A Machine learning technique for spatial interpolation of solar radiation observations

Thomas Leirvik[1] and Menghan Yuan[1]

[1]Nord University

November 21, 2022

Abstract

This paper applies statistical methods to interpolate missing values in a dataset 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) dataset. 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 neighboring variable is almost 40 times better than the best climatic variable in predicting SSR. Furthermore, our analysis shows that the Mean Absolute Error is 10.2 on average using RF, with a standard deviation of 1.5. Conventional methods have an average MAE of 21.3, with an average standard deviation of 6.4. This highlights the benefits of using machine learning in environmental research.

A Machine learning technique for spatial interpolation of solar radiation observations

Thomas Leirvik*[a] and Menghan Yuan†[a]

[a] Graduate School of Business, Nord University, 8049 Bodø, Norway

Abstract

This paper applies statistical methods to interpolate missing values in a dataset 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) dataset. 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 better than the best climatic variable in predicting SSR. Furthermore, our analysis shows that the Mean Absolute Error is 10.2 on average using RF, with a standard deviation of 1.5, which is significantly less than conventional methods; the average MAE of 21.3, more than twice the size of the RF method, and an average standard deviation of 6.4, more than four times larger than the RF standard deviation. This highlights the benefits of using machine learning in environmental research.

Keywords: Spatial interpolation; Surface Solar Radiation; Random Forest; Ordinary Kriging; Regression Kriging; Comparison

1 Introduction

Spatial data is the foundation for climate research (Zhu et al., 2016; Tao et al., 2018; Pfeifroth et al., 2018). In particular, downward shortwave radiation is a fundamental determinant of the Global Energy Balance, a crucial driving force for temperature change and hydrological cycles

*thomas.leirvik@nord.no

†Corresponding author: menghan.yuan@nord.no

variation (Budyko, 1969). Obryk et al. (2018) showed that a significant decrease of solar radiation in Antarctica has affected the polar desert processes substantially and that it is important to project the trend of solar radiation in order to determine the predictions of future desert processes. Furthermore, the impact of solar radiation on various aspects of a country's economy, for example the relationship to crop yields, is also highly important to secure food distribution both globally and locally, in particular as Earth's human population is expected to increase significantly over the next hundred years, see, for example, Yang et al. (2019); Thornton et al. (2009); Aaheim et al. (2015) and Tollenaar et al. (2017). As the population grows, urban planners rely on simulations for energy production and consumption when constructing future urban areas, see, for example Calcabrini et al. (2019). Therefore the importance of solar radiation on all aspects of the environment is difficult to underestimate. In this paper, we investigate the trends in solar radiation, both globally, and for each continent on Earth. We furthermore explore the best opportunities to improve a current dataset which has observations spanning several decades over hundreds of geographical areas on Earth. However, the physical stations collecting observations of climatological variables can have unexpected failures, and might not report values for a month. This leads to a variable with a probability of one, or many, so-called *missing values*. Ignoring, and deleting, missing values is one option. However, this approach reduces statistical power and increases estimation bias, see Nakagawa and Freckleton (2008). Thus, ignoring missing values is not acceptable when methods exists for dealing with this issue. Furthermore, due to strong seasonal, latitudinal, longitudinal, altitude, and other geographical dependencies of the observation stations, simple linear interpolation is suboptimal. For example, if the December value for solar radiation is missing in the far north of the Northern hemisphere, a simple average between November and January will not represent a good estimate of the true value. We propose to use sophisticated statistical tools, in particular a specific machine learning approach called Random Forest (RF), to deal with missing values. We show that RF is far better than any of the other, more conventional methods, often applied for interpolation. Our end result is a balanced panel dataset, i.e. information over time and space, with monthly values spanning almost five decades. The resulting dataset makes it possible to carry out a wide range of analyses on climate change, as solar radiation is shown to be a key driver of temperature, see Storelvmo et al. (2016) and Phillips et al. (2020).

The variability of solar radiation has been analyzed in many studies (Zhang et al., 2015; Sanchez-Lorenzo et al., 2017; Stephens et al., 2012; Wang et al., 2013; Wild, 2009), and is observed in the form of "global dimming and brightening" for a large-scale decrease and increase of solar radiation respectively. A significant dimming was observed between 1950s to 1980s (Wild, 2009; Stanhill and Cohen, 2001; Zhang et al., 2015), followed by general brightening since the mid-1980s, especially in developed regions such as Europe (Wild, 2009). In particular, Sanchez-Lorenzo et al. (2017) found that there had been a mean increase of SSR of at least 2 Wm^{-2} per decade from 1983 to 2010 over European land areas. Our imputed dataset exhibits strong non-linear trends in the SSR, corroborating earlier results on the matter, see, for example, Liepert (2002) and Stanhill and

Cohen (2001). However, we find that these trends vary significantly geographically, whereas some similarities are present: for our time-range there is at first a strong decrease in SSR from 1964 to 1990s, followed by a reversal, broadly consistent with the periods for which global dimming and brightening have been reported in the literature. We further show that the trends, and changes in trends, varies significantly continent by continent.

Solar radiation has been a key input in many climate studies, for example related to climate sensitivity, its effect on agriculture, and other economic sectors. Based on a climate econometric model that uses surface radiation as one of the inputs, Phillips et al. (2020) estimated the transient climate response, which is the change in global mean surface temperature for a doubling of CO_2. Storelvmo et al. (2016) decomposed observed temperature development into components attributable to changes in greenhouse gas concentrations (CO_2) and surface radiation. The increase in greenhouse gas pulls the temperature up, while a reduction in downward solar radiation drags the temperature down. We construct a more complete dataset than has been available previously, which simplifies analysis of trends in SSR and its impact on climate, environment, and the economies around the world.

To study the importance and impact of solar radiation on climate, environment, and other areas, it is of utmost importance to have reliable data on solar radiation, which can be used for analyzing a wide range of global, and local, aspects of the environment. Unfortunately, there are few, if any, complete datasets for solar radiation over long time periods for many locations. One such dataset is the Global Energy Balance Archive (GEBA). The GEBA dataset has a long time range, with the first observations from the early 1950's, continuing to 2013. However, many values are missing due to maintenance and operational failures of observational devices. To fill missing values, many different techniques have been tested, with linear interpolation as one of the simplest. This, however, does not perform very well if the data is highly spatially correlated, and moreover if the data is seasonal, which is certainly true for solar radiation. Other, more advanced, techniques used by researchers are, for example, Inverse Distance Weighting (IDW), Kriging, splining, regression and etc. (Collins, 1995; Scudiero et al., 2016; Erxleben et al., 2002). We will apply advanced statistical methods, often called machine learning techniques, together with conventional spatial interpolation methods, to simulate values in a solar radiation dataset. We compare all methods described above and find that machine learning methods are by far the best choice for the kind of dataset we analyze in this paper. The methods and our contribution is explained in the next paragraph.

IDW is a deterministic estimation method where the value at a missing location is determined by weighted average of values at neighboring points, where the weights are assigned as inversely proportional to the distance from the target location. Although its simplicity makes it easy to implement, a drawback is that it will have a discontinuous slope of the estimation surface at each data point (Collins, 1995). The Kriging method is an interpolation technique where the interpolated values are modeled by a more sophisticated relationship between observed values with target missing values. The relationship is depicted by a variogram function, which calculates the semi-variance

differences (spatial autocorrelation) between the neighboring values. The variogram function can be exponential, spherical, logarithmic or any other function. Spline models, a class of functions often called polynomial interpolation models, use mathematical functions to connect the sampled data points in order to produce a continuous elevation and grade surface while minimizing the curvature of the surface. It is useful when the surface varies smoothly without sharp fluctuation. The models mentioned above are all univariate, which means that they apply only information about the variable of interest. Regression methods break this limitation by characterizing the response's relationship with other potential explanatory variables, such as the location (latitude, longitude) of the measurement station, and other correlated meteorologic variables. However, it is difficult for regression models to take into account spatial autocorrelation between nearby data points. This deficiency leads to the emergence of Regression Kriging (RK) which combines regression predicted trends with Kriging fitted residuals, generating a prediction that brings together the merits of the two methods.

