A Machine Learning Technique for Spatial Interpolation of Solar Radiation Observations

Thomas Leirvik¹ and Menghan Yuan¹

¹Graduate School of Business, Nord University, Bodø, Norway

Abstract This study applies statistical methods to interpolate missing values in a data set of radiative energy fluxes at the surface of Earth. We apply Random Forest (RF) and seven other conventional spatial interpolation models to a global Surface Solar Radiation (SSR) data set. We apply three categories of predictors: climatic, spatial, and time series variables. Although the first category is the most common in research, our study shows that it is actually the last two categories that are best suited to predict the response. In fact, the best spatial variable is almost 40 times more important than the best climatic variable in predicting SSR. Furthermore, the 10-fold cross validation shows that the RF has a Mean Absolute Error (MAE) of 10.2 Wm⁻² and a standard deviation of 1.5 Wm⁻². On the other hand, the average MAE of the conventional interpolation methods is 21.3 Wm⁻², which is more than twice as large as the RF method, in addition to an average standard deviation of 6.4 Wm⁻², which is more than four times larger than the RF standard deviation. This highlights the benefits of using machine learning in environmental research.

1. Introduction

Downwelling surface shortwave radiation, or surface solar radiation (hereafter referred to as SSR), is a fundamental determinant of the global energy balance, and a crucial driving force for temperature change and hydrological cycle variation (Budyko, 1969; Obryk et al., 2018). Furthermore, the impact of SSR on various aspects of a country's economy is widespread, including, but not limited to agricultural productivity, energy use, food security, and human health risk from increased poverty risk (Aaheim et al., 2015; Calcabrini et al., 2019; Thornton et al., 2009; Tollenaar et al., 2017; Yang et al., 2019). Therefore, the importance of SSR on all aspects of the environment and economy is difficult to underestimate.

In order to study the importance and impact of SSR on climate, environment, and other areas, it is of utmost importance to have reliable data on it, which can be used for analyzing a wide range of global and local aspects of the environment. The most commonly used SSR data are observational data collected from ground climate stations and satellite-derived alternatives which emerged in the 1980s, thanks to the significant progress in satellite algorithms (Schiffer & Rossow, 1985; Sellers et al., 1990). Ground-based observations are deemed as the most reliable and are referred to as “ground-truth” by Pinker et al. (1995), whereas observations are afflicted with the problem of missing values due to maintenance and operational failures of observational devices. On the other hand, despite of the complete spatial and temporal coverage of satellite-derived datasets, since they are only available long enough to cover the “brightening period” after the 1980s, they lack the long-term records to fully represent the multidecadal historical trends. Furthermore, enormous efforts are still needed in validation and calibration of satellite data before they can be widely applied in environmental research because the data commonly suffer from spurious changes resulting from the use of different instruments or satellites to derive the datasets, as well as the degradation algorithms of satellite sensors (Pfeifroth et al., 2018; Sanchez-Lorenzo et al, 2013, 2017). Thereby, the present study is focused on SSR estimation based on ground observations, aiming to fill missing values and obtain a temporally continuous data set with reliability and accuracy.

Considerable effort has been devoted to filling gaps in spatial datasets, which is referred to as spatial interpolation in the literature (Collins, 1995; Erxleben et al., 2002; Scudiero et al., 2016). Conventionally used spatial interpolation techniques include Inverse Distance Weighting (IDW), Kriging, splining, regression, etc. More details of the methods are referred to the data and method section. Schwandt et al. (2014) proposed a so-called “basic gap filling” method, which fills gaps of one category of solar radiation, for instance diffuse horizontal irradiance. This method applies an equation relating other types of solar radiation, such
as global horizontal irradiance and direct normal irradiance. Other gap-filling methods are also available, see Weiss et al. (2014) and the references therein. Nevertheless, machine learning methods have seen increasing applications in the area of environmental research, for example, predicting storm movement trend and growth (Han et al., 2017), halogenated substances (Wang et al., 2019), and solar radiation (Jiang, 2008). Machine learning consists of many different statistical models that computer systems use to effectively perform a specific task without using explicit instructions and strong assumptions; it relies mostly on pattern recognition and inference instead (Jiang, 2008; Sun et al., 2016; Zhou et al., 2017). In our study, we carry out a comprehensive performance comparison between a machine learning model, the Random Forest (RF), and seven other conventional spatial interpolation methods, including Ordinary Kriging (OK), three regression models, and their respective Regression Kriging (RK) methods in the application of predicting SSR values in a global scale solar forcing observation-based data set. It is worth noting that these methods have different approaches to estimate a missing value. This means that for some models it is only possible to use a particular set of variables to derive an estimate of the missing value, whereas another method might use the same variables and also other. However, the main objective in this study is to investigate which method gives the most accurate estimate of the missing values regardless of the number of input variables.

The highlight of our model is that we explore not only the explanatory ability of climatic and geographical variables, but also the contribution of intrinsic spatial and temporal dependence of the response. Specifically, three groups of predictors are constructed: i) climatic and geographical variables, such as cloud coverage, temperature, etc., to describe the climatic status of stations; ii) spatial neighboring variables, to reflect spatial autocorrelation of SSR in the vicinity of a station; and iii) time series variables, which are variants of lagged SSR, to account for seasonality and temporal autocorrelation. To the best of our knowledge, the first group of predictors are commonly used (Sun et al., 2016; Zhou et al., 2017), whereas the latter two groups are rarely contemporarily included as a part of a regressor ensemble. Nevertheless, we show that the groups ii) and iii) have significant explanatory power for predicting SSR. As a matter of fact, the RF permutation variable importance shows that the neighboring variables actually carry the highest importance, in particular, the best neighboring variable is almost twice important as the best time series variable and nearly 40 times important as the best climatic variable.

2. Data and Method

2.1. Graphical Illustration of Data and Method Output

In this section, we illustrate the SSR data set, and how the output of our methods fills missing values of SSR observations.

Panel (a) in the left column of Figure 1 illustrates our initial data set, where the y-axis is the station and the x-axis is time. Thus, each geographical station measures SSR at time t. An observation is illustrated with red colored circles. However, some observations are missing, and are represented by the green colored circles. Our objective is to apply various quantitative methods to estimate the green circles. Estimates of SSR are illustrated in Panel (a) in the right column of Figure 1. Panels (b) in the left and right column illustrate three hypothetical stations with SSR values, points in time where the SSR is missing (left column), and how the SSR time series would be after inserting the estimates of SSR (right column) for any missing values. Furthermore, Panel (c) in the left column takes a randomly selected station with actual observations and missing values, and plots the SSR of the time period analyzed. Panel (c) in the right column illustrates a complete SSR time series with both estimated and observed values of the SSR. It is worth noting the seasonality of SSR, which implies that a simple linear interpolation will not work well if the missing value is located in the maximums (minimums) during the year, for example, month t: a linear interpolation will produce the mean of the months t − 1 and t + 1. A more sophisticated method can, for example, apply information about the SSR in that particular month the year before, t − 12, in addition to other information (neighboring stations, climatic information, etc).

The middle column in Figure 1 illustrates the machine learning process applied in the study and how it is used to achieve the results.
2.2. Climatic Datasets

2.2.1. The GEBA Global Radiation Data set

The SSR (measured in Wm\(^{-2}\)) used in this study is provided by the Global Energy Balance Archive (GEBA, Wild et al., 2017), which is a data set maintained by the Institute for Climate and Atmospheric Science at ETH Zurich. It stores energy fluxes at the surface, mainly SSR measurements. These data are collected from climate stations worldwide and it provides the longest record for monthly SSR starting from the early 1950s and up to recent time. Its spatial and temporal coverage is exceptional, and therefore makes GEBA one of the best choices among datasets for this type of variable for testing our spatial interpolation models.

