

The regression algorithm presented in Section 2.5 requires a training dataset for training the model and a test dataset to validate the trained model. In this study, the main inputs to the model are the surface solar radiations from CLARA and ERA5. In addition to these, clear sky indices were obtained by using shortwave surface radiation downward clear-sky (SWSDC) from ERA5 and GHI from ground measurements. By using clear sky indices, the RFR algorithm can take advantage of the sky stratification in different conditions. The daytime averages of solar zenith angle were also used as an input as it can provide the regression algorithm with the variation in solar elevation and its effects on surface radiations. Furthermore, latitudes and altitudes of the locations were used as input to the algorithm. In the training phase, 20% of randomly selected data was used from Norwegian locations, while the rest of the 80% data and data from Swedish locations were used in testing phase for validation of the model. The size of the training data was selected after running multiple runs with different sizes of data. Using more than 20% of data did not result in significant improvements. The model was tested with a number of trees ranging from 32 to 256 and pruning from 1 to 10 data points per leaf node. After multiple runs, 128 trees were selected with 10 data points per leaf node. The results presented in the next section are for the whole dataset.

Two main pre-processing procedures were applied in the training data of the regression model. Because of problems with convergence of the regression model, the missing data in CLARA and ground measurements was treated. First, training data with missing values in the ground measurements were discarded. This step eliminates the missing values in the ground data so that the regression model can converge, and also reduces the number of missing values in CLARA. This process was not performed on the test dataset, as missing values in the ground-measured data used in validation would not affect the errors statistics. Following previous studies that have shown that reanalyses can be used to fill the gaps in satellite datasets, we replaced in the second step the missing values of CLARA by corresponding values from ERA5 (Babar et al., 2018b; Urraca et al., 2017b; Urraca et al., 2018). These pre-processing steps enable the regression model to converge although with less training data.

2.7 Validation

In order to evaluate the performance of the RFR model, we introduce some common statistical measures. We first introduce the deviation (sometimes called error or residual) as the difference between the estimated (or predicted) and the observed global horizontal irradiance: $\delta = GHI_{estimated,i} - GHI_{observed,i}$, where the subscript i is a data point index.

A widely used measure of dispersion is the root mean square deviation (RMSD), computed from a sample of N data points as

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(GHI_{estimated,i} - GHI_{observed,i})^2}\,. \tag{1}$$

This measure combines both accuracy and precision.

The bias (or mean deviation) is used in the evaluation to quantify under- or overestimation. The bias is a measure of accuracy and is computed from the sample as

$$Bias = \frac{1}{N}\sum_{i=1}^{N}\left(GHI_{estimated,i} - GHI_{observed,i}\right) = \overline{GHI_{est\imath mated}} - \overline{GHI_{observed}}\ , \tag{2}$$

where $\overline{GHI_{estimated}}$ and $\overline{GHI_{observed}}$ are the sample means of the estimated and the observed GHI values, respectively.

The mean absolute deviation (MAD) is another measure of dispersion, which give less weight to and is therefore less sensitive to outliers than the RMSD (and the variance). The sample MAD is computed as (Sanchez-Lorenzo et al., 2013; Willmott and Matsuura, 2005)

$$MAD = \frac{1}{N}\sum_{i=1}^{N}\left|GHI_{estimated,i} - GHI_{observed,i}\right|. \tag{3}$$

Following Karlsson et al. (2017), the standard deviation of δ (STD) is also used in the evaluation. The sample STD is computed as

$$STD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\left(GHI_{estimated,i} - GHI_{observed,i}\right) - \left(\overline{GHI_{estimated}} - \overline{GHI_{observed}}\right)\right)^2}. \tag{4}$$

In addition, a bias-variance decomposition was used to obtain the optimal configuration of the RFR, with respect to the number of trees and the number of leaves. Moreover, R^2 and scatter plots are used to indicate the spread and overall correlation of the datasets with ground measurements.

3. Results

Table 1 compares performance of the models in terms of RMSD, MAD and bias for CLARA, ERA5 and the proposed RFR model. The RFR model performs better than the models that were used to construct it.

We start by looking at accuracy. For monthly averages of GHI at Norwegian locations, CLARA and ERA5 produced a bias of -1.6 Wm^{-2} and 3.9 Wm^{-2}, respectively. The RFR model delivered a bias of $-0.2\ Wm^{-2}$. The underestimation of the satellite model and the overestimation of the reanalysis is in agreement with previous studies (Babar et al., 2018a; Babar et al., 2018b; Urraca et al., 2017b; Urraca et al., 2018). The regression model underestimates the GHI, but the magnitude of the bias is reduced with 88% with respect to CLARA and with 95% with respect to ERA5, proving that the RFR model substantially improves the accuracy. These percentages are, as we will see, somewhat exaggerated when compared to seasonal values of the bias. Nonetheless, the seasonal biases are also much improved. The underestimation of the RFR model indicates that it weights CLARA higher than ERA5 on the whole, although the algorithm clearly adapts to exploit the strengths of either source under different conditions, as we will discuss below.