Recently, machine learning methods have seen an increasing number of applications in environmental research, see, for example Jiang (2008); Sun et al. (2016); Zhou et al. (2017). 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. An effective application in predicting evapotranspiration is established by Xu et al. (2018), who upscaled site observations to a regional grid scale over a continuous time period. Among the five machine learning methods (including artificial neural network, Cubist, deep belief network, random forest, and support vector machines) applied in the paper, all have almost identical performance quantified by root-mean-square error, however, the random forest shows lower uncertainty than other methods based on the three-corned hat algorithm, which measures the predictive stability by capturing the variability for the prediction against the reference. More applications can be found in predicting storm movement trend and growth (Han et al., 2017), halogenated substances (Wang et al., 2019), solar radiation (Jiang, 2008), etc. Machine learning approaches are also commonly applied in pattern detection and factor quantification. Stirnberg et al. (2020) identified the contributors in the levels of particles in the air using a machine learning method, which is shown to be capable of reproducing concentrations of particles with high accuracy and reliability, and important drivers of the series are identified through the process.

Among a variety of machine learning models, the Random Forest (RF) model provides a distinctive solution for accommodating a high dimension of the covariates; specifically, it is effective, straightforward, and relatively less computationally intensive (Grabska et al., 2020). This makes it practical, and feasible, to implement and analyze large datasets, which can be difficult for convergence in other machine learning algorithms. For example, Sun et al. (2016) presented a RF model to estimate solar radiation based on three types of input variables, sunshine hours, air temperature, and their derivative types. Surface albedo, emissivity and vegetation indices data were used as predictors by Zhou et al. (2017) in a RF model in order to generate an accurate prediction of solar

radiation. Most of the research on spatial interpolation are on regional scale and/or only consider a relatively short period of time for testing model quality, such as a one year period. The main contribution of this paper is to apply an effective machine learning method in spatial interpolation of Surface Solar Radiation (SSR), that is applicable for the global scale over a long time-span. Moreover, spatial and temporal dependency is incorporated in the model by constructing a group of neighboring variables and time series variable.

In this paper, we apply eight models, including random forest, ordinary Kriging, three regression models, and their respective Regression Kriging methods, to predict SSR values in the GEBA dataset and evaluate their performance. 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 characteristic of the response. Specifically, three groups of predictors are constructed: *i*) climatic and geographical variables, such as cloud coverage, temperature, etc., to describe station climatic status; *ii*) spatial neighboring variables to reflect spatial autocorrelation of SSR in the vicinity of the station; and *iii*) time series variables, which are variants of lagged SSR, to account for seasonality and temporal autocorrelation. To our best knowledge, the first group of predictors are commonly used, the latter two groups are rarely contemporarily included as a part of a regressor ensemble. Nevertheless, we show that groups *ii*) and *iii*) have significant explanatory power for predicting SSR. Furthermore, an analysis of 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. Intuitively, neighboring variables and time series variables provide direct information about the response, whereas the climatic variables only function as assistant predictors, which serves as drivers for the variation in the response. It is not surprising that direct information is the best predictor among all, though a limit of neighboring and time series variables is that these variables require high data completeness, which could be demanding in practice.

A 10-fold cross validation for each continent is implemented to assess model performance. The results shows that (1) the RF generates the smallest prediction errors and is least affected by observation density. The relative density independence of RF is an essential property to ensure a high prediction accuracy even in a setting with scarce observations. (2) Ordinary Kriging's performance is strongly determined by a high 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 prediction. (3) Regression Kriging is more preferred than just having the OK or regression models alone, however the accuracy improvement is not considerable. On the contrary, the improvement of RF is quite substantial. The RF model has a 44% lower MAE than the best regression Kriging 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 rest of the paper is organized as follows: Section 2 describes the study data and methods. Section 3 gives an assessment of model performance in prediction. Section 4 investigates SSR trends both globally as well as for each continent. The study is discussed and concluded in Section 5. Finally, technical notes are addressed in Section 6.

2 Data and Method

This section describes the data source we used in the empirical study, as well as the spatial interpolation methods which we apply to the SSR estimation.

2.1 Climatic Datasets

A. The GEBA global radiation dataset

The Surface Solar Radiation (SSR, measured in Wm^{-2}) used in this study is provided by the Global Energy Balance Archive (GEBA), which is a dataset maintained by the Institute for Climate and Atmospheric Science at ETH Zurich. It stores energy fluxes at the surface, mainly SSR measurements (Sanchez-Lorenzo et al., 2017). 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 large 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.

B. The CRU-TS climate dataset

The CRU-TS climate dataset comprises monthly time series of meteorological variables that might be used as predictors for SSR prediction. Data is available for all global land areas, excluding Antarctica, and can be downloaded from the Climate Research Unit (CRU) website. It has a spatial resolution of 0.5×0.5 degree 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:** vapour pressure in hPa
- **wet:** number of rain days

C. The GRUMP station location dataset

The Global Rural-Urban Mapping Project (GRUMP) dataset 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 dataset are classified as rural or urban based on the value of the grid cell containing the location of each climate station. The variables from GRUMP are shown in the following list:

- **LAT:** latitude of station
- **LON:** longitude of station
- **ALT:** altitude of station
- **URB:** whether station is in an urban/rural area

The three datasets are joined together to provide a climatic and geographical description for each climate station in the GEBA dataset.

2.2 Random Forest Model

The Random Forest 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*. In our study, we include climatic and geographical variables encompassing the following indicators: *year*, *month*, the $CRU - TS$ and the $GRUMP$ variables. Spatial neighboring variables as a group describes the spatial distribution of SSR in the neighborhood, with respects to its level, spread and tendency, which will be elaborated in detail below. 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 data from nearby areas provided that these sections share similar spatial dynamics with the target location (Ohashi and 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 paper, 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, since 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 dataset $Z = \{z_1, z_2, \cdots, z_n\}$, where z_i is the value of radiation z at location i. Let $N_o^k = \{z_i \in Z : d(o,i) \leq d(o,j_k), \forall i \neq o\}$ be the neighborhood of station o, where $d(o,i)$ indicates the distance between station o and station i, $d(o,j_k)$ denotes the distance between station o and its kth-nearest station j_k.

We use the notation $\overline{z}(N_o^k)$ for the average of z_i in the neighborhood N_o^k, and $\tilde{z}(N_o^k)$ for the weighted average counterpart, based on the Inverse Distance Weighting (IDW). The weighted average is then given by:

$$\begin{aligned} \tilde{z}(N_o^k) &= \sum_{z_i \in N_o^K} \lambda_i z_i \\ \lambda_i &= \frac{d_{io}^{-r}}{\sum_{z_i \in N_o^k} d_{io}^{-r}} \end{aligned} \tag{1}$$

This means that λ_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 (Zaki et al., 2019) states an exponent of 5 outperforms an exponent value of 2. We thus take $r = 5$ in this study. For the arithmetic spatial average, $\overline{z}(N_o^k)$, 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_o^k. Figure 1 illustrates an example for the 2-nearest neighborhood, N_o^2. 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

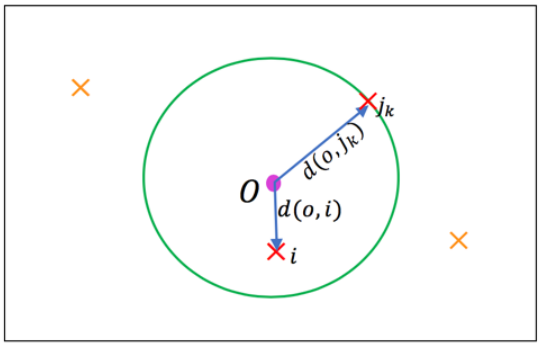

Fig. 1 Neighborhood illustration. o is the target station, × indicates a station within the boundary and × a station outside.

neighborhood, given by

$$\sigma_z(N_o^k) = \sqrt{\frac{1}{|N_o^k|} \sum_{z_i \in N_o^k} (z_i - \overline{z}(N_o^k))^2} \tag{2}$$