2.2.2. The CRU-TS Climate Data set

The Climate Research Unit (CRU)-TS 4.04 climate data set comprises monthly time series of meteorological variables that might be used as predictors for SSR prediction (Harris et al., 2020). Data are available for all global land areas, excluding Antarctica, and can be downloaded from the CRU website. It has a spatial resolution of 0.5 × 0.5° grid cells. For each grid cell, we obtain from CRU-TS observations for the following climatic variables:

- cld: cloud cover as percentage
- dtr: diurnal temperature range in °C
- frs: number of days with ground frost
- pre: monthly total precipitation in mm/month
- tmn: minimum temperature in °C
- tmp: mean temperature in °C
- tmx: maximum temperature in °C
- vap: vapor pressure in hPa
- wet: number of rain days

2.2.3. Urbanization and Elevation Data set

The Global Rural-Urban Mapping Project (GRUMP, CIESIN, 2004) data set provides a 30 arc-second urban extends grid for all land, except Antarctica and parts of the Greenland ice sheet. Locations of the climate stations contributing to the GEBA data set are classified as rural or urban based on the value of the grid cell containing the location of each climate station.
Elevation data are available from the “TerrainBase” (TBASE) data set, which is supported by the National Center for Atmospheric Research; a 0.5 by 0.5° resolution elevation file is available from http://research.jisao.washington.edu/data_sets/elevation/.

The datasets are joined together by extracting the grids in which at least one GEBA station exists such that for each station in the GEBA data set a number of climatic and geographical variables are available to be used as predictors for SSR.

2.3. Random Forest Model

Among a variety of machine learning models, the RF model provides a distinctive solution for accommodating a high dimension of the covariates; specifically, it is effective, straightforward, and has a relatively low computational intensity (Grabska et al., 2020). This makes it practical and feasible to implement in our setting, and suitable for analyzing large datasets which, on the other hand, can be difficult to reach a convergence for other machine learning algorithms' (Gradient Boosting Machine (GBM) and Support Vector Machine (SVM) were also applied to our data set trying to find machine learning alternatives; however, due to computational complexity resulting from our large global data set and the number of covariates, these models do not converge to a global optimal solution in our case. We conclude that these models are not appropriate to our problem, and therefore will not be discussed further in our analysis. There are, however, many settings where SVM and GBM provide excellent possibilities for applications, see, for example, Ma and Guo (2014)).

The RF algorithm is an ensemble classifier that applies bootstrap samples to construct multiple decision trees, where each decision tree is built on a random subset of the training samples, see Breiman (2001) for a thorough introduction to the method. During the tree growing process, the best split of the data is determined through \( n \) randomly selected features. The advantages of RF are that it can select the most important variables at each node split and it can deliver good predictive performance even when most predictive variables are noisy (Li, 2013). Moreover, a variable importance analysis can easily be conducted so that it is possible to obtain an overview of the contribution for each explanatory variable.

The input variables of RF are selected based on previous studies and empirical analyses. The predictors can be categorized into three classes: climatic and geographical variables, spatial neighboring radiation variables, and time series variables. The first group of variables encompasses the following indicators: the meteorological variables from the CRU-TS, the temporal stamps (year and month of the observation) together with the geographical characteristics (longitude, latitude, elevation, and rural/urban). Spatial neighboring variables as a group describes the spatial distribution of SSR in the neighborhood, with respect to its level, spread, and tendency, which will be elaborated in detail at what follows. Time series variables are lagged SSR observations to account for autocorrelation and seasonality in the data.

Neighborhood is constructed such that we are able to use the data from nearby areas provided that these sections share similar spatial dynamics with the target location (Ohashi & Torgo, 2012). Our intuition is that radiation within a neighborhood is correlated, also known as spatial autocorrelation, and therefore spatial dependency might be useful in predicting the value at a target location. In this study, we define three layers of neighborhoods with the search scope limited by \( k_1, k_2, k_3 \), respectively. Neighborhood size should be limited to relatively small values because as the distance between climate stations increases, the similarity decreases and there is most likely less useful information we could extract from the neighbors. We define a spatial data set \( Z = \{ z_1, z_2, ..., z_n \} \), where \( z_i \) is the value of radiation \( z \) at location \( i \). Let \( N_o^k = \{ z_i \in Z : d(o, i) \leq d(o, j), \forall i \neq o \} \) be the neighborhood of station \( o \), where \( d(o, i) \) indicates the distance between station \( o \) and \( z_i \), and \( d(o, j) \) denotes the distance between station \( o \) and its \( k \)th nearest station \( j_k \).

We use the notation \( \bar{z}(N_o^k) \) for the average of \( z_i \) in the neighborhood \( N_o^k \), and \( \bar{z}(N_o^k) \) for the weighted average counterpart, based on the IDW. The weighted average is then given by:
Fig. 2. Neighborhood illustration. o is the target station, × indicates a station within the boundary, and × a station outside.

This means that \( \lambda_i \) measures the relative inverse distance weights between stations, where \( r \) is the specified exponent parameter that amplifies the relevance of nearer neighbors, while reducing the weights of further apart neighbors. A higher value of \( r \) indicates a larger influence of nearby points on the predicted values. Rojas-Avellaneda (2007) suggests an exponent ranging from 1 to 4, whereas a more recent study (Faiz et al., 2019) states an exponent of 5 outperforms an exponent value of 2.

To give more weight to stations closer to the target station, we use the same equation as above, though with \( \lambda_i = \frac{1}{k} \), where \( k \) is the number of spatial stations in the neighborhood \( N^k_o \). Figure 2 illustrates an example for the two nearest neighborhoods, \( N^k_o \). The average thereby consists of the mean of the two red stations within the green circle.

To capture the spread of the values, we use the standard deviation of the values within this neighborhood, given by

\[
\sigma_z(N^k_o) = \sqrt{\frac{1}{N^k_o} \sum_{i \in N^k_o} \left( z_i - \bar{z}(N^k_o) \right)^2 }
\]

\( \sigma_z(N^k_o) \) measures the volatility of SSR within the neighborhood \( N^k_o \). The ratio between two averages of two spatial neighborhoods are calculated to describe the tendency, when the target variable values evolve in the vicinity of this location. The ratio is defined as follows,

\[
\bar{Z}^{k_1,k_2}_o = \frac{\bar{z}(N^{k_1}_o)}{\bar{z}(N^{k_2}_o)}
\]

where \( k_1 \) and \( k_2 \) are two neighborhood sizes \( (k_1 < k_2) \) and \( \bar{z}() \) is the average of a set of the values in the neighborhood of \( o \). This means that \( \bar{Z}^{k_1,k_2}_o \) captures spatial trends, and can provide very useful information in determining any missing values. Such trends can be the latitude and longitude, as well as elevation of the grid, and thus captured by the specific variables measuring this. Such a trend can also be moving toward, or away, from the ocean, closer to densely populated areas, etc. A variation of this indicator is the weighted average counterpart given by

\[
\bar{Z}^{k_1,k_2}_o = \frac{\bar{z}(N^{k_1}_o)}{\bar{z}(N^{k_2}_o)}
\]

where \( \bar{z}() \) is the weighted average of a set of the values in the neighborhood of \( o \).