Regarding the dispersion measures, CLARA and ERA5 gave an MAD of 6.3 Wm^{-2} and 7.0 Wm^{-2}, respectively. The RFR model produced an MAD of 4.3 Wm^{-2}, which is a relative improvement of 32% and 39% with respect to CLARA and ERA5. Similarly, an RMSD of 6.6 Wm^{-2} was observed for the RFR model, while the RMSD of CLARA and ERA were 9.6 Wm^{-2} and 10.2 Wm^{-2}, respectively. The relative improvement in the RMSD was 31% and 35%, respectively. From the bias-variance decomposition of mean squared error ($MSE = RMSD^2$), the variance can be computed as: $Var = RMSD^2 - Bias^2$. We can use this to use that the variances of CLARA and ERA5 are very similar, and the variance of the RFR model is less half of these. This proves that the RFR model also provides a large improvement in precision. Table 1 also lists bias, MAD and RMSD for daily averages of GHI that show similar patterns as for the monthly averages.

Table 1 lists the error metrics after geographically grouping the ground measurement sites as explained in section 2.3. A brief overview of Table 1 shows that the proposed regression model improved all the four groups (above 65°N, below 65°N, coastal and inland). Like CLARA and ERA5, the proposed RFR model performed better at above 65°N than below 65°N. Nevertheless, the accuracy and precision is improved in both of these groups.

Table 1: The RMSD, MAD and bias of the input datasets and the presented model are shown. The error metrics for all locations in addition to providing an analysis on below 65°N, above 65°N, coastal and inland locations are shown. Numbers without parentheses are monthly averaged errors while those in parentheses are daily averaged errors. Best results are indicated in bold.

	RMSD (Wm^{-2})			MAD (Wm^{-2})			Bias (Wm^{-2})		
	CLARA	ERA5	RFR	CLARA	ERA5	RFR	CLARA	ERA5	RFR
NIBIO sites	9.6 (19.1)	10.2 (26.7)	**6.6** **(15.7)**	6.3 (13.1)	7.0 (16.7)	**4.3** **(10.2)**	-1.6 (-2.0)	3.9 (3.9)	**-0.2** **(-0.2)**
Above 65^{o}N	9.6 (16.0)	10.1 (26.3)	**6.5** **(13.7)**	6.3 (9.7)	6.9 (14.5)	**4.2** **(8.2)**	-1.6 (-2.9)	3.8 (5.6)	**-0.2** **(-0.1)**
Below 65^{o}N	9.7 (19.5)	12.7 (26.8)	**8.0** **(15.9)**	6.5 (13.6)	9.4 (17.3)	**5.4** **(10.5)**	-1.8 (-1.8)	5.7 (3.9)	**0.1** **(-0.1)**
Coastal	9.7 (16.7)	10.1 (26.7)	**6.6** **(14.8)**	6.4 (11.4)	7.0 (16.3)	**4.3** **(9.4)**	-1.7 (-1.1)	3.8 (4.9)	**-0.2** **(0.4)**
Inland	8.2 (20.8)	11.2 (26.7)	**6.6** **(16.4)**	5.7 (14.4)	7.9 (17.5)	**4.6** **(10.8)**	-0.6 (-2.6)	4.5 (3.4)	**0.1** **(-0.4)**

In addition, a seasonal error analysis was performed after dividing the yearly time series in groups of three months, i.e. February to April in FMA, May to July in MJJ, August to October in ASO, and November to January in NDJ. This type of grouping was preferred in this analysis because most locations analysed in this study are high latitude locations and at such locations the spread of solar radiation density is not as uniform as at other regions closer to the equator. At high latitude locations, most of the sun hours occur in summer months and least sun hours occur in winter months. By having such a grouping, summer and winter seasons are analysed separately. The seasonal analyses in Table 2 shows that errors decreased in all of the seasonal groups with the RFR model. However, the largest improvements were seen in FMA and MJJ. An analysis of the results of CLARA and ERA5 in NDJ and FMA shows that ERA5 performed better than CLARA in this period. This is mainly because of the low solar elevation in winter months, which increases errors in satellite-based estimates. However, CLARA performed better than ERA5 in MJJ and ASO.

The RFR model improves the accuracy and precision through all seasons. Nonetheless, the seasonal analysis reveals some interesting features: The bias of the RFR model varies over the year. The model underestimates in winter and overestimates in summer. However, we see that the biases of CLARA and ERA5 also fluctuate, and the RFR model succeeds in maintaining a much lower bias throughout the year. We may take this as a sign that the RFR model is flexible and adaptive, and manages to weight the input datasets in an appropriate way and combine their strengths to obtain good performance under various conditions. When it comes to the dispersion measures, the values of the RFR model follow the pattern of CLARA and ERA and largely decrease over the year. The largest relative improvements are seen in the FMA quarter, when the RFR model produces a 25% improvement in RMSD and a 39% improvement in MDA with respect to ERA5 (the best alternative). The magnitude of the bias reduction also over 70% for both models. The seasonal improvements are lower than the improvement in monthly averaged values, but the RFR model has much more consistent performance over the year than the input datasets. This is evident if one studies and compares the ranges or totals of the seasonal error metrics for the three models.

Table 2: The seasonal error analysis of CLARA, ERA5 and the RFR model are shown here. Major improvements occur in the FMA and MJJ quarters. Numbers without parentheses are monthly averaged errors while those in parentheses are daily averaged errors. Best results are indicated in bold.