$\sigma_z(N_o^k)$ measures the volatility of SSR within the neighborhood N_o^k. 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,

$$\overline{Z}_o^{k_1,k_2} = \frac{\overline{z}(N_o^{k_1})}{\overline{z}(N_o^{k_2})} \tag{3}$$

where k_1 and k_2 are two neighborhood sizes ($k_1 < k_2$) and $\overline{z}()$ is the average of a set of the values in the neighborhood of o. This means that $\overline{Z}_o^{k_1,k_2}$ 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 towards, or away, from the ocean, closer to densely populated areas, etc. A variation of this indicator is the weighted average counterpart given by

$$\tilde{Z}_o^{k_1,k_2} = \frac{\tilde{z}(N_o^{k_1})}{\tilde{z}(N_o^{k_2})} \tag{4}$$

where $\tilde{z}()$ 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, $\overline{z}(N_o^{k_1})$, $\overline{z}(N_o^{k_2})$, $\overline{z}(N_o^{k_3})$, $\overline{Z}_o^{k_1,k_2}$, $\overline{Z}_o^{k_2,k_3}$, $\tilde{z}(N_o^{k_1})$, $\tilde{z}(N_o^{k_2})$, $\tilde{z}(N_o^{k_3})$, $\tilde{Z}_o^{k_1,k_2}$, $\tilde{Z}_o^{k_2,k_3}$, $\sigma_z(N_o^{k_1})$, $\sigma_z(N_o^{k_2})$, $\sigma_z(N_o^{k_3})$, 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, i.e. $RAD_{t-1}, RAD_{t-2}, RAD_{t-3}, \ldots$. Meteorological variables are likely to show the propensity of seasonality that previous years' radia-

tion in the same month has significant correlation with that of the current month. To account for this, we introduce annual lagged values. i.e $RAD_{t-12}, RAD_{t-24}, RAD_{t-36}, \ldots$, where for example RAD_{t-12} is the observed RAD twelve 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 RAD_{mamk} representing the moving average of the previous k months' radiation and RAD_{mayK} for the counterpart of the yearly lagged radiation with K annual lags.

2.3 Ordinary Kriging

Ordinary Kriging uses 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

$$\gamma(h) = \frac{1}{2} average \left[(z(x_i) - z(x_j))^2 \right] \tag{5}$$

where h is the distance between points x_i and x_j, and $z(x_i)$ 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 autocorrealation 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 $Var(z^*(x_o) - z(x_o))$. The weights can be used to make a prediction by

$$z^*(x_o) = \sum_{i=1}^{n} \omega_i^{OK} z(x_i) \tag{6}$$

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

2.4 Regression Kriging

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. Here we have a RK model given by

$$\begin{aligned}\hat{z}(x_o) &= \hat{m}(x_o) + \hat{e}(x_o) \\ &= \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(x_o) + \sum_{i=1}^{n} \hat{\omega}_i \cdot e(x_i)\end{aligned} \tag{7}$$

where $\hat{m}(x_o)$ is the fitted trend, $\hat{e}(x_o)$ is the interpolated residual, $\hat{\beta}_k, k = 0, \ldots, p$, is the estimated regression coefficients, p is the number of auxiliary variables, $q_0(x_o) = 1$ is the constant term, and $\hat{\omega}_i, i = 1, \ldots, n$, are the Kriging weights.

In the RK, we use the group of *climatic and geographical variables* as regressors, and apply three forms of regression, i.e. simple linear regression (LR), generalized additive regression (GAM) and least squares dummy variable (LSDV) model. Figure 2 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 a GAM model. Another way of capturing the various levels of SSR is to introducing eleven new dummy variables to represent the twelve months, when the eleven dummy variables all equal to zero, it indicates the reference month.

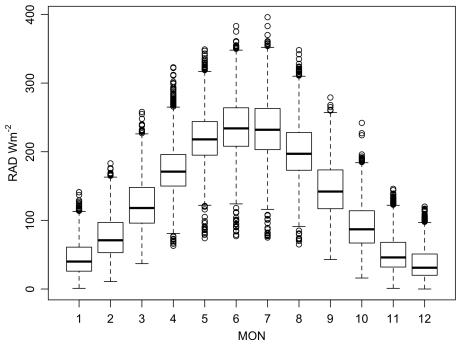

Fig. 2 Box plot for each month. It plots the monthly SSR distribution for Europe.

3 Empirical Results

We start out with a dataset of 1523 stations scattered around land areas on Earth. Figure 3 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 sparsely, though even, distribution, while in South America and Australia stations are more located at the borders of the continents. There are also a few stations scattered on islands across the ocean. Each station registers the average downward shortwave solar radiation each month during the year. The first stations were starting to operate in the early 1950's. 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 paper is to estimate those missing values using advanced statistical methods. We implement a data quality control process that prepares the dataset for the input of Random Forest. The following stations are dropped in the dataset: 1) stations that exist for less than three years; 2) stations that don't 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 1227 stations, see Figure 4 for the number of observations every year.

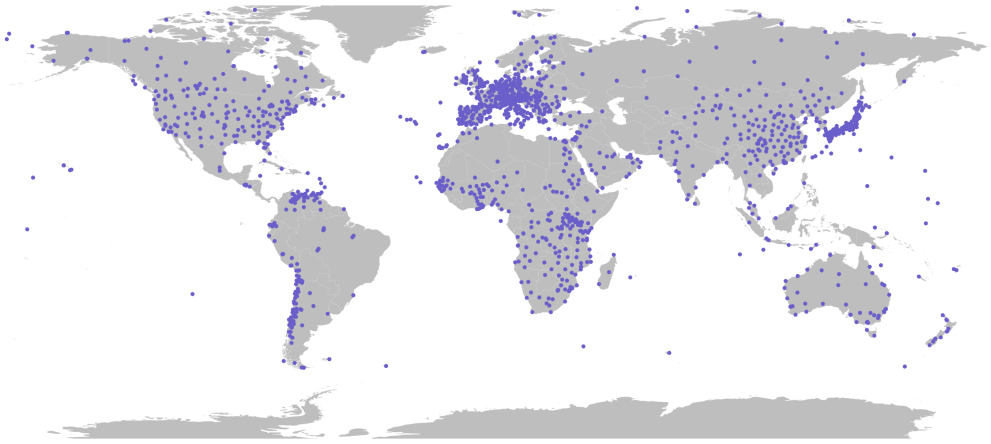

Fig. 3 Locations of all stations contributing to the GEBA dataset

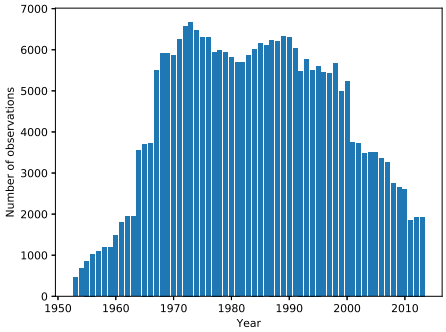

Fig. 4 Number of observations every year

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).

$$
\begin{aligned}
R^2 &= 1 - \frac{\sum_{i=1}^{n}(\hat{z}_i - z_i)^2}{\sum_{i=1}^{n}(z_i - \bar{z}_i)^2} \\
MAE &= \frac{1}{n}\sum_{i=1}^{n}|\hat{z}_i - z_i| \\
MAPE &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{z}_i - z_i}{z_i}\right| \\
RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{z}_i - z_i)^2}
\end{aligned}
\tag{8}
$$

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

A. 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 cutoff distance that intends to avoid bias from outlier stations. Figure 5 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. In summary, we set $k_1 = 2, k_2 = 3, k_3 = 5$, with cutoff distances being $d_1 = 691km, d_2 = 790km, d_3 = 1047km$ respectively.

B. Order of seasonal autoregression

To see the characteristics of radiation evolution in the past, Figure 6 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 7 plots the AutoCorrelation Function (ACF) and Partial AutoCorrelation Function (PACF) of the radiation time series. Both figures show significant annual

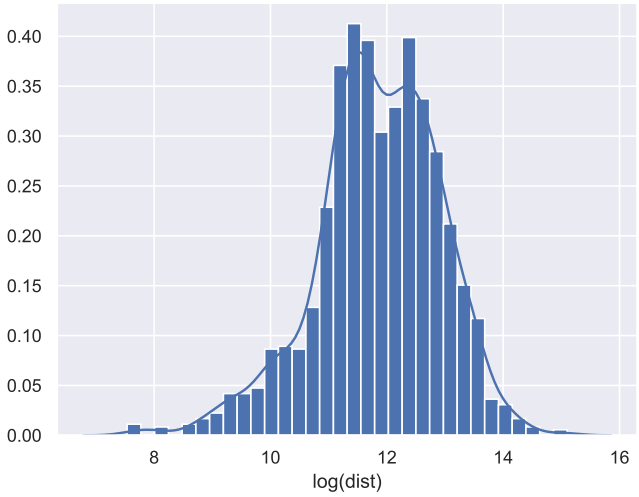

Fig. 5 Distribution of log distance between the second nearest stations, measured in km

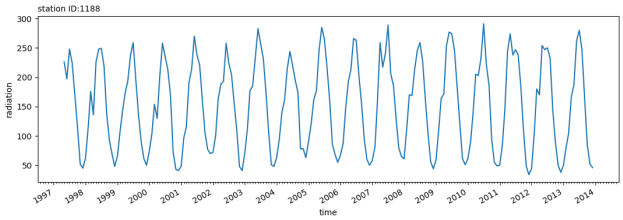

Fig. 6 Radiation time series of a random station, in Wm^{-2}

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 $RAD_{t-K\times s}$ seasonal lagged terms as predictors for RAD_t, where $s = 12$ is the seasonal length, and $K = 1, 2, \ldots, P$, where P is the optimal order of seasonal lagged terms.