To summarize, we have spatial neighboring variables containing the following predictors, \( \bar{z}(N^{k_1}_o) \), \( \bar{z}(N^{k_2}_o) \), \( \bar{z}(N^{k_3}_o) \), \( \bar{Z}^{k_1,k_2}_o \), \( \bar{Z}^{k_1,k_3}_o \), \( \bar{Z}^{k_2,k_3}_o \), \( \bar{Z}^{k_1,k_2,k_3}_o \), \( \sigma_1(N^{k_1}_o) \), \( \sigma_1(N^{k_2}_o) \), \( \sigma_1(N^{k_3}_o) \), where \( k_1 \), \( k_2 \), and \( k_3 \) satisfying \( k_1 < k_2 < k_3 \).

Moreover, autoregressive terms are introduced to remove trend and seasonality. For the same reason, we also include moving average terms to reduce the effect of short-term fluctuation. The trend in radiation is removed by including monthly lagged values, that is, RAD_{t-1}, RAD_{t-2}, RAD_{t-3}, ..., Meteorological varia-
bles are likely to show the propensity of seasonality that previous years’ radiation in the same month has significant correlation with that of the current month. To account for this, we introduce annual lagged values, that is, $\text{RAD}_{t-12}$, $\text{RAD}_{t-24}$, $\text{RAD}_{t-36}$, ..., where, for example, $\text{RAD}_{t-12}$ is the observed RAD 12 months before month $t$.

Moving average terms are often used for smoothing out short-term fluctuations and highlight long-term trends or cycles, therefore we introduce moving average terms for both monthly lagged radiation and annual lagged radiation, with $\text{RAD}_{\text{mamk}}$ representing the moving average of the previous $k$ months’ radiation and $\text{RAD}_{\text{mayK}}$ for the counterpart of the yearly lagged radiation with $K$ annual lags.

2.4. Ordinary Kriging

OK applies a particular function, called a semivariogram, to measure the spatial dependence between observations. Assumed to be intrinsically stationary, the semivariogram $\gamma(h)$ is defined as

$$\gamma(h) = \frac{1}{2} \text{average} \left[ (z(x_i) - z(x_j))^2 \right]$$

where $h$ is the distance between points $x_i$ and $x_j$, whereas $z(x_i)$ is the measured value at sample point $x_i$. Equation 5 means finding all pairs of points $x_i$ and $x_j$ that are separated by a distance $h$, and then calculating the average squared difference and dividing by 2. The empirical semivariogram provides insights on the spatial autocorrelation of observations in datasets. It represents the average rate of change in the target variable with distance $h$. The semivariogram is fitted with sample data and used to derive the weights of an unbiased linear interpolator by means of minimizing the error estimate variance $\text{Var}(z^*(x_o) - z(x_o))$. The weights can be used to make a prediction by

$$z^*(x_o) = \sum_{i=1}^{n} \omega_{OK}^{i} z(x_i)$$

where $\omega_{OK}^{i}$ is the semivariogram weight, $z^*(x_o)$ is the linear estimator of the unknown true value $z(x_o)$.

2.5. Regression Kriging

RK combines regression predicted trends with Kriging predicted residuals where the trend is obtained by regressing the response on auxiliary variables (such as climatic and geographical variables) and the spatial dependence of the response is captured by a Kriging model. The RK model is given by

$$\hat{z}(x_o) = \hat{m}(x_o) + \hat{e}(x_o)$$

$$= \sum_{k=0}^{p} \beta_k \cdot q_k(x_o) + \sum_{i=1}^{n} \hat{\omega}_i \cdot e(x_i)$$

where $\hat{m}(x_o)$ is the fitted trend that is explained by the variation of climatic and geographical variables, and $\hat{e}(x_o)$ is the interpolated residual that incorporates spatial neighboring information of the values of its neighbors. $\hat{\beta}_k, k = 0, ..., p$, is the estimated regression coefficients, $p$ is the number of auxiliary variables, $q_k(x_o) = 1$ is a constant term, and $\hat{\omega}_i, i = 1, ..., n$, are the Kriging weights.

In the RK, the group of climatic and geographical variables is used as trend regressors; the spatial dependency within the neighborhood is captured by the Kriging fitted residuals. Three methods of regression are applied, that is, simple linear regression (LR), generalized additive regression (GAM), and least squares dummy variable (LSDV) model. Figure 3 shows the distribution of SSR per month for Europe. It is noteworthy that SSR varies with month nonlinearly, therefore it makes sense to include highly polynomial terms for month, which is realized by the GAM model. Another way of capturing the various levels of SSR is to introduce 11 new dummy variables to represent the 12 months (LSDV); when the 11 dummy variables all equal to zero, it indicates the reference month.
3. Empirical Results

We start out with a data set of 1,523 stations scattered around land areas on Earth. Figure 4 shows an overview of how the stations are distributed. We can see that the stations are not equally distributed in all regions. Middle and Southern Europe and Japan own the most concentration of stations, whereas northern North America, Asia, and Africa have relatively sparse, though even, distribution, while in South America and Australia stations are mostly located at the land-sea borders of the continents. There are also a few stations scattered on islands across the oceans. Each station registers the average downward shortwave solar radiation each month during the year. The first stations were starting to operate in the early 1950s. Due to data maintenance and quality checks, a substantial number of observations are only available through 2013. This leaves us with about 60 years of data. However, due to instrument failure and other errors, some
months’ observed values are not registered. This leads to a so-called missing value. One of the main tasks in this study is to estimate those missing values using advanced statistical methods. We implement a data quality control process that prepares the data set for the input of RF. The following stations are dropped in the data set: (1) stations that exist for less than three years; (2) stations that do not have at least once three consecutive months of observations; and (3) stations for which the layers of neighborhoods we defined above before cannot be found. After the data quality control, we are left with data from 1,227 stations, see Figure 5 for the number of observations every year.

The following evaluation metrics are used to assess model prediction performance: $R^2$, mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared error (RMSE).

\[
R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{z}_i - z_i)^2}{\sum_{i=1}^{n}(z_i - \bar{z}_i)^2}
\]

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{z}_i - z_i|
\]

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{z}_i - z_i}{z_i} \right|
\]

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{z}_i - z_i)^2}
\]

where $n$ is the number of predictions, $\hat{z}_i$ is the predicted value by the model, and $z_i$ is the observed value.

### 3.1. Parameter Setup for Regressors

#### 3.1.1. Define the Neighborhood Size

Defining the size of the neighborhood is challenging, as it is desirable to include as many relevant stations as possible while excluding irrelevant stations. Since we are most interested in how radiation from
a number of the nearest stations coincides with the target station, the neighborhood should not be too large, nor too small. We define neighborhood size by the number of nearest stations that are included, moreover, we add a cut-off distance that intends to avoid bias from outlier stations. Figure 6 shows the distribution of natural logarithmic distance between the second nearest stations. Given a roughly symmetric distribution of logarithmic distance, it indicates a fat right tail in the raw distance distribution, which confirms our statement about the existence of distant stations. Since these outlier stations are isolated, even their nearest stations do not share similar characteristics with them, leading us to provide cutoff distances to confine the definition of neighborhoods. The cutoff distances are determined by the 95 percentile of distance between the nearest 2, 3 and 5 neighbors (The largest neighborhood is defined by the five stations nearest to the target location. The definition refers to the Inverse Distance Weighting (IDW) in that, generally, when the size of neighborhood stretches to the sixth nearest station, the weight given to the station is small enough to be negligible as compared to other nearer stations). In summary, we set \( k_1 = 2, k_2 = 3, k_3 = 5 \), with cutoff distances being \( d_1 = 691 \) km, \( d_2 = 790 \) km, \( d_3 = 1047 \) km, respectively.