	RMSD (Wm^{-2})				MAD (Wm^{-2})				Bias (Wm^{-2})			
	FMA	MJJ	ASO	NDJ	FMA	MJJ	ASO	NDJ	FMA	MJJ	ASO	NDJ
CLARA	15.3 (21.4)	8.8 (21.9)	8.9 (15.8)	6.9 (11.0)	10.4 (14.7)	6.7 (16.5)	4.9 (11.0)	4.4 (5.3)	-6.9 (-8.3)	1.3 (1.2)	**1.4** **(1.1)**	0.3 (-2.3)
ERA	12.9 (23.5)	14.3 (40.7)	9.8 (23.7)	6.2 (9.3)	9.2 (16.4)	11.4 (30.9)	6.7 (16.5)	2.7 (4.2)	7.0 (7.0)	7.2 (7.1)	2.0 (2.1)	0.2 (0.3)
RFR Model	**9.7** **(15.9)**	**7.4** **(21.2)**	**7.8** **(14.4)**	**5.6** **(8.8)**	**5.6** **(11.1)**	**5.6** **(15.9)**	**4.4** **(10.0)**	**2.3** **(3.9)**	**-1.8** **(-1.7)**	**-0.1** **(-0.2)**	1.5 (1.5)	**0.0** **(0.0)**

Finally, the R^2 values and the standard deviation (STD) of the Norwegian locations is analysed. Values of the coefficient of determination, R^2, are computed from the ground-measured and model data. The standard deviation is a measure of the spread of the prediction errors around their mean value. Table X shows the R^2 values and standard deviation for all Norwegian locations, in addition to below 65°N, above 65°N, coastal and inland regions. The standard deviation in Table 3 has units of Wm^{-2}, whereas R^2 has no units. For standard deviation, the smaller the value, the better the model estimates and for R^2, the larger the value, the better are the estimates.

Table 3: The R^2 and error standard deviation analysis of CLARA, ERA5 and the proposed RFR model for Norwegian locations is shown here. The RFR model improves the estimates in all types of geographical categories. The units of the standard deviation (STD) is Wm^{-2} and R^2 is unit-less. Best results are indicated in bold.

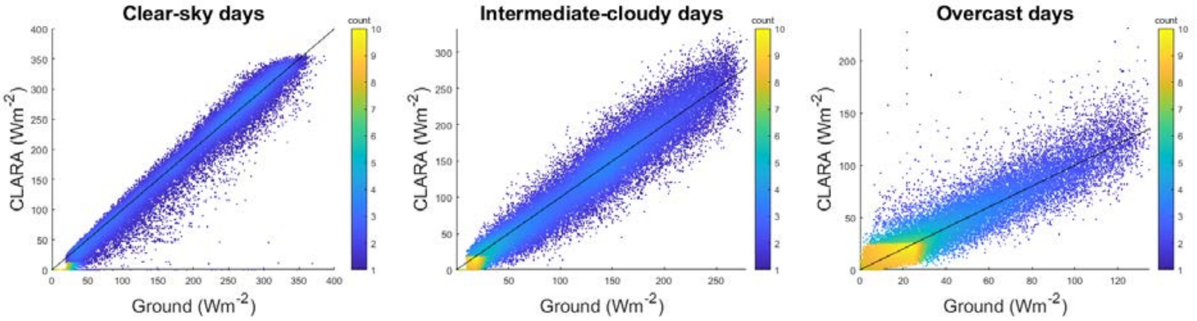
	NIBIO sites		Above 65ºN		Below 65ºN		Coastal		Inland	
	R^2	STD	R^2	STD	R^2	STD	R^2	STD	R^2	STD
CLARA	0.96	23.8	0.96	18.4	0.95	25.0	0.97	21.1	0.95	25.9
ERA	0.92	26.9	0.89	28.5	0.92	26.7	0.91	27.1	0.92	26.7
RFR model	**0.97**	**16.0**	**0.97**	**15.3**	**0.97**	**16.1**	**0.97**	**15.3**	**0.97**	**16.5**

It can be observed that the proposed regression model improves the solar radiation estimates at all Norwegian locations. The largest improvements were observed in location above 65°N, although the differences are small. The proposed model had lower standard deviation than CLARA and ERA5 in all geographical groups. Note that CLARA performs better in coastal regions than in inland regions, while the opposite is true for ERA5.