C. Order of autoregression

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

D. 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

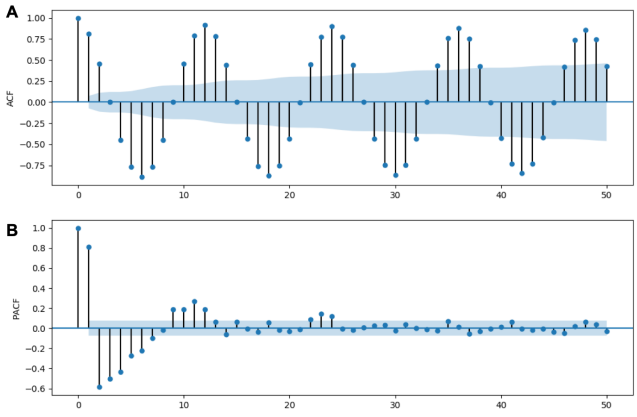

Fig. 7 ACF and PACF of the sample Radiation time series

average of radiation:

$$RAD_{t\_mayK} = \frac{\sum_{i=1}^{K} RAD_{t-i\times s}}{K} \tag{9}$$

where $K = 2, 3, \ldots$, and s is the seasonal length equal to 12. An optimal number K of observation could to be found by calculating the MAE for n arguments and choosing the one with the minimum MAE.

E. Moving average

A smoothed trend of past few months radiation is represented as the average of the last k terms, i.e., 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:

$$RAD_{t\_mamk} = \frac{\sum_{i=1}^{k} RAD_{t-i}}{k} \tag{10}$$

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.

F. 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 Random Forest 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: $p = 1$, $k = 3$, $P = 3$, $K = 3$.

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, i.e., 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 with the two model parameters can be found in the Supplement Material S1.

Figure 8 plots the natural logarithm of permutation variable importance for the first ten most important variables in the Random Forest 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 result also highlights the importance of using a sophisticated statistical model which allows for other types of explanatory variables. The time series variables (labeled in magenta), especially the three-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 differing, or even opposing, climatic and geographical characteristics, we implement model training and testing on a continental level. A 10-fold Cross Validation (CV) is implemented for the RF. For regression Kriging, 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

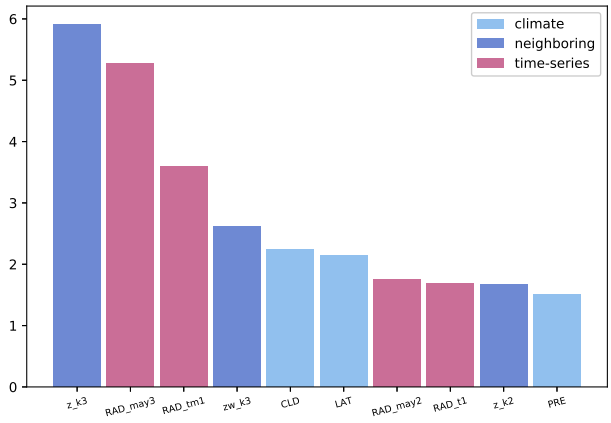

Fig. 8 Random Forest permutation logarithm scaled variable importance. Note that the importance is measured in natural logarithm scale, actual importance could be obtained by taking its natural exponent. z_k3 is the simple average of radiation for its nearest five neighbors.

trends. On the other hand, Ordinary Kriging 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 Ordinary Kriging by a monthly 10-fold CV, and generating SSR residual predictions. By adding regression trends and Kriging residuals together, we eventually obtain our final predictions.

Table 1 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 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 and a standard deviation of 1.5, while the ensemble of the other seven methods generates an average MAE of 21.3, and an average of standard deviation of 6.4. 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. In particular, 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 become the good-performed models under MAPE. One possible explanation could be that the percentage error tend 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 high levels of SSR in Africa, therefore showing a false high accuracy. Consequently, the abso-

lute error measures, i.e., MAE and RMSE, are more appropriate indicators for model evaluation regarding SSR in our case. It is worth noticing that another preferable advantage for the RF is that observation density is not as influential on model performance as the other models. We see from the station distribution map that South America has the most clustered and uneven station placement, and we hence observe a significant decrease in performance in the continent for the conventional regression Kriging methods, whereas the error measures under the RF model are not deteriorating as much, compared to other continents. The relative stability of RF prediction errors indicates 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 striking larger errors for LR indicate that regarding MON as a regular numeric variable is not adequate, 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 effectively in South America. The outperformance of OK is nonetheless noticeable in Europe, which indicates that the performance of the OK model is highly determined by observation density. The regression Kriging methods incorporate spatial dependence by adding a spatial correlated residual term in the trend prediction. The effectiveness of this method is confirmed by its lower error metrics. In Figure 9, predicted values in Europe are plotted against observed values for the four best performance models. We observe that all coefficient of determination (R squared) 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 the most stable predictions.

Figure 10 shows monthly anomalies for one chosen station, located at Locarno-Monti (46.17N, 8.78E) 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 11, monthly anomalies are plotted for Europe (See Supplement Material S2 for monthly anomalies for the other continents). The aggregated predictions for all the three models replicate trends of observations reasonably well. By comparing Figure 10 and Figure 11, 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).

Table 1 Evaluation metrics. Numbers in bold face represent to best fit. We see that the Random forest methods is better for all continents, for all performance metrics, whereas linear regression performs the worst.

		Models							
Measure	Continent	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
	Std. dev	**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	12.44	12.46
	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	25.20	15.15	15.15	13.34	15.86	12.64	12.65
	Std. dev	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
	Std. dev	1.57	2.86	8.78	8.78	8.58	5.95	8.52	8.52
R^2	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	0.90	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
	Std. dev	0.04	0.21	0.28	0.28	0.27	0.23	0.25	0.25

[a] MAE and RMSE are measured in Wm^{-2}; [b] MAPE is measured in %; [c] R^2 is the coefficient of determination by regres
on observed values;

4 Trends in Global and Continental Average Surface Solar Radiation

In this section, annual global and continental average trends for SSR are explored. Figure 12 shows the cumulative annual global average radiation change from 1964 to 2013. 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

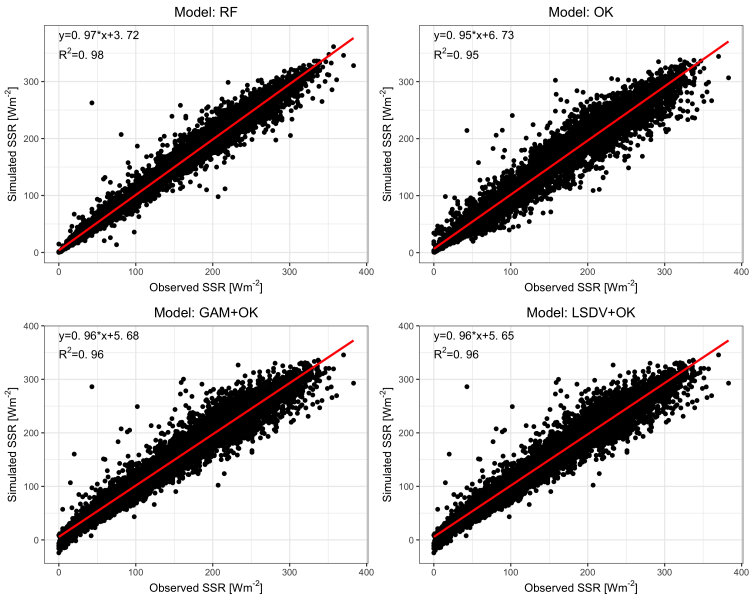

Fig. 9 Simulation against observation values in Europe. The figure illustrates that the RF model produces simulated values with a lower variability than the other models. See the Supplement Material for corresponding figures for the other continents.