### 3.1.2. Order of Seasonal Autoregression

To see the characteristics of radiation evolution in the past, Figure 7 shows the time series of radiation from a randomly picked station. As expected, we observe that the radiation demonstrates strong annual cycles, and we hypothesize that previous radiation from the same month from previous years may have close correlation with that of any future months for the station.

In order to test the seasonality, Figure 8 plots the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the radiation time series. Both figures show significant annual seasonality. The ACF shows spikes at lags for multiples of 12, and the PACF shows exponential decay in seasonal lags, which means that the importance of lagged values decreases as the lag increases. Based on these results, we add \( \text{RAD}_{t-K\times s} \) seasonal lagged terms as predictors for \( \text{RAD}_t \), where \( s = 12 \) is the seasonal length, and \( K = 1, 2, ..., P \), where \( P \) is the optimal order of seasonal lagged terms.
3.1.3. **Order of Autoregression**

We also want to investigate how previous monthly lagged radiation affects future values. Therefore, we add \( \text{RAD}_{t-k}, k = 1, \ldots, p \), where \( k \) stands for \( k \) months' lagged radiation, where \( p \) is the optimal lag-length, as predictors for \( \text{RAD}_t \).

3.1.4. **Seasonal Moving Average**

Moving average is a calculation to smooth data, it uses the arithmetic mean of the last several observations to forecast the next observation. We want to compute the moving average of the last \( K \) SSR observations for the same month. The following formula is used to calculate yearly moving average of radiation:

\[
\text{Moving Average} = \frac{1}{K} \sum_{i=1}^{K} \text{RAD}_{t-i}
\]
where $K = 2, 3, \ldots$, and $s$ is the seasonal length equal to 12. An optimal number $K$ of observation could be found by calculating the MAE for $n$ arguments and choosing the one with the minimum MAE.

### 3.1.5. Moving Average

A smoothed trend of past few months’ radiation is represented as the average of the last $k$ terms, that is, the moving average term. We add this term to reflect how previous months’ trend affect the future value. The monthly moving average is defined as:

$$RAD_{t, \text{mam}} = \frac{\sum_{i=1}^{k} RAD_{t-i}}{k}$$

where $k = 2, 3, \ldots$. The number of previous months, $k$, is also optimized by minimizing the error MAE.

To summarize the parameters, we use lower case $k$ and $p$ to represent the optimal order of moving average terms and lagged terms respectively in ordinary monthly autoregression; whereas the upper case $K$ and $P$ represent the counterparts in seasonal autoregression.

### 3.1.6. Implementation of the Random Forest Model

We use a “grid search” to iteratively explore different combinations of parameters for the autoregressive terms and the moving average terms. For each combination of parameters, we fit a RF model and calculate the MAE. Once we have explored the entire set of parameters, one optimal combination of parameters will be the one that yields the best performance for the estimation. In our case, this turns out to be: $p = 1$, $k = 3$, $P = 3$, $K = 3$. Table 1 provides a summary of the input predictors for all models.

Two sources of randomness in the RF model are the number of trees and the number of candidate independent variables to split at in each node. The first parameter determines how large the forest is, that is, the number of bootstraps to generate independent samples, and the latter parameter decides the number of randomly preselected variables to split at in each tree. The number of trees is specified as 700 and the number of candidate independent variables is set as 6 according to a tuning process. The relationship between MAE and the two model parameters can be found in the supplement material.

### Table 1

| Variable category                        | Constituent variable                                      |
|------------------------------------------|-----------------------------------------------------------|
| Response/Dependent variable¹              | SSR                                                       |
| Climatic and geographical Variables²     | CRU climatic variables: cld, dtr, frs, pre, tmn, tmp, tmx, vap, wet |
|                                          | Geographical variables: longitude, latitude, elevation, rural/urban |
|                                          | Time indicators: year, mon |
| Spatial neighboring variables³            |                                                                 |
|                                          | Level indicators: $\sum(N^i_k)$, $\sum(N^{i-1}_k)$, $\sum(N^{i-2}_k)$, $Z^{i-1}_u$, $Z^{i-2}_u$ |
|                                          | Spread indicators: $\sum(N^i_k)$, $\sum(N^{i-1}_k)$, $\sum(N^{i-2}_k)$, $Z^{i-1}_u$, $Z^{i-2}_u$ |
|                                          | Tendency indicators: $\sum(N^i_k)$, $\sum(N^{i-1}_k)$, $\sum(N^{i-2}_k)$, $Z^{i-1}_u$, $Z^{i-2}_u$ |
| Time series variables⁴                    |                                                                 |
|                                          | Autocorrelation variables: $RAD_{t-1}$, $RAD_{t-2}$, $RAD_{t-3}$, $RAD_{t-4}$, $RAD_{t-5}$, $RAD_{t-6}$ |
|                                          | Seasonality variables: $RAD_{t-12}$, $RAD_{t-24}$, $RAD_{t-36}$, $RAD_{t-2}$, $RAD_{t-3}$, $RAD_{t-4}$ |

Random Forest input variables: 2, 3, 4; Ordinary Kriging input variable: 1; Regression models input variables: 2; Regression Kriging models input variables: 1 and 2; RAD and SSR are used interchangeably throughout the text. CRU, climate research unit; SSR, surface solar radiation.
Figure 9 plots the natural logarithm of permutation variable importance for the first 10 most important variables in the RF model. The importance is first computed as the R squared percentage change if we permute the target variable and keep the remainders unchanged, and then scale the importance to its logarithm in order to mitigate the importance difference in visualization. We observe that the most influential variable is one of the spatial neighboring variables (labeled in violet), $z_{k3}$, which is the simple average of radiation of the $k_3$ neighborhood, consisting of the nearest five climate stations. It indicates that the stations within one neighborhood share significant similar dynamics of radiation, and therefore their observations could be used as powerful explanatory variables to predict the value at any missing locations. Another variable within the same class is its weighted average variant, $zw_{k3}$, which is the fourth most useful variable that also explains a large part of the radiation variation. The time series variables (labeled in magenta), especially the 3-year moving average radiation on the same month (RAD_may3) and the previous month’s radiation (RAD_tm1), are the second and third most important indicators, which accords with our previous observation of autoregression and seasonality in the time series of radiation. If we transform to actual importance by taking natural exponent, we observe that the most important neighboring variable is indeed almost twice as important as the best time series variable. By contrast, the group of climatic variables is not as important as the preceding two groups, as the importance decays very fast as ranks decrease. Among the group of meteorological variables (labeled in light blue) that describes the state of climatic and geophysical information of stations, the ones that provide the most explanatory power are cloud coverage (CLD), latitude (LAT), and precipitation (PRE). Nonetheless, if we focus on the actual importance, the best neighboring variable ($z_{k3}$) is almost 40 times more important than the best climatic variable (CLD).

3.2. Evaluation for Prediction Errors

Due to the differing, or even opposing, climatic and geographical characteristics for various continents, we implement model training and testing on the continental level. A 10-fold cross validation (CV) is implemented for the RF. For RK, trend regression and spatial dependence Kriging needs to be trained separately. A 10-fold CV is first implemented and generates out-of-sample predictions for SSR trends. On the other hand, OK can only be trained for a static cross sectional map, which allows only one observation for each unique location. Hence data training and testing for Kriging is carried out monthly. Thereafter, spatial dependence, represented by residuals of regression models, is fed to OK by a monthly 10-fold CV, and gener-
By adding regression trends and Kriging residuals together, we eventually obtain our final predictions.