3.1 Sky stratification in CLARA, ERA5 and the regression model

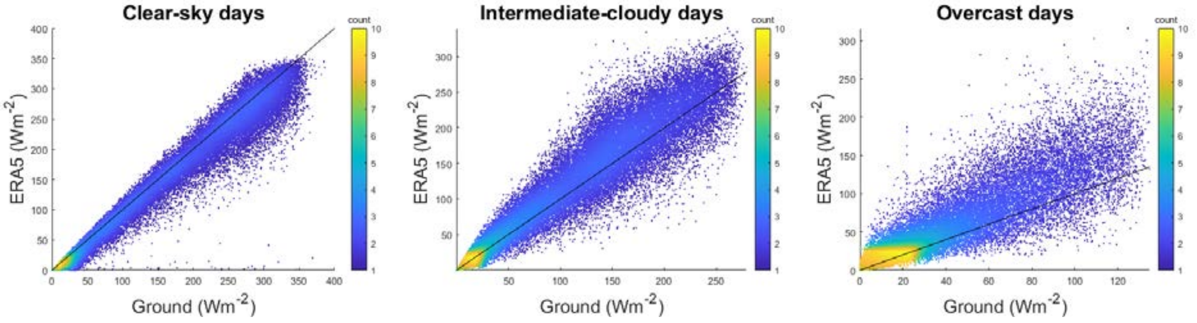
To evaluate the datasets for their performances in different sky conditions, the datasets were divided into clear-sky, intermediate-cloudiness and overcast categories. This division was established based on the clear-sky index (*Kc*), which is defined as the ratio of clear-sky GHI to the GHI recorded on the ground. Shortwave solar radiation clear-sky downwards (SWSCD) from ERA5 was used to obtain the clear-sky index. After calculating clear-sky index, *Kc*, following Smith et al. (2017) and Widén et al. (2017), values higher than 0.8 were considered as indicating a clear-sky day, values of *Kc* between 0.4 and 0.8 were considered as intermediate-cloudy, and values below 0.4 were considered as overcast. This kind of categorization is quite arbitrary, as days with *Kc* value of 0.8 or higher are not necessarily days with completely clear sky, but a majority of these days are expected to have a clear sky. This analysis is used here to roughly divide the sky conditions followed by a rigorous analysis. Any misclassification based on the clear sky indices will have similar effects on all the datasets.

Figures 2-4 show the errors in the datasets under different sky categories. It can be seen from the figures and the tables that the RFR model improves the results in the clear-sky and intermediate cloudy categories. However, in the overcast category, CLARA and the RFR model performed similarly besides that CLARA had a lower bias. On average, CLARA underestimated radiation in clear and cloudy conditions, while an overestimation was observed in overcast conditions. On the contrary, ERA5 overestimated radiation in cloudy and overcast conditions, while it was underestimated in clear-sky condition. ERA5 is reported to have a positive bias towards estimating days as clear sky and a negative bias towards estimating overcast days (Babar et al., 2018b). The reason for these biases is the higher concentration of total cloud water content in the ERA5 model on rather clear sky days and a lower concentration of total cloud water content in cloudy conditions. The underestimation in CLARA in clear sky and intermediate-cloudy days is possibly due to the use of an optically thick aerosol climatology – in this case the Global Aerosol Data Set/Optical Properties of Aerosols and Cloud (GADS/OPAC) climatology (Babar et al., 2018b; Mueller and Träger-Chatterjee, 2014). The RFR model underestimated solar radiation in clear sky condition and overestimated radiation in intermediate-cloudy and overcast conditions. Nevertheless, large improvements were observed in clear-sky and cloudy conditions. However, from a solar energy harvesting point of view, in overcast conditions smaller amounts of energy is produced as compared to clear-sky and intermediate-cloudy days.



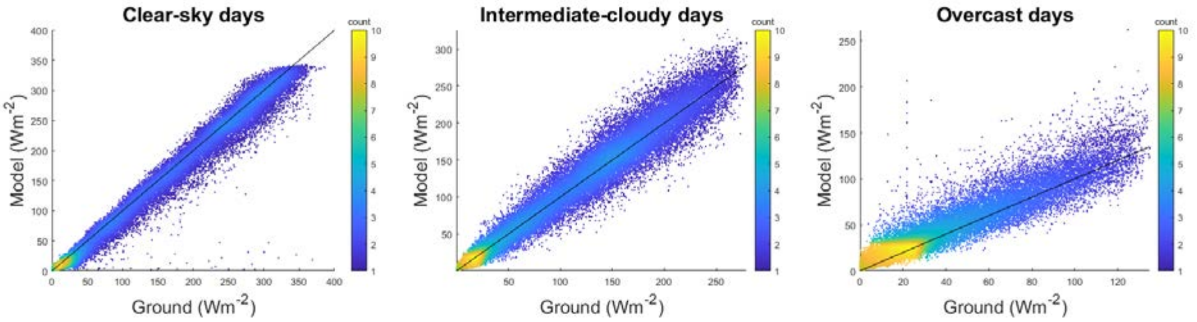
CLARA	RMSD (Wm^{-2})	MAD (Wm^{-2})	Bias (Wm^{-2})
Clear-sky	21.3	14.4	-7.1
Intermediate-cloudy	20.0	14.9	-2.8
Overcast	12.4	8.2	0.7

Figure 2: CLARA errors under clear-sky, intermediate-cloudy and overcast conditions for Norwegian sites. The scatter plots for different sky categories are also shown. The coloured legend bar shows the density of points.



ERA5	RMSD (Wm^{-2})	MAD (Wm^{-2})	Bias (Wm^{-2})
Clear-sky	25.0	15.9	-11.2
Intermediate-cloudy	28.2	19.4	9.5
Overcast	28.3	17.2	14.4

Figure 3: Same as Figure 2, but for ERA5.