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 afterwards. The exact correlation between SSR and sulfur emissions is examined further in what follows.

Annual cumulative continental average change is shown in Figure 13 and Figure 14 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, it seems that SSR enters an accelerating increase in the 10-15-year 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 trend ceases regardless.

The trends in Asia are fairly alike that of North America, the distinct decrease prevails until the beginning of 1990s and SSR fluctuates around a zero-trend afterwards. The remarkable decrease is mostly attributed to the increase in aerosol loading caused by rapid economic expansion and

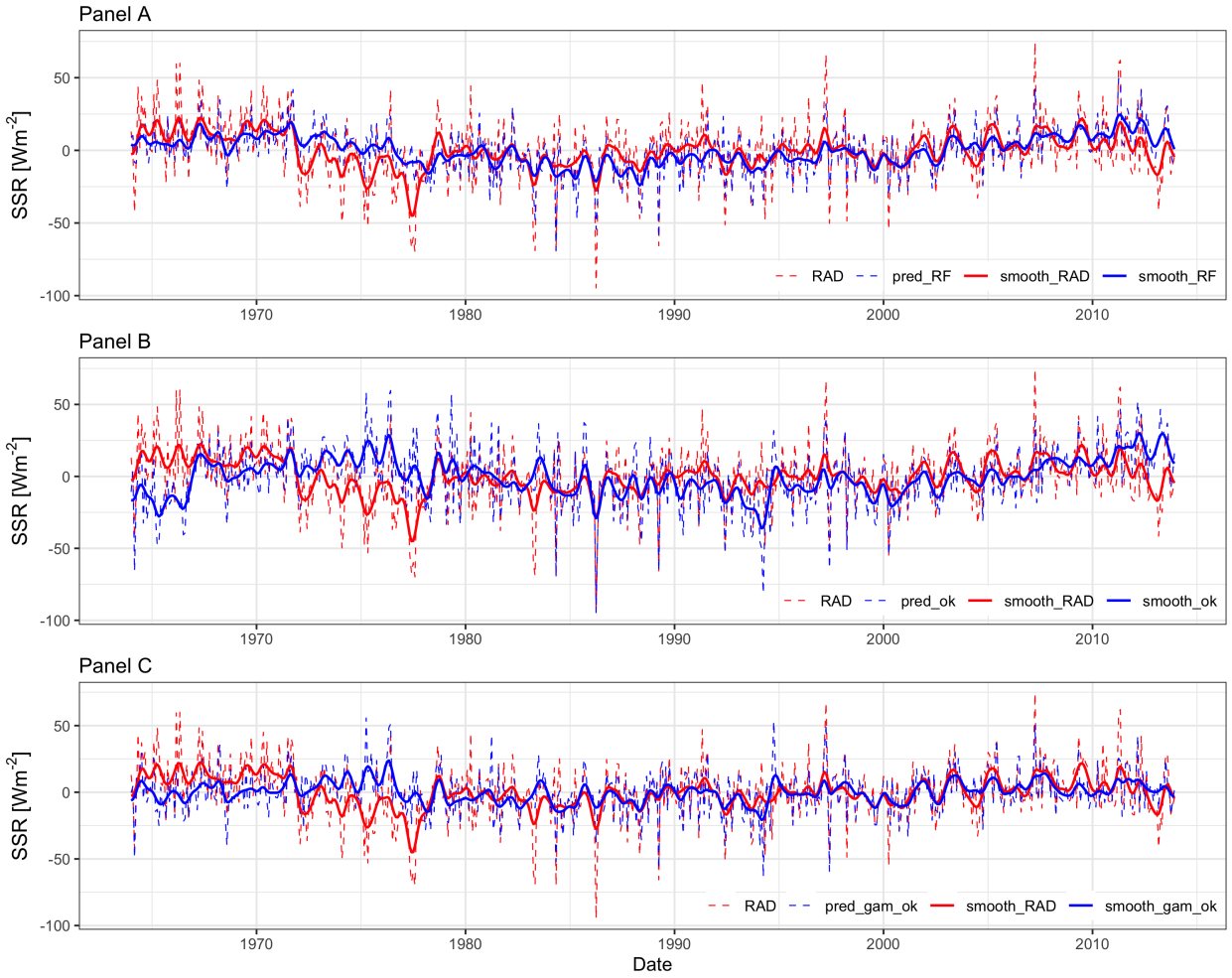

Fig. 10 SSR monthly anomalies for one chosen station. Each panel encompasses monthly SSR anomalies (dashed line) of observations (red) and model simulations (blue), together with their corresponding smoothed series (solid line) by a 12-month Gaussian kernel. The series are expressed as anomalies from the 1964-2013 mean. Corresponding prediction models for each panel: **Panel A** - RF, **Panel B** - OK, **Panel C** - GAM+OK. Since LSDV+Ok prediction has very similar trends with GAM+OK, see Supplement Material Figure S4 for details.

population growth during that period (Liang and Xia, 2005). At the beginning of the 1990s, when major economic crises occurred in Asia, the decreasing trend stops persisting and SSR fluctuates with no evident trend.

As atmospheric sulfur aerosol loading is known to have 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. (Storelvmo et al., 2016), and thereby attenuate SSR and cause global dimming. An inspection of the relationship between SSR and SO_2, a precursor of sulfate aerosol, is discussed thereafter. Figure 15 illustrates the relationship between SSR and SO_2 emissions, see also (Smith et al., 2011; Klimont et al., 2013) for a thorough treatment of the matter. Significant correlation coefficients are observed as -0.601 for the period