Table 2 shows error metrics for each continent. By comparing the values across columns, we observe the performance difference among various models. The four models with the best overall performance are LEIRVIK AND YUAN.

Table 2  
Evaluation Metrics

| Measure | Continent | Models | RF | LR | GAM | LSDV | OK | LR + OK | GAM + OK | LSDV + OK |
|---------|-----------|--------|----|----|-----|------|----|--------|----------|------------|
| MAE     | Africa    | 11.73  | 30.78 | 30.15 | 30.14 | 26.7 | 25.4 | 25.37 | 25.35 |
|         | Asia      | 10.7   | 25.23 | 19.34 | 19.34 | 16.82 | 17.25 | 15.81 | 15.81 |
|         | Europe    | 7.93   | 30.74 | 14.15 | 14.14 | 11.83 | 18.18 | 11.29 | 11.28 |
|         | North America | 9.69 | 30.91 | 15.8 | 15.79 | 13.75 | 16.04 | 12.91 | 12.91 |
|         | Oceania   | 9.52   | 30.96 | 14.98 | 14.99 | 20.49 | 20.91 | 13.98 | 14.05 |
|         | South America | 11.70 | 31.18 | 30.52 | 30.53 | 29.98 | 29.73 | 29.69 | 29.70 |
| Average |          | 10.21  | 29.97 | 20.82 | 20.82 | 19.93 | 21.25 | 18.18 | 18.18 |
| Standard deviation | 1.46 | 2.33 | 7.58 | 7.58 | 7.22 | 5.33 | 7.52 | 7.51 |
| MAPE (%) | Africa    | 5.31   | 14.3 | 13.95 | 13.94 | 12.7 | 11.63 | 11.6 | 11.59 |
|         | Asia      | 7.17   | 18.22 | 13.43 | 13.43 | 11.67 | 11.59 | 10.73 | 10.73 |
|         | Europe    | 7.08   | 45.27 | 17.27 | 17.26 | 12.39 | 23.72 | 12.58 | 12.59 |
|         | North America | 6.65 | 32.57 | 16.62 | 16.62 | 10.87 | 15.3 | 14.44 | 14.46 |
|         | Oceania   | 5.66   | 22.11 | 11.4 | 11.42 | 14.79 | 15.7 | 11.3 | 11.32 |
|         | South America | 6.44 | 18.74 | 18.23 | 18.22 | 17.62 | 17.21 | 17.2 | 17.2 |
| Average |          | 6.39   | 25.2 | 15.15 | 15.15 | 13.34 | 15.86 | 12.64 | 12.65 |
| Standard deviation | 0.76 | 11.63 | 2.63 | 2.62 | 2.47 | 4.47 | 2.34 | 2.34 |
| RMSE    | Africa    | 15.88  | 39.08 | 38.63 | 38.63 | 35.6 | 33.54 | 33.54 | 33.5 |
|         | Asia      | 14.79  | 32.38 | 25.87 | 25.87 | 23.56 | 24.5 | 22.12 | 22.12 |
|         | Europe    | 12.12  | 39.05 | 19.72 | 19.72 | 18.06 | 26.57 | 16.94 | 16.95 |
|         | North America | 13.75 | 39.13 | 21.98 | 21.97 | 20.4 | 22.07 | 18.34 | 18.35 |
|         | Oceania   | 13.29  | 39.75 | 21.54 | 21.56 | 29.99 | 30.25 | 22.81 | 22.9 |
|         | South America | 16.17 | 39.72 | 39.02 | 39.03 | 39.43 | 37.99 | 37.96 | 38.02 |
| Average |          | 14.33  | 38.19 | 27.79 | 27.80 | 27.84 | 29.15 | 25.29 | 25.31 |
| Standard deviation | 1.57 | 2.86 | 8.78 | 8.78 | 8.58 | 5.95 | 8.52 | 8.52 |
| R²      | Africa    | 0.90   | 0.38 | 0.39 | 0.39 | 0.48 | 0.54 | 0.54 | 0.54 |
|         | Asia      | 0.94   | 0.70 | 0.81 | 0.81 | 0.84 | 0.83 | 0.86 | 0.86 |
|         | Europe    | 0.98   | 0.77 | 0.94 | 0.94 | 0.95 | 0.9 | 0.96 | 0.96 |
|         | North America | 0.97 | 0.77 | 0.93 | 0.93 | 0.94 | 0.93 | 0.95 | 0.95 |
|         | Oceania   | 0.97   | 0.78 | 0.93 | 0.93 | 0.87 | 0.87 | 0.93 | 0.93 |
|         | South America | 0.88 | 0.31 | 0.34 | 0.34 | 0.31 | 0.36 | 0.36 | 0.36 |
| Average |          | 0.94   | 0.62 | 0.72 | 0.72 | 0.73 | 0.74 | 0.77 | 0.77 |
| Standard deviation | 0.04 | 0.21 | 0.28 | 0.28 | 0.27 | 0.23 | 0.25 | 0.25 |

Note. Numbers in bold face represent best fit. We see that the Random forest methods is better for all continents, for all performance metrics, whereas linear regression performs the worst. MAE and RMSE are measured in Wm⁻². MAPE is measured in %. R² is the coefficient of determination by regressing predicted values on observed values.

GAM, generalized additive regression; LR, linear regression; LSDV, least squares dummy variable; MAE, mean absolute error; MAPE, mean absolute percentage error; OK, ordinary kriging; RF, random forest; RMSE, root mean squared error.
RF, OK, GAM + OK, and LSDV + OK, we will mainly focus on these models thereafter. The RF model has the most accurate prediction with the highest stability among all the models no matter at which measure we are comparing. For instance, if we look at the MAEs, the RF has the lowest global average MAE of 10.2 Wm\(^{-2}\) and a standard deviation of 1.5 Wm\(^{-2}\), while the ensemble of the other seven methods generates an average MAE of 21.3 Wm\(^{-2}\), and an average of standard deviation of 6.4 Wm\(^{-2}\). By comparing the values row-wisely, we see how model performance varies across continents. Note that the relative accuracy among continents, as indicated by MAE, RMSE and \(R^2\), generally coincides, while contradicts what is reflected by MAPE. Taking the RF model as an example, Europe has the lowest error indicated by MAE, RMSE, and \(R^2\), whereas it has the second largest error, only smaller than Asia, if measured by MAPE. Conversely, Africa and South America perform relatively poor among all the continents in terms of absolute error terms, yet they perform better using the MAPE. One possible explanation could be that the percentage error tends to be biased by the absolute levels of observations. In other words, the small MAPE value (5.31%) in Africa does not necessarily reflect a high accuracy, the value is merely biased downward by the average high levels of SSR in Africa, therefore showing a potential false high accuracy. Consequently, the absolute error measures, that is, MAE and RMSE, are more appropriate indicators for model evaluation regarding SSR in our case. It is worth noting that another preferable advantage of RF is that observation density is not as influential on model performance as the other models. We see from the station map that South America has the most clustered and uneven station placement, and subsequently, we observe a significant decrease in performance in the continent for the conventional RK methods (e.g., MAE increases from Europe's 11.3 to South America's 29.7, with an increase of 18.4 Wm\(^{-2}\)), whereas the errors of the RF model are not deteriorating as much (e.g., MAE increases from Europe's 7.9 to South

**Figure 10.** Simulation against observation values in Europe. The figure illustrates that the random forest (RF) model produces simulated values with a lower variability than the other models. See the supplement material for corresponding figures for the other continents.
America’s 11.7, with a much smaller increase of 3.8 Wm\(^{-2}\)). The relative stability of RF’s performance indicates the capacity for its potential applications in spatial interpolation in sparsely and unevenly scattered data points settings.