RFR model	RMSD (Wm^{-2})	MAD (Wm^{-2})	Bias (Wm^{-2})
Clear-sky	17.4	11.3	-6.6
Intermediate-cloudy	16.8	11.8	1.7
Overcast	12.8	8.2	5.3

Figure 4: Same as Figure 2, but for the RFR model.

3.2 Testing the regression model on Swedish locations

In this section, the regression model is tested on five Swedish locations. Data from these locations were not used in the training of the model, therefore this analysis tests the robustness of the regression model proposed in this study. Table A2 in Appendix A lists the information on the Swedish locations used in the analysis.

Table 3 lists the errors for CLARA, ERA5 and the RFR model for individual Swedish locations. The errors for all locations are summarized in the last row of the table. In this analysis, it was found that the RFR model improved the solar radiation estimates for Swedish locations as well. The monthly MAD for all Swedish locations for CLARA and ERA5 was found to be 6.3 Wm^{-2} and 5.6 Wm^{-2}, respectively. At these locations, the RFR model gave a MAD of 4.5 Wm^{-2}. Similarly, the daily averages were also improved in the RFR model. As previously observed for Norwegian locations, CLARA underestimated the solar radiation and ERA5 overestimated it for Swedish locations. The proposed RFR model underestimated the solar radiation as well, but the magnitude of the bias was smaller than for CLARA and ERA5. This analysis shows that the proposed model can at least be used for Swedish locations that may have a similar climate in terms of cloud, snow and sunlight conditions.

Table 3: The RMSD, MAD and Bias of the input datasets and the RFR model for Swedish locations is shown here. These locations were not used in the training of the regression model. Numbers without parentheses are monthly averaged errors while those in parentheses are daily averaged errors. Best results are indicated in bold.

	RMSD (Wm^{-2})			MAD (Wm^{-2})			Bias (Wm^{-2})		
	CLARA	ERA5	RFR	CLARA	ERA5	RFR	CLARA	ERA5	RFR
Kiruna	17.2 (26.6)	**7.6** (24.0)	11.0 **(18.7)**	10.1 (16.6)	**4.9** (14.4)	6.8 **(11.7)**	-7.0 (-8.2)	**-2.3** **(-2.5)**	-5.9 (-6.0)
Luleå	10.6 (24.4)	10.4 (25.1)	**5.6** **(17.5)**	6.9 (14.9)	6.6 (15.3)	**3.8** **(11.0)**	-4.4 (-4.2)	5.1 (4.9)	**-2.1** **(-2.1)**
Umeå	8.3 (16.4)	7.1 (23.0)	**5.5** **(13.5)**	6.1 (11.5)	4.4 (14.2)	**3.8** **(9.1)**	-3.2 (-3.5)	**2.0** **(2.1)**	-2.6 (-2.5)
Stockholm	6.8 (16.4)	7.0 (23.6)	**5.9** **(14.6)**	5.1 (11.5)	4.8 (15.7)	**4.5** **(10.0)**	**2.6** **(2.5)**	3.1 (3.1)	3.9 (4.0)
Göteborg	**4.7** (14.9)	9.5 (26.1)	4.8 **(14.4)**	3.5 (10.5)	7.3 (17.0)	3.7 (9.9)	**1.6** **(1.8)**	6.9 (6.8)	3.0 (2.9)
SMHI locations	10.4 (20.3)	8.4 (24.4)	**6.9** **(15.9)**	6.3 (13.0)	5.6 (15.3)	**4.5** **(10.3)**	-2.1 (-2.3)	2.9 (2.9)	**-0.8** **(-0.7)**

4 Conclusion

Studies have shown that satellite estimation of solar radiation provide reasonable estimates and reanalyses can be used to fill the gaps when satellite datasets are not available or they contain missing data. It has also been observed that at high latitude locations there are a larger number of missing values in satellite-derived data, as in CLARA. Some previous studies have reported that prediction errors increase with latitude, so the available datasets have a systematic bias that grows with latitude. This study proposes a novel method to construct an improved dataset by combining a surface solar radiation dataset based on satellite measurements (CLARA-A2) and a newly published global reanalysis dataset (ERA5). The assumption used in this study is that the underestimation in satellite models and the overestimation in reanalyses can be largely cancelled and overcome if they are fused in a regression model to improve the estimates of surface solar radiation. The proposed regression model is constructed by using the random forest regression method, which is a machine learning algorithm based on regression trees and ensemble learning.

It is seen that on monthly and daily averages of radiation, the regression model provided more accurate estimations than CLARA and ERA5. On monthly averages of surface solar radiation for Norwegian locations, CLARA provided an MAD of 6.3 Wm^{-2} while ERA5 provided an MAD of 7.0 Wm^{-2}. The regression model reduced the error to a MAD of 4.3 Wm^{-2}. Similarly, on daily averages, CLARA and ERA5 provided MADs of 13.1 Wm^{-2} and 16.7 Wm^{-2}, respectively, while the regression model gave a MAD of 10.2 Wm^{-2}. Similar improvements were seen in RMSE values, proving that the RFR model has significantly improved precision with respect to the input datasets. In addition the RFR model was seen to provide large reductions in both annual and seasonal bias, showing that the accuracy improves as much as the precision.