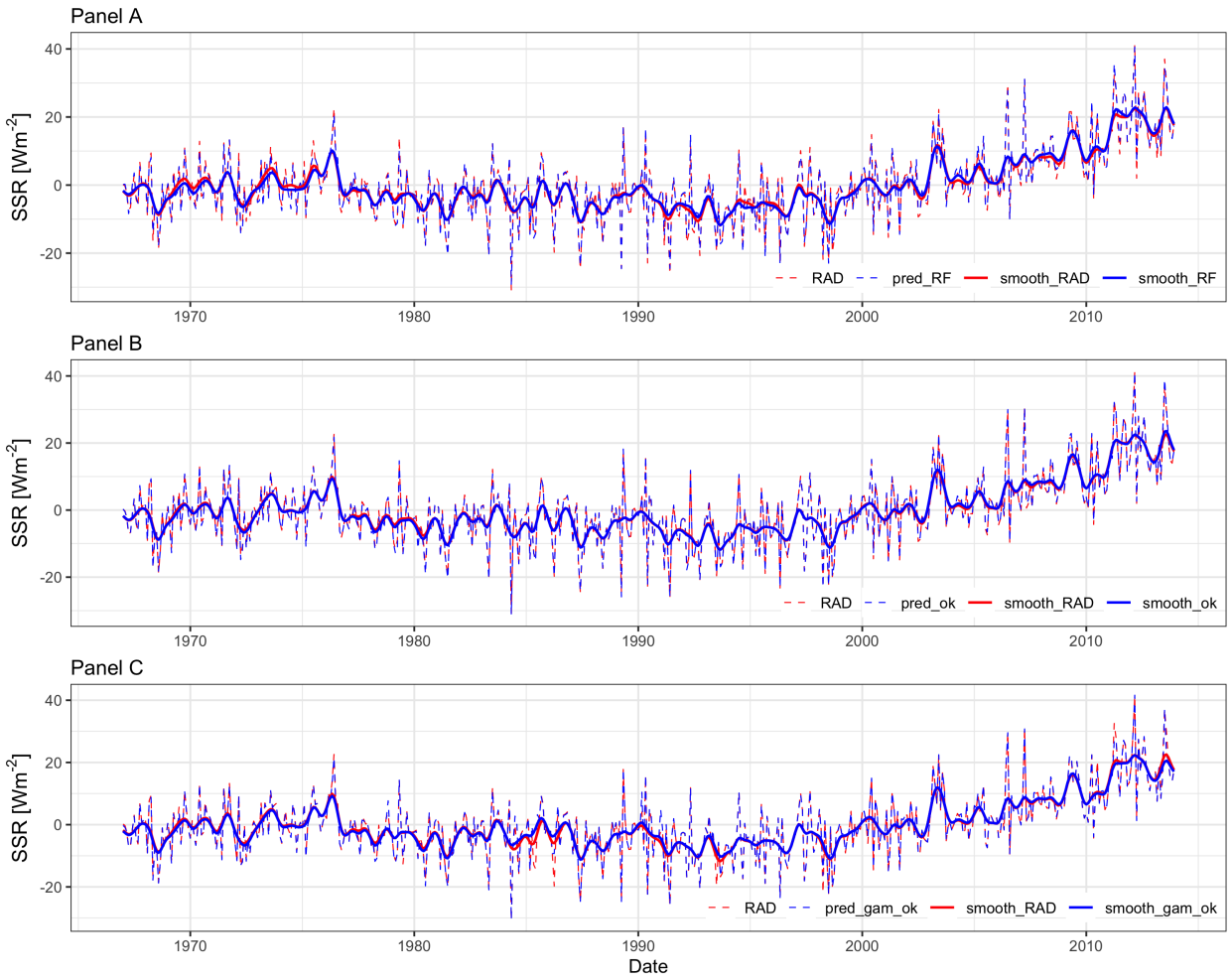

Fig. 11 SSR monthly anomalies for Europe. Each panel encompasses monthly SSR anomalies (dashed line) of observations (red) and model simulations (blue), together with their corresponding smoothed series (solid line) by a 12-month Gaussian kernel. The series are expressed as anomalies from the 1967-2013 mean. Corresponding prediction models for each panel: **Panel A** - RF, **Panel B** - OK, **Panel C** - GAM+OK. Since LSDV+OK prediction has very similar trends with GAM+OK, see Supplement Material Figure S5 for details.

1964-1985, and -0.607 for 1986-2011, which are in agreement with the result shown in Storelvmo et al. (2018). A similar analysis over Europe is shown in Figure 16. A jump at 2000 is noticed due to change of dataset. 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 hard to reconcile for the two datasets, as some countries locate transcontinentally, for example Russia. This aggregation problem does not exist on a global level, thereby we observe a rather smooth curve in Figure 15. To reduce bias from the jump, correlation coefficients are calculated for sub periods split by 2000. From 1964 to 1999, correlation coefficient is 0.137, and insignificant, in contrast with a very significant correlation coefficient of -0.971 for 2000-2011.

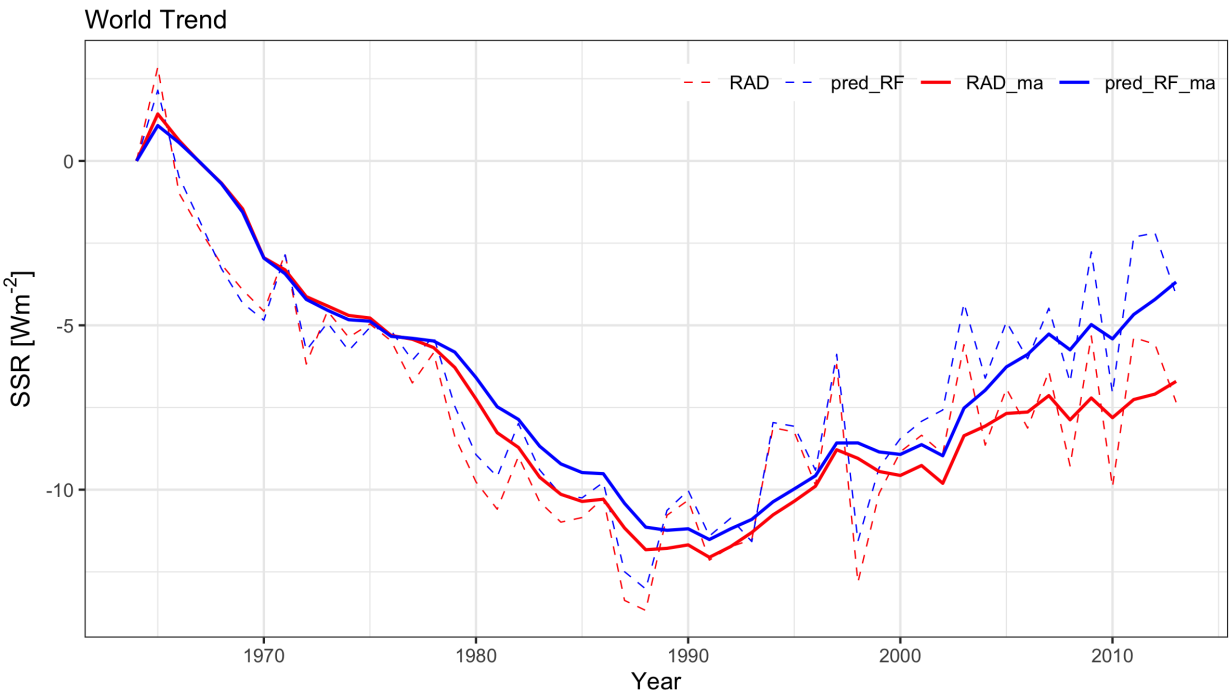

Fig. 12 Cumulative annual global average trend for SSR from 1964 to 2013 for observations (rad) and RF predictions (blue), plotted together with five-year moving average (solid line). The series are expressed as cumulative change from 1964.

5 Conclusion

This paper explores the potential of applying a machine learning approach, Random Forest, to spatial interpolation. A wide range of explanatory variables is explored in the interpolation, including climatic and geographical variables, spatial neighboring variables, and time series variables, among which spatial neighboring variables are proved to be the determinant factor for the RF. RF shows the highest predictive accuracy among all eight implemented models in the study, in addition to that, its performance is less contingent on data density, which makes RF an preferable alternative in spatial interpolation applications. 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 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 investigate the global SSR trend and find that Earth first experienced global dimming from 1964 to 1990, and then a significant trend reversal or recovery until 2013. Each continents' SSR is analyzed and a significant geographical discrepancy is observed. The imputed SSR dataset could be applied in further climate research. One of the applications could be a comparison of observed

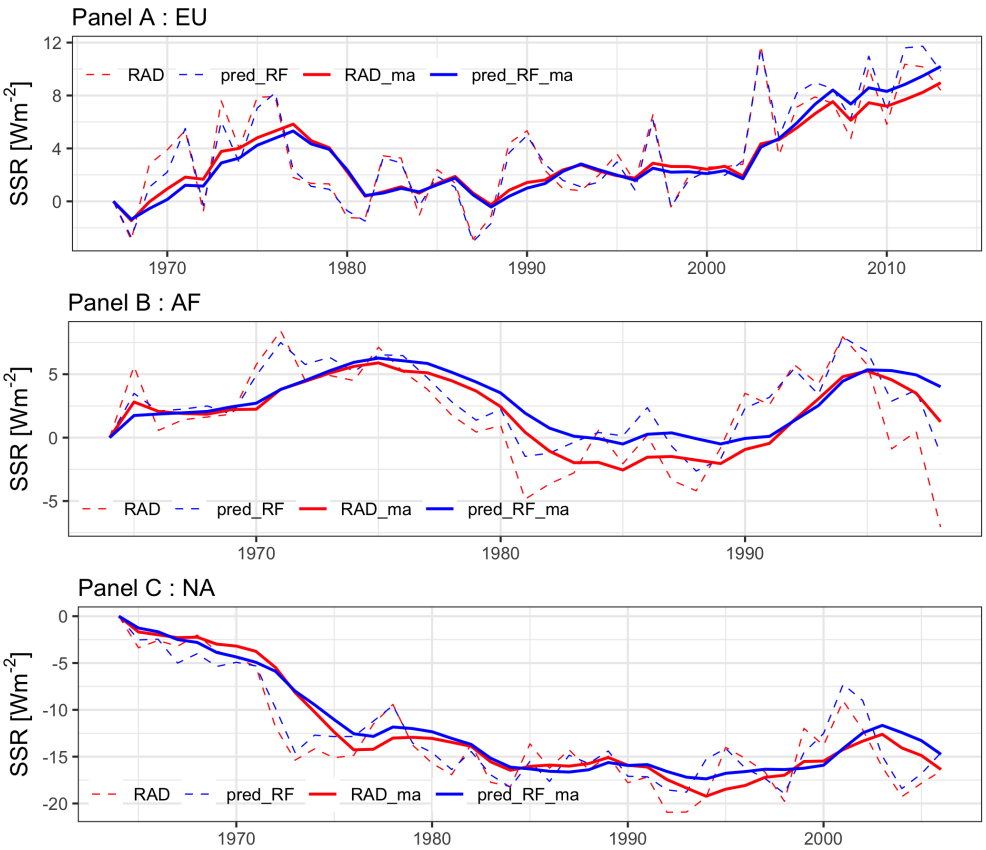

Fig. 13 Cumulative annual continent average SSR trend - part I. Cumulative annual continent average trend for SSR for observations (rad) and RF predictions (blue), plotted together with five-year moving average (solid line). The series are plotted for various time periods for each continent such that a valid number of stations exist within the time (referring to Supplementary Material for each continent's recorded number of stations). **Panel A** for Europe (1967-2013), **Panel B** for Africa (1964-1998), **Panel C** for North America (1964-2006).