The three regression models (LR, GAM, and LSDV) generate profoundly different errors, with LR having the largest errors, and GAM and LSDV having substantially lower errors. LR differs from the other two models in how it copes with the MON (month) variable. The strikingly larger errors for LR indicates that regarding MON as a regular numeric variable is not adequate, and a special treatment is apparently needed. Moreover, the results show that it has quite similar beneficial effects either introducing MON as high polynomial terms (in GAM) or treating each month like they have different levels (in LSDV). The OK model generally outperforms regression models to a certain degree, which indicates the essential importance of including spatial autocorrelation in prediction. The OK method fails to outperform regression in Oceania in which stations are scattered especially sparsely. Additionally, it does not work as effectively in South America given that the error reduction for OK is nuanced (less than 1 Wm\(^{-2}\) for MAE). The outperformance of OK is nonetheless noticeable in Europe. The good performance in data rich continents and deteriorating performance in data sparse continents indicates that the performance of the OK model is highly determined by data density. The RK methods incorporate spatial dependence by adding a spatial correlated residual term to the the trend prediction by regression. The effectiveness of this method is confirmed by its lower

---

**Figure 11.** Surface solar radiation (SSR) monthly anomalies for one chosen station. Each panel encompasses monthly SSR anomalies (dashed line) of observations (red) and model simulations (blue), together with their corresponding smoothed series (solid line) by a 12-month Gaussian kernel. The series are expressed as anomalies from the 1964–2013 mean. Corresponding prediction models for each panel: Panel (a) RF, Panel (b) OK, Panel (c) GAM + OK. Since LSDV + OK predictions have very similar trends with GAM + OK, we do not show it here. Refer to supplement material for the details of LSDV + OK. GAM, generalized additive regression; LSDV, least squares dummy variable; OK, ordinary kriging; RF, random forest.
error metrics as compared to regression methods alone. In Figure 10, predicted values in Europe are plotted against observed values for the four best performance models. We observe that all coefficients of determination ($R^2$) are above 95%, indicating the models' remarkable ability in predicting more than 95% of the response variance. The narrow scatter of points alongside the regression line in RF suggests that the RF model generates predictions with low variability.

Figure 11 shows monthly anomalies for one chosen station, located at Locarno-Monti (46.17°N, 8.78°E) in Switzerland, which is one of the most long-standing climate observation stations in Europe. RF model shows a relative advantage in reproducing SSR trends, in particular during the highly volatile period 1970–1980, during which SSR fell largely. Although the RF predictions do not exactly reproduce the observations, it captures the general trends better than either the OK or GAM + OK (LSDV + OK) model.

In Figure 12, monthly anomalies are plotted for Europe (See Supplement Material for monthly anomalies for the other continents). The continental aggregated predictions for all the three models replicate trends of observations reasonably well. By comparing Figures 11 and 12, it is worth noting that RF is more capable of capturing idiosyncratic trends of single stations, while the aggregated performance is equally comparable with that of OK and GAM + OK (LSDV + OK).
4. Trends in Global and Continental Average Surface Solar Radiation

In this section, annual global and continental average trends for SSR are explored. Figure 13 shows the cumulative annual global average radiation change from 1964 to 2013. Year 1964 is chosen as radiation observations are few, clustered in a few continents, and fluctuate a lot before this year. Including data before that might thus bias the SSR trends. The global SSR first experienced a decrease (global dimming) until the beginning of the 1990s, and started to increase onward (global brightening). The trend could be broadly explained by the changes in global anthropogenic sulfur emissions shown in Stern (2006). According to Stern, global anthropogenic sulfur emissions experienced drastic increase during the period from 1960 to 1980, thereafter the emission level fluctuated from 1980 to 1990, and started to decrease afterward. The exact correlation between SSR and sulfur emissions is examined further in what follows.

Annual cumulative continental average change is shown in Figure 14 and 15 for all continents except for Antarctica, as it has too few observations to form a valid basis for trend analysis. Note that Storelvmo et al. (2018) presented a similar continental SSR figure, where the authors used a balanced GEBA data set together with interpolated missing values. This contrasts our paper, where we aim to reproduce existing observations as precisely as possible. The trends reported in Storelvmo et al. (2018) are generally in agreement with our results, though with a relatively stronger trend in our study. It is interesting to notice that the spatial difference of SSR trends is intriguingly significant among continents. SSR over Europe has a trend of increase during the period 1967–1976, at what follows, a decrease has been observed until the mid of the 1980s. Beyond that, SSR enters an accelerated increasing period until 2013. The recent brightening period corroborates the pronounced effect of sulfur emissions regulations imposed in Europe starting from 1990 (Stern, 2006). Similar regulations were enforced in North America around the same time, though the result is not as profound, the declining trends cease nevertheless.

The trends in Asia resemble that of North America; the distinct decrease prevails until the beginning of 1990s and SSR fluctuates around a zero-trend afterward. The remarkable decrease is mostly attributed to the increase in aerosol loading caused by rapid economic expansion and population growth during that period (Liang & Xia, 2005). At the beginning of the 1990s, when a major economic crisis occurred in Asia, the decreasing trend ceased and SSR fluctuated onwards with no evident trend.

Figure 13. Cumulative annual global average trend for SSR from 1964 to 2013 for observations (red) and RF predictions (blue), plotted together with 5-year moving average (solid line). The series are expressed as cumulative change from 1964. RF, random forest; SSR, Surface solar radiation.
There are many different atmospheric aerosols affecting SSR, for example sulfur, soot, salt-particles from the ocean, etc. In particular, atmospheric sulfur aerosol loading is known to have the most essential impacts on SSR variation, see Wild (2016). The sulfur emissions increase Earth's albedo either by direct interaction with solar radiation, or by increasing cloud reflectivity, areal extent and etc., and thereby attenuate SSR and cause global dimming (Storelvmo et al., 2016). An inspection of the global relationship between SSR and SO$_2$, a precursor of sulfate aerosol, is discussed thereafter. Figure 16 illustrates the relationship between SSR and SO$_2$ emissions, see also Smith et al. (2011) and Klimont et al. (2013) for a thorough treatment of the matter. Significant correlation coefficients are observed as −0.601 for the period 1964–1985, and −0.607 for 1986–2011, which are in agreement with the result shown in Storelvmo et al. (2018). A similar analysis of Europe is shown in Figure 17. A jump at 2000 is noticed due to the change of data set (Two datasets are joined at 2000, where a country level is used for 1964–1999 and a regional level for 2000–2011, see Smith et al. (2011) and Klimont et al. (2013)). A discrepancy emerges when we try to aggregate datasets to generate continental emissions for Europe. Europe is difficult to reconcile for the two datasets, as some countries locate transcontinentally, for example, Russia. This aggregation problem does not exist on the global level, thereby we observe a rather smooth curve in Figure 16). To reduce bias from the jump, correlation coefficients are calculated for the subperiods split by 2000. From 1964 to 1999,
Figure 15. Cumulative annual continent average SSR trend: part II. Series representations are referred to part I of Figure 14. Panel (d) for South America (1967–1998), Panel (e) for Asia (1964–2010), Panel (f) for Oceania (1971–1991). SSR, surface solar radiation.