A discussion of the seasonal analysis concluded that the RFR model succeeds in combining the input datasets in an adaptive fashion, such that the strengths of both models are exploited to produce consistently high performance under all conditions and throughout the whole year. Moreover, from a geographical analysis of errors it was observed that large improvements were obtained in locations above 65°N and coastal regions. A seasonal error analysis is performed and it is observed that the regression model provided better estimates than CLARA and ERA5 in all seasons of the year with large improvements in the period of November to April. A sky stratification analysis was performed on Norwegian locations to assess the datasets in different sky conditions. It was observed that the regression model improved solar radiation estimates in all sky condition, especially in clear-sky and intermediate-cloudy conditions. Additionally, in terms of standard deviation, large improvements were found inland and below 65°N. The proposed model was also tested on Swedish locations, that were not included in the training set, and very similar improvements were observed.

Overall, the regression model provides an improved alternative to the available reanalyses and satellite based estimates of surface solar radiation. In addition to an improved dataset, this study also highlights the important role of machine learning algorithms in the production of sophisticated databases for high latitude locations.

Appendix A

Table A1

Lists of Norwegian locations with their coordinates, altitudes and land type.

	Station	Latitude	Longitude	Altitude	Land type
1	Holt	69.65	18.91	12	Coastal
2	Sortland	68.65	15.28	14	Coastal
3	Vågønes	67.28	14.45	26	Coastal
4	Tjøtta	65.83	12.43	10	Coastal
5	Skogmo	64.51	12.02	32	Inland
6	Rissa	63.59	9.97	23	Coastal
7	Kvithamar	63.49	10.88	28	Inland
8	Skjetlein	63.34	10.3	44	Coastal
9	Surnadal	62.98	8.69	5	Inland
10	Tingvoll	62.91	8.19	23	Coastal
11	Fåvang	61.46	10.19	184	Inland
12	Fureneset	61.29	5.04	12	Coastal
13	Gausdal	61.22	10.26	375	Inland
14	Løken	61.12	9.06	527	Inland
15	Ilseng	60.8	11.2	182	Inland
16	Kise	60.77	10.81	129	Inland
17	Apelsvoll	60.7	10.87	262	Inland
18	Hønefoss	60.14	10.27	126	Inland
19	Årnes	60.13	11.39	162	Inland
20	Etne	59.66	5.95	8	Inland
21	Ås	59.66	10.78	94	Inland
22	Bø	59.42	9.03	105	Inland
23	Rakkestad	59.39	11.39	102	Inland
24	Ramnes	59.38	10.24	39	Coastal
25	Tomb	59.32	10.81	12	Coastal
26	Gjerpen	59.23	9.58	41	Coastal
27	Hjelmeland	59.23	6.15	43	Inland
28	Tjølling	59.05	10.13	19	Coastal
29	Særheim	58.76	5.65	90	Coastal
30	Landvik	58.34	8.52	10	Coastal
31	Lyngdal	58.13	7.05	4	Inland

Table A2

Lists of Swedish locations with their coordinates, altitudes and land type.

	Station	Latitude	Longitude	Altitude	Land type
1	Kiruna	67.83	20.43	408	Inland
2	Luleå	65.55	22.13	17	Coastal
3	Umeå	63.82	20.25	10	Coastal
4	Stockholm	59.35	18.07	30	Coastal
5	Goteborg	57.70	12.00	5	Coastal

Appendix B

Table B1

The following years are not included in the study.

	Station	Years having more than 5% missing data	Years failing Long and Dutton test	Years having operational error (snow/frost/ shading/soiling)	Years having equipment error
1	Holt	2001,2002,2006,2007,2008,2010	2013		2000
2	Sortland	2000,2006,2007,2010,2013			
3	Vågønes	2006,2007		2002	
4	Tjøtta	2006,2007			2008, 2012
5	Skogmo	2006,2007,2008,2015		2011	2013, 2014
6	Rissa	2006,2007	2000		
7	Kvithamar	2006,2007,2013			
8	Skjetlein	2006,2007	2000		
9	Surnadal	2006,2007,2014			
10	Tingvoll	2006,2007,2012			
11	Fåvang	2006,2007			2001
12	Fureneset	2006,2007,2011,2012			
13	Gausdal	2006,2007,2009			2015
14	Løken	2006,2007			
15	Ilseng	2006,2007,2004	2000	2009	
16	Kise	2002,2006,2007,2015		2013	
17	Apelsvoll	2006,2007		2002,2003,2004	2009
18	Hønefoss	2006,2007	2000		
19	Årnes	2006,2007			
20	Etne	2006,2007		2004,2012	
21	Ås	2006,2007			
22	Bø	2000,2006,2007			
23	Rakkestad	2006,2007			
24	Ramnes	2006,2007		2009	
25	Tomb	2006,2007	2009		
26	Gjerpen	2006,2007,2015			
27	Hjelmeland	2006,2007			2002, 2015
28	Tjølling	2006,2007,2008,2014		2012,2015	2009, 2010
29	Særheim	2000,2006,2007			
30	Landvik	2006,2007		2005,2010,2014, 2015	
31	Lyngdal	2006,2007	2001		

References

Anderson, D.W., Kish, L., Cornell, R.G., 1980. On stratification, grouping and matching. J Scandinavian Journal of Statistics, 61-66.