SSR with simulated SSR from Global Climate Models.

6 Software and technical notes

The modelling is implemented in the statistical environment R (R Core Team, 2018) and Python (Van Rossum and Drake Jr, 1995). We apply the randomForest (Wright and Ziegler, 2017) to train the random forest model. GAM model is estimated using package mgcv (Wood, 2003). Ordinary Kriging is conducted with geostatistical packages sp (Pebesma and 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`.

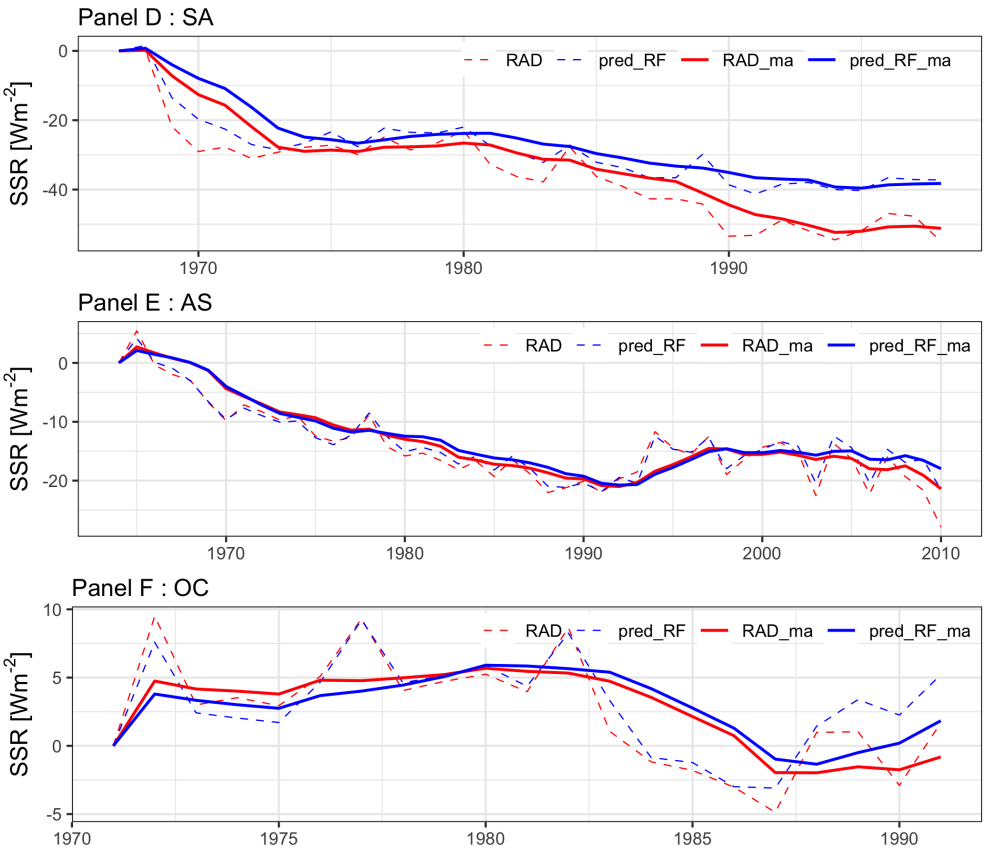

Fig. 14 Cumulative annual continent average SSR trend - part II. Cumulative annual continent average trend for SSR for observations (rad) and RF predictions (blue), plotted together with five-year moving average (solid line). The series are plotted for various time periods for each continent such that a valid number of stations exist within the time (referring to Supplementary Material for each continent's recorded number of stations). **Panel D** for South America (1967-1998), **Panel E** for Asia (1964-2010), **Panel F** for Oceania (1971-1991).

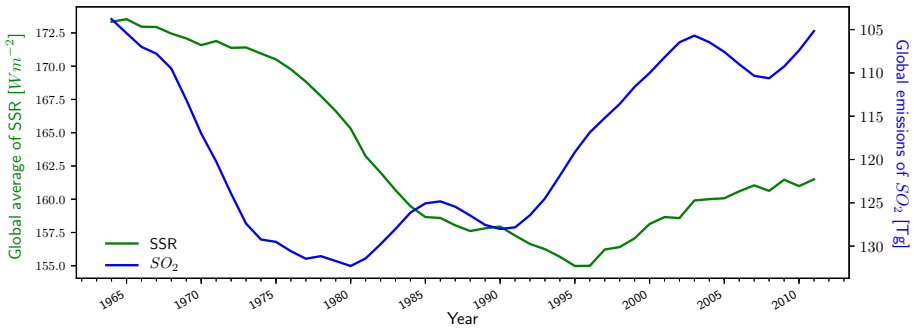

Fig. 15 Annually global average of SSR (green line, left axis) and global emissions of SO_2 (blue line, right axis). Both curves show 5-year running averages. SO_2 data are from ScocioEconomic Data and Applications Center (SEDAC).

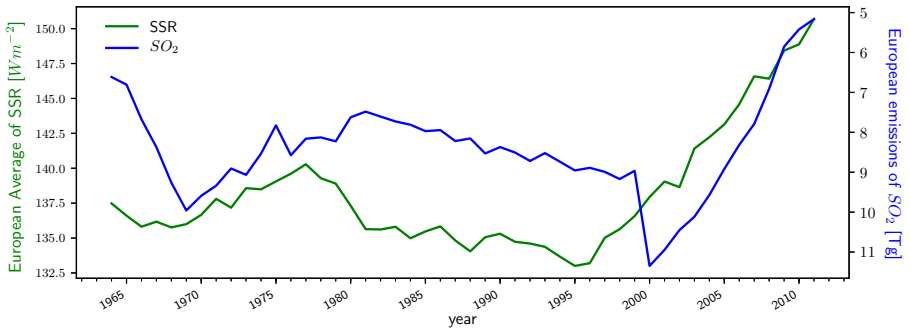

Fig. 16 Annually European average of SSR (green line, left axis) and European emissions of SO_2 (blue line, right axis). Both curves show 5-year running means. SO_2 data are from SEDAC.

Acknowledgements

This research was funded by Norwegian Research Council (grant number: 281071), under the project of "Climate Change Modelling and Prediction of Economic Impact".

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

References

Aaheim, A., Romstad, B., Wei, T., Kristjánsson, J. E., Muri, H., Niemeier, U., and Schmidt, H. (2015). An economic evaluation of solar radiation management. *Science of The Total Environment*, 532:61 – 69.

Breiman, L. (2001). Random forests. Machine Learning. 45:5–32.

Budyko, M. I. (1969). The effect of solar radiation variations on the climate of the Earth. *Tellus*, 21(5):611–619.

Calcabrini, A., Ziar, H., Isabella, O., and Zeman, M. (2019). A simplified skyline-based method for estimating the annual solar energy potential in urban environments. *Nature Energy*, 4(3):206–215.

Collins, F. C. (1995). *A comparison of spatial interpolation techniques in temperature estimation*. PhD thesis, Virginia Tech.

Erxleben, J., Elder, K., and Davis, R. (2002). Comparison of spatial interpolation methods for estimating snow distribution in the Colorado Rocky Mountains. *Hydrological Processes*, 16(18):3627–3649.

Grabska, E., Frantz, D., and 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.

Gräler, B., Pebesma, E., and Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *The R Journal*, 8:204–218.

Han, L., Sun, J., Zhang, W., Xiu, Y., Feng, H., and Lin, Y. (2017). A machine learning nowcasting method based on real-time reanalysis data. *Journal of Geophysical Research: Atmospheres*, 122(7):4038–4051.

Hijmans, R. J. (2019). *raster: Geographic Data Analysis and Modeling*. R package version 3.0-7.

Jiang, Y. (2008). Prediction of monthly mean daily diffuse solar radiation using artificial neural networks and comparison with other empirical models. *Energy Policy*, 36(10):3833–3837.

Klimont, Z., Smith, S. J., and Cofala, J. (2013). The last decade of global anthropogenic sulfur dioxide: 2000–2011 emissions. *Environmental Research Letters*, 8(1):014003.