Figure 16. Annually global average of SSR (green line, left axis) and global emissions of SO$_2$ (blue line, right axis). Both curves show 5-year running averages. SO$_2$ data are from Socioeconomic Data and Applications Center (SEDAC). SSR, surface solar radiation.
Figure 17. Annually European average of SSR (green line, left axis) and European emissions of SO$_2$ (blue line, right axis). Both curves show 5-year running means. SSR, surface solar radiation.

the correlation coefficient is 0.137, yet insignificant, in contrast to the significant correlation coefficient of −0.971 over the period 2000–2011.

5. Conclusion

This study explores the potential of applying a machine learning approach, RF, to spatial interpolation. A wide range of explanatory variables are explored: climatic and geographical variables, spatial neighboring variables, and time series variables; among which spatial neighboring variables are proved to be the most determinant factors for the RF. A 10-fold cross validation for each continent was implemented to evaluate the model performance. The results show that (1) the RF generates the smallest prediction errors and is least affected by the data density. The relative density independence of RF is an essential property to ensure a satisfactory prediction accuracy even in a setting with scarce observations; (2) OK’s performance is strongly dependent on the density of data points. When the stations are sparsely scattered, the performance of the OK deteriorates significantly. However, the OK has higher accuracy than regression models in most cases, which highlights the essential necessity of introducing spatial autocorrelation in the prediction; (3) RK is more preferred than either having the OK or regression models alone; however, the accuracy improvement is not considerable. On the contrary, the improvement of RF is substantial. The RF model has a 44% lower MAE than the best RK model. The results show that we obtain estimates broadly consistent with Li et al. (2011) in terms of the relative better performance of RF, compared to OK, and regression Kriging methods. Li et al. asserts the superior predictive accuracy and sensitivity to inputs of RF, especially its combination with OK or IDW, out of 23 methods (including RF, support vector machines, OK, IDW and their combinations), in a spatial interpolation application of mud content samples in the southwest Australian coast. However, the combination of RF and OK show no significant improvement in accuracy in our application. One possible explanation could be that the spatial dependency is well captured by introducing the spatial neighboring variables, therefore combining OK contributes little in adding new information to the final estimation.

The performance advantage is evident, whereas there are shortcomings as well. When including a large number of explanatory variables, the criteria for admitting sample data become strict. In our particular model, a geographical station should not be isolated, and needs to have existed for a certain time in order for our model to reach an accurate prediction. This limits the model’s extrapolation ability to predict values in unsampled areas. In addition, the complexity of machine learning models makes it behave like a black box and thereby restrains our understanding in terms of interpreting exactly how explanatory variables affect radiation.

We show that RF outperforms conventional spatial interpolation methods with unparalleled accuracy and performance stability even in sparse data settings. Our end result is a panel data set, that is, information
over time and space, with monthly values spanning almost five decades. The imputed SSR data set could be applied in further climate research.

6. Software and Technical Notes

The modeling is implemented in the statistical environment R (R Core Team, 2018) and Python (Van Rossum & Drake, 1995). We apply the randomForest (Wright & Ziegler, 2017) to train the RF model. GAM model is estimated using package mgcv (Wood, 2003). OK is conducted with geostatistical packages sp (Pebesma & Bivand, 2005), gstat (Gräler et al., 2016), and raster (Hijmans, 2019). Ggplot2 (Wickham, 2016) is used in visualization. Python is used in covariates parameter estimation and data quality control. The codes for spatial interpolation models are available at https://github.com/my1396/SSR-Spatial-Interpolation.git.

Data Availability Statement

The data related to the conclusions of this study are available at http://doi.org/10.17632/wzz2pvzcrz.1.

References

Aaheim, A., Romstad, B., Wei, T., Kristjánsson, J. E., Muri, H., Niemeier, U., & Schmidt, H. (2015). An economic evaluation of solar radiation management. The Science of the Total Environment, 532, 61–69. https://doi.org/10.1016/j.scitotenv.2015.05.106

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/a:1010933404324

Budyko, M. I. (1969). The effect of solar radiation variations on the climate of the Earth. Tellus, 21, 611–619. https://doi.org/10.3402/tellusa.v21i5.10109

Calcabruni, A., Ziar, H., Isabella, O., & Zeman, M. (2019). A simplified skyline-based method for estimating the annual solar energy potential in urban environments. Nature Energy, 4, 206–215. https://doi.org/10.1038/s41560-018-0318-6

CIESIN. (2004). Global rural–urban mapping project (GRUMP), alpha version: Urban extents grids. NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4GH9FVG

Collins, F. C. (1995). A comparison of spatial interpolation techniques in temperature estimation. Virginia Tech. Retrieved from http://hdl.handle.net/10919/38139

Erxleben, J., Elder, K., & Davis, R. (2002). Comparison of spatial interpolation methods for estimating snow distribution in the Colorado Rocky Mountains. Hydrological Processes, 16(18), 3627–3649. https://doi.org/10.1002/hyp.1239

Faiz, M., Zaki, M., Ashraf, M., Ismail, M., Govindasamy, D., & Zainalabidin, M. H. (2019). Interpretation and development of top-surface grid in subsurface ground profile using inverse distance weighting (IDW) method for twin forest stand in Kenny hill formation. Bulletin of the Geological Society of Malaysia, 67, 103–109. https://doi.org/10.7186/bgsms67201911

Grabska, E., Frantz, D., & Ostapowicz, K. (2020). Evaluation of machine learning algorithms for forest stand species mapping using Sentinel-2 imagery and environmental data in the Polish Carpathians. Remote Sensing of Environment, 251, 112103. https://doi.org/10.1016/j.rse.2020.112103

Gräler, B., Pebesma, E., & Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. The R Journal, 8, 204–218. https://doi.org/10.32614/rj-2016-014

Han, L., Sun, J., Zhang, W., Xiu, Y., Feng, H., & Lin, Y. (2017). A machine learning nowcasting method based on real-time reanalysis data. Journal of Geophysical Research: Atmospheres, 122, 4038–4051. https://doi.org/10.1002/2016JD025783

Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRUTS monthly high-resolution gridded multivariate climate dataset. Scientific Data, 7, https://doi.org/10.1038/s41597-020-0453-3

Hijmans, R. J. (2018). Raster: Geographic data analysis and modeling. R package version 3.0-7. Retrieved from https://CRAN.R-project.org/package=raster

Jiang, Y. (2008). Prediction of monthly mean daily diffuse solar radiation using artificial neural networks and comparison with other empirical models. Energy Policy, 36, 3833–3837. https://doi.org/10.1016/j.enpol.2008.06.030

Klimont, Z., Smith, S. J., & Cofala, J. (2013). The last decade of global anthropogenic sulfur dioxide: 2000-2011 emissions. Environmental Research Letters, 8, 014003. https://doi.org/10.1088/1748-9326/8/1/014003

Li, J. (2013). Predicting the spatial distribution of seabird gravel content using random forest, spatial interpolation methods and their hybrid methods. In 20th international congress on modelling and simulation (pp. 394–400). Adelaide, Australia. https://doi.org/10.36334/modsim.2013.a9.li

Li, J., Heap, A. D., Potter, A., & Daniell, J. J. (2011). Application of machine learning methods to spatial interpolation of environmental variables. Environmental Modelling & Software, 26, 1647–1659. https://doi.org/10.1016/j.envsoft.2011.07.004

Liang, F. & Xia, X. A. (2005). Long-term trends in solar radiation and the associated climatic factors over China for 1961-2000. Annales Geophysicae, 23, 2425–2432. https://doi.org/10.5194/angeo-23-2425-2005