Babar, B., Graversen, R., Boström, T., 2018a. Evaluating CM-SAF solar radiation CLARA-A1 and CLARA-A2 datasets in Scandinavia. Solar Energy 170, 76-85.

Babar, B., Rune, G., Tobias, B., 2018b. Solar radiation estimation at high latitudes: Assessment of the CMSAF databases, ASR and ERA5. under-review in Solar Energy.

Bagirov, A.M., Mahmood, A., Barton, A., 2017. Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach. J Atmospheric Research

188, 20-29.

Breiman, L.J.M.l., 2001. Random forests. Machine learning 45(1), 5-32.

Bylander, T.J.M.L., 2002. Estimating generalization error on two-class datasets using out-of-bag estimates. Machine Learning 48(1-3), 287-297.

Cano, D., Monget, J.-M., Albuisson, M., Guillard, H., Regas, N., Wald, L., 1986. A method for the determination of the global solar radiation from meteorological satellites data. Solar energy 37(1), 31-39.

Cebecauer, T., Šúri, M., 2012. Correction of satellite-derived DNI time series using locally-resolved aerosol Data, Proceedings of the SolarPACES Conference, Marrakech, Morocco.

Davy, R.J., Huang, J.R., Troccoli, A.J.S.E., 2016. Improving the accuracy of hourly satellite-derived solar irradiance by combining with dynamically downscaled estimates using generalised additive models. Solar Energy 135, 854-863.

Dee, D.P., Uppala, S.M., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d.P., 2011. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Quarterly Journal of the royal meteorological society 137(656), 553-597.

ECMWF, 2018. ERA5 data documentation.

Gueymard, C.A.J.S.E., 2012. Temporal variability in direct and global irradiance at various time scales as affected by aerosols. Solar Energy 86(12), 3544-3553.

Hersbach, H., Dee, D., 2016. ERA5 reanalysis is in production. ECMWF newsletter 147(7).

Hsu, D.J.A.e., 2015. Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. Applied energy 160, 153-163.

Karlsson, K.-G., Anttila, K., Trentmann, J., Stengel, M., Meirink, J.F., Devasthale, A., Hanschmann, T., Kothe, S., Jääskeläinen, E., Sedlar, J., 2017. CLARA-A2: the second edition of the CM SAF cloud and radiation data record from 34 years of global AVHRR data. Atmospheric Chemistry and Physics 17(9), 5809.

Kennedy, A.D., Dong, X., Xi, B., Xie, S., Zhang, Y., Chen, J., 2011. A comparison of MERRA and NARR reanalyses with the DOE ARM SGP data. Journal of Climate 24(17), 4541-4557.

Long, C.N., Dutton, E.G., 2010. BSRN Global Network recommended QC tests, V2. x.

Luppino, L.T., Bianchi, F.M., Moser, G., Anfinsen, S.N., 2018. Remote sensing image regression for heterogeneous change detection. arXiv preprint arXiv:1807.11766.

Mueller, R., Matsoukas, C., Gratzki, A., Behr, H., Hollmann, R.J.R.S.o.E., 2009. The CM-SAF operational scheme for the satellite based retrieval of solar surface irradiance—A LUT based eigenvector hybrid approach. Remote Sensing of Environment 113(5), 1012-1024.

Mueller, R., Träger-Chatterjee, C., 2014. Brief accuracy assessment of aerosol climatologies for the retrieval of solar surface radiation. Atmosphere 5(4), 959-972.

Müller, R., Pfeifroth, U., Träger-Chatterjee, C., Trentmann, J., Cremer, R., 2015. Digging the METEOSAT treasure—3 decades of solar surface radiation. Remote Sensing 7(6), 8067-8101.

NIBIO, 2018.

Noia, M., Ratto, C., Festa, R., 1993. Solar irradiance estimation from geostationary satellite data: II. Physical models. Solar Energy 51(6), 457-465.

Persson, T., 2000. Measurements of solar radiation in Sweden 1983-1998. na.

Pfeifroth, U., Kothe, S., Müller, R., Trentmann, J., Hollmann, R., Fuchs, P., Werscheck, M., 2017. Surface Radiation Data Set–Heliosat (SARAH)–Edition 2, Satellite Application Facility on Climate Monitoring.

Pinker, R., Laszlo, I., 1992. Modeling surface solar irradiance for satellite applications on a global scale. Journal of Applied Meteorology 31(2), 194-211.

Polo, J., Martín, L., Vindel, J.J.R.E., 2015. Correcting satellite derived DNI with systematic and seasonal deviations: application to India. Renewable Energy 80, 238-243.

Polo, J., Wilbert, S., Ruiz-Arias, J.A., Meyer, R., Gueymard, C., Suri, M., Martin, L., Mieslinger, T., Blanc, P., Grant, I., 2016. Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets. Solar Energy 132, 25-37.

Rigollier, C., Lefèvre, M., Wald, L., 2004. The method Heliosat-2 for deriving shortwave solar radiation from satellite images. Solar Energy 77(2), 159-169.