Li, J. (2013). Predicting the spatial distribution of seabed gravel content using random forest, spatial interpolation methods and their hybrid methods. In *20th International Congress on Modelling and Simulation, Adelaide, Australia*.

Li, J., Heap, A. D., Potter, A., and Daniell, J. J. (2011). Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software*, 26(12):1647–1659.

Liang, F. and Xia, X. A. (2005). Long-term trends in solar radiation and the associated climatic factors over China for 1961-2000. *Annales Geophysicae*, 23(7):2425–2432.

Liepert, B. G. (2002). Observed reductions of surface solar radiation at sites in the United States and worldwide from 1961 to 1990. *Geophysical Research Letters*, 29(10):61–1–61–4.

Nakagawa, S. and Freckleton, R. P. (2008). Missing inaction: the dangers of ignoring missing data. *Trends in Ecology Evolution*, 23(11):592 – 596.

Obryk, M. K., Fountain, A. G., Doran, P. T., Lyons, W. B., and Eastman, R. (2018). Drivers of solar radiation variability in the McMurdo Dry Valleys, Antarctica. *Scientific Reports*, 8(1):5002.

Ohashi, O. and Torgo, L. (2012). Spatial Interpolation Using Multiple Regression. In *2012 IEEE 12th International Conference on Data Mining*, pages 1044–1049. IEEE.

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, 5(2):9–13.

Pfeifroth, U., Sanchez-Lorenzo, A., Manara, V., Trentmann, J., and 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(3):1735–1754.

Phillips, P. C., Leirvik, T., and Storelvmo, T. (2020). Econometric estimates of Earth's transient climate sensitivity. *Journal of Econometrics*, 214(1):6–32.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rojas-Avellaneda, D. (2007). Spatial interpolation techniques for stimating levels of pollutant concentrations in the atmosphere. *Revista mexicana de física*, 53(6):447–454.

Sanchez-Lorenzo, A., Enriquez-Alonso, A., Wild, M., Trentmann, J., Vicente-Serrano, S. M., Sanchez-Romero, A., Posselt, R., and Hakuba, M. Z. (2017). Trends in downward surface solar radiation from satellites and ground observations over Europe during 1983–2010. *Remote Sensing of Environment*, 189:108–117.

Scudiero, E., Corwin, D. L., Morari, F., Anderson, R. G., and Skaggs, T. H. (2016). Spatial interpolation quality assessment for soil sensor transect datasets. *Computers and Electronics in Agriculture*, 123:74–79.

Smith, S. J., van Aardenne, J., Klimont, Z., Andres, R. J., Volke, A., and Delgado Arias, S. (2011). Anthropogenic sulfur dioxide emissions: 1850–2005. *Atmospheric Chemistry and Physics*, 11(3):1101–1116.

Stanhill, G. and Cohen, S. (2001). Global dimming: a review of the evidence for a widespread and significant reduction in global radiation with discussion of its probable causes and possible agricultural consequences. *Agricultural and forest meteorology*, 107(4):255–278.

Stephens, G. L., Li, J., Wild, M., Clayson, C. A., Loeb, N., Kato, S., L'Ecuyer, T., Stackhouse, P. W., Lebsock, M., and Andrews, T. (2012). An update on Earth's energy balance in light of the latest global observations. *Nature Geoscience*, 5(10):691–696.

Stern, D. I. (2006). Reversal of the trend in global anthropogenic sulfur emissions. *Global Environmental Change*, 16(2):207–220.

Stirnberg, R., Cermak, J., Fuchs, J., and Andersen, H. (2020). Mapping and Understanding Patterns of Air Quality Using Satellite Data and Machine Learning. *Journal of Geophysical Research: Atmospheres*, 125(4).

Storelvmo, T., Heede, U. K., Leirvik, T., Phillips, P. C. B., Arndt, P., and Wild, M. (2018). Lethargic Response to Aerosol Emissions in Current Climate Models. *Geophysical Research Letters*, 45(18):9814–9823.

Storelvmo, T., Leirvik, T., Lohmann, U., Phillips, P. C. B., and Wild, M. (2016). Disentangling greenhouse warming and aerosol cooling to reveal Earth's climate sensitivity. *Nature Geoscience*, 9(4):286–289.

Sun, H., Gui, D., Yan, B., Liu, Y., Liao, W., Zhu, Y., Lu, C., and Zhao, N. (2016). Assessing the potential of random forest method for estimating solar radiation using air pollution index. *Energy Conversion and Management*, 119:121 – 129.

Tao, P., Ni, G., Song, C., Shang, W., Wu, J., Zhu, J., Chen, G., and Deng, T. (2018). Solar-driven interfacial evaporation. *Nature Energy*, 3(12):1031–1041.

Thornton, P., van de Steeg, J., Notenbaert, A., and Herrero, M. (2009). The impacts of climate change on livestock and livestock systems in developing countries: A review of what we know and what we need to know. *Agricultural Systems*, 101(3):113–127.

Tollenaar, M., Fridgen, J., Tyagi, P., Stackhouse Jr, P. W., and Kumudini, S. (2017). The contribution of solar brightening to the US maize yield trend. *Nature Climate Change*, 7(4):275–278.

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

Wang, K., Dickinson, R. E., Ma, Q., Augustine, J. A., Wild, M., Wang, K., Dickinson, R. E., Ma, Q., Augustine, J. A., and Wild, M. (2013). Measurement Methods Affect the Observed Global Dimming and Brightening. *Journal of Climate*, 26(12):4112–4120.

Wang, S., Kinnison, D., Montzka, S. A., Apel, E. C., Hornbrook, R. S., Hills, A. J., Blake, D. R., Barletta, B., Meinardi, S., Sweeney, C., Moore, F., Long, M., Saiz-Lopez, A., Fernandez, R. P., Tilmes, S., Emmons, L. K., and Lamarque, J. F. (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(22):12319–12339.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wild, M. (2009). Global dimming and brightening: A review. *Journal of Geophysical Research: Atmospheres*, 114(D10).

Wild, M. (2016). Decadal changes in radiative fluxes at land and ocean surfaces and their relevance for global warming. *WIREs Climate Change*, 7(1):91–107.

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, 65(1):95–114.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

Xu, T., Guo, Z., Liu, S., He, X., Meng, Y., Xu, Z., Xia, Y., Xiao, J., Zhang, Y., Ma, Y., and Song, L. (2018). Evaluating Different Machine Learning Methods for Upscaling Evapotranspiration from Flux Towers to the Regional Scale. *Journal of Geophysical Research: Atmospheres*, 123(16):8674–8690.

Yang, Y., Xu, W., Hou, P., Liu, G., Liu, W., Wang, Y., Zhao, R., Ming, B., Xie, R., Wang, K., and Li, S. (2019). Improving maize grain yield by matching maize growth and solar radiation. *Scientific Reports*, 9(1):3635.

Zaki, M. F. M., Ismail, M. A. M., Govindasamy, D., and Zainalabidin, M. H. (2019). Interpretation and development of top-surface grid in subsurface ground profile using inverse distance weighting (IDW) method for twin tunnel project in Kenny hill formation. *Bulletin of the Geological Society of Malaysia*, 2019(67):103–109.

Zhang, X., Liang, S., Wild, M., and Jiang, B. (2015). Analysis of surface incident shortwave radiation from four satellite products. *Remote Sensing of Environment*, 165:186–202.

Zhou, Q., Flores, A., Glenn, N. F., Walters, R., and 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 us. *PloS one*, 12(8):e0180239.

Zhu, P., Wild, M., van Ruymbeke, M., Thuillier, G., Meftah, M., and Karatekin, O. (2016). Interannual variation of global net radiation flux as measured from space. *Journal of Geophysical Research: Atmospheres*, 121(12):6877–6891.

Supplementary Material

S1 RF model parameter tuning

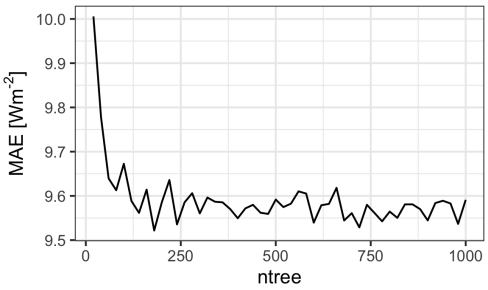

Fig. S1 MAE changing with the number of decision trees