Ma, Y. & Guo, G. (2014). Support vector machines applications (p. 302). Springer International Publishing. https://doi.org/10.1007/978-3-319-02300-7

Obryk, M. K., Fountain, A. G., Doran, P. T., Lyons, W. B., & Eastman, R. (2018). Drivers of solar radiation variability in the McMurdo Dry Valleys, Antarctica. Scientific Reports, 8, 5002. https://doi.org/10.1038/s41598-018-23390-7

Ohashi, O. & Torgo, L. (2012). Spatial interpolation using multiple regression. In 2012 IEEE 12th international conference on data mining (pp. 1044–1049). IEEE. https://doi.org/10.1109/ICDM.2012.48

Pebesma, E. J. & Bivand, R. S. (2005). Classes and methods for spatial data in R. R News, 5, 9–13.
Pfeifroth, U., Sanchez-Lorenzo, A., Manara, V., Trentmann, J., & Hollmann, R. (2018). Trends and variability of surface solar radiation in Europe based on surface- and satellite-based data records. Journal of Geophysical Research: Atmospheres, 123, 1735–1754. https://doi.org/10.1002/2017JD027418

Pinker, R. T., Frouin, R., & Li, Z. (1995). A review of satellite methods to derive surface shortwave irradiance. Remote Sensing of Environment, 51, 108–124. https://doi.org/10.1016/0034-4257(94)00069-Y

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Rojas-Avellaneda, D. (2007). Spatial interpolation techniques for stimulating levels of pollutant concentrations in the atmosphere. Revista Mexicana de Física, 53, 447–454. Retrieved from http://www.scielo.org.mx/scielo.php?script=sci_abstract&pid=S0035-001X2007000600041lng=en&nr=1&iso=es

Sanchez-Lorenzo, A., Enriquez-Alonso, A., Wild, M., Trentmann, J., Vicente-Serrano, S. M., Sanchez-Romero, A., et al. (2017). Trends in downward surface solar radiation from satellites and ground observations over Europe during 1983–2010. Remote Sensing of Environment, 189, 108–117. https://doi.org/10.1016/J.ERS.2016.11.018

Sanchez-Lorenzo, A., Wild, M., & Trentmann, J. (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. https://doi.org/10.1016/J.RSE.2013.03.012

Schiller, R. A. & Rosow, W. B. (1985). ISCCP global radiance data set: A new resource for climate research. Bulletin of the American Meteorological Society, 66, 1498–1509. https://doi.org/10.1175/1520-0477(1985)066<1498:IGRDSA>2.0.CO;2

Schwandt, M., Chhathar, K., Meyer, R., Fross, K., Mitra, J., Vashistha, R., et al. (2014). Development and test of gap filling procedures for solar radiation data of the Indian SRRA measurement network. Energy Procedia, 57, 1100–1109. https://doi.org/10.1016/j.egypro.2014.10.096

Scudiero, E., Corwin, D. L., Morari, F., Anderson, R. G., & Skaggs, T. H. (2016). Spatial interpolation quality assessment for soil sensor transect datasets. Computers and Electronics in Agriculture, 123, 74–79. https://doi.org/10.1016/J.CEC.2016.02.016

Sellers, P. J., Rasool, S. I., & Bolle, H.-J. (1990). A review of satellite data algorithms for studies of the land surface. Bulletin of the American Meteorological Society, 71, 1429–1447. https://doi.org/10.1175/1520-0477(1990)71<1429:AROSDA>2.0.CO;2

Smith, S. J., van Aardenne, J., Klimont, Z., Andres, R. J., Volke, A., & Delgado Arias, S. (2011). Anthropogenic sulfur dioxide emissions: 1850–2005. Atmospheric Chemistry and Physics, 11, 1101–1116. https://doi.org/10.5194/acp-11-1101-2011

Stern, D. I. (2006). Reversal of the trend in global anthropogenic sulfur emissions. Global Environmental Change, 16, 207–220. https://doi.org/10.1016/J.GLOENVCHA.2006.01.001

Storelvmo, T., Heede, U. K., Leirvik, T., Phillips, P. C. B., Arndt, P., & Wild, M. (2018). Lethargic response to aerosol emissions in current climate models. Geophysical Research Letters, 45, 9814–9823. https://doi.org/10.1029/2018GL078298

Storelvmo, T., Leirvik, T., Lohmann, U., Phillips, P. C. B., & Wild, M. (2016). Disentangling greenhouse warming and aerosol cooling to reveal Earth’s climate sensitivity. Nature Geoscience, 9, 286–289. https://doi.org/10.1038/ngeo2670

Sun, H., Gui, D., Yan, B., Liu, Y., Liao, W., Zhu, Y., et al. (2016). Assessing the potential of random forest method for estimating solar radiation using air pollution index. Renewable and Sustainable Energy Reviews, 57, 1498–1505. https://doi.org/10.1016/J.RSER.2016.02.016

Tollenaar, M., Fridgen, J., Tyagi, P., Stackhouse Jr, P. W., & Kumudini, S. (2017). The contribution of solar brightening to the US maize yield trend. Nature Climate Change, 7, 275–278. https://doi.org/10.1038/nclimate3234

Van Rossum, G. and Drake, F. L. (1995). Python reference manual (p. 140). Centrum voor Wiskunde en Informatica Amsterdam.

Wang, S., Kinnison, D., Montzka, S. A., Apel, E. C., Horbrook, R. S., Hills, A. J., et al. (2019). Ocean biogeochemistry control on the marine emissions of brominated very short-lived ozone-depleting substances: A machine learning approach. Journal of Geophysical Research: Atmospheres, 124, 12319–12339. https://doi.org/10.1029/2019JD031288

Weiss, D. J., Atkinson, P. M., Bhatt, S., Mappin, R., Hay, S. I., & Gething, P. W. (2014). An effective approach for gap-filling continental scale remotely sensed time-series. ISPRS Journal of Photogrammetry and Remote Sensing, 90, 106–118. https://doi.org/10.1016/J.ISPRSJRPS.2014.10.001

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis (p. 260). Springer-Verlag New York. https://doi.org/10.1007/978-3-319-24277-4

Wild, M. (2016). Decadal changes in radiative fluxes at land and ocean surfaces and their relevance for global warming. Wiley Interdisciplinary Reviews: Climate Change, 7, 91–107. https://doi.org/10.1002/wcc.372

Wild, M., Ohmura, A., Schär, C., Müller, G., Folini, D., Schwierz, M., et al. (2017). The global energy balance archive (GEBA) version 2017: A database for worldwide measured surface energy fluxes. Earth System Science Data, 9, 601–613. https://doi.org/10.5194/essd-9-601-2017

Wood, S. N. (2003). Thin plate regression splines. Journal of the Royal Statistical Society Series B, 65(1), 95–114. https://doi.org/10.1111/1467-9868.00374

Wright, M. N. & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software, 77, 1–17. https://doi.org/10.18637/jss.v077.i01

Yang, Y., Xu, W., Hou, P., Liu, G., Liu, W., Wang, Y., et al. (2019). Improving maize grain yield by matching maize growth and solar radiation. Scientific Reports, 9, 3635. https://doi.org/10.1038/s41598-019-40081-z

Zhou, Q., Flores, A., Glenn, N. F., Walters, R., & Han, B. (2017). A machine learning approach to estimation of downward solar radiation from satellite-derived data products: An application over a semi-arid ecosystem in the U.S. PLoS ONE, 12, e0180239. https://doi.org/10.1371/journal.pone.0180239