Riihelä, A., Carlund, T., Trentmann, J., Müller, R., Lindfors, A.V., 2015. Validation of CM SAF surface solar radiation datasets over Finland and Sweden. Remote Sensing 7(6), 6663-6682.

SAF, C., 2016. 2: Algorithm Theoretical Basis Document–CM SAF Cloud, Albedo, Radiation data record, AVHRR-based, Edition 2 (CLARA-A2)–Cloud Fraction. SAF/CM/DWD/ATBD/CMA_AVHRR version 2.0, https://doi. org/10.5676/EUM_SAF_CM ….

Sanchez-Lorenzo, A., Wild, M., Trentmann, J.J.R.S.o.E., 2013. Validation and stability assessment of the monthly mean CM SAF surface solar radiation dataset over Europe against a homogenized surface dataset (1983–2005). Remote Sensing of Environment 134, 355-366.

Schmetz, J., Pili, P., Tjemkes, S., Just, D., Kerkmann, J., Rota, S., Ratier, A., 2002. An introduction to Meteosat second generation (MSG). Bulletin of the American Meteorological Society 83(7), 977-992.

Schulz, J., Albert, P., Behr, H.-D., Caprion, D., Deneke, H., Dewitte, S., Durr, B., Fuchs, P., Gratzki, A., Hechler, P., 2009. Operational climate monitoring from space: the EUMETSAT Satellite Application Facility on Climate Monitoring (CM-SAF). Atmospheric Chemistry and Physics 9(5), 1687-1709.

Segal, M.R., 2004. Machine learning benchmarks and random forest regression.

Sengupta, M., Habte, A., Gueymard, C., Wilbert, S., Renne, D., 2017. Best practices handbook for the collection and use of solar resource data for solar energy applications. National Renewable Energy Lab.(NREL), Golden, CO (United States).

Siroky, D.S.J.S.S., 2009. Navigating random forests and related advances in algorithmic modeling. Statistics Surveys 3, 147-163.

Smith, C.J., Bright, J.M., Crook, R., 2017. Cloud cover effect of clear-sky index distributions and differences between human and automatic cloud observations. Solar Energy 144, 10-21.

Späth, H.J.C., 1979. Algorithm 39 Clusterwise linear regression. Computing 22(4), 367-373.

Stoffel, T., Renne, D., Myers, D., Wilcox, S., Sengupta, M., George, R., Turchi, C., 2010. Concentrating solar power. Golden: National Renewable Energy Laboratory.

Suri, M., Cebecauer, T., 2014. Satellite-based solar resource data: Model validation statistics versus user's uncertainty, ASES SOLAR 2014 Conference, San Francisco. pp. 7-9.

Tarpley, J., 1979. Estimating incident solar radiation at the surface from geostationary satellite data. Journal of Applied Meteorology 18(9), 1172-1181.

Tso, G.K., Yau, K.K.J.E., 2007. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy 32(9), 1761-1768.

Urraca, R., Gracia-Amillo, A.M., Huld, T., Martinez-de-Pison, F.J., Trentmann, J., Lindfors, A.V., Riihelä, A., Sanz-Garcia, A., 2017a. Quality control of global solar radiation data with satellite-based products. Solar Energy 158, 49-62.

Urraca, R., Gracia-Amillo, A.M., Koubli, E., Huld, T., Trentmann, J., Riihelä, A., Lindfors, A.V., Palmer, D., Gottschalg, R., Antonanzas-Torres, F., 2017b. Extensive validation of CM SAF surface radiation products over Europe. Remote sensing of environment 199, 171-186.

Urraca, R., Huld, T., Gracia-Amillo, A., Martinez-de-Pison, F.J., Kaspar, F., Sanz-Garcia, A., 2018. Evaluation of global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalyses using ground and satellite-based data. Solar Energy 164, 339-354.

Widén, J., Shepero, M., Munkhammar, J., 2017. On the properties of aggregate clear-sky index distributions and an improved model for spatially correlated instantaneous solar irradiance. Solar Energy 157, 566-580.

Wild, M., 2008. Short-wave and long-wave surface radiation budgets in GCMs: A review based on the IPCC-AR4/CMIP3 models. Tellus A: Dynamic Meteorology and Oceanography 60.5 60(5), 932-945.

Wild, M., Folini, D., Henschel, F., Fischer, N., Müller, B., 2015. Projections of long-term changes in solar radiation based on CMIP5 climate models and their influence on energy yields of photovoltaic systems. Solar Energy 116, 12-24.

Willmott, C.J., Matsuura, K.J.C.r., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research 30(1), 79-82.

You, Q., Sanchez-Lorenzo, A., Wild, M., Folini, D., Fraedrich, K., Ren, G., Kang, S., 2013. Decadal variation of surface solar radiation in the Tibetan Plateau from observations, reanalysis and model simulations. Climate dynamics 40(7-8), 2073-2086.

Yu, Z., Haghighat, F., Fung, B.C., Yoshino, H.J.E., Buildings, 2010. A decision tree method for building energy demand modeling. 42(10), 1637-1646.

Zhao, L., Lee, X., Liu, S., 2013. Correcting surface solar radiation of two data assimilation systems against FLUXNET observations in North America. Journal of Geophysical Research: Atmospheres 118(17), 9552-9564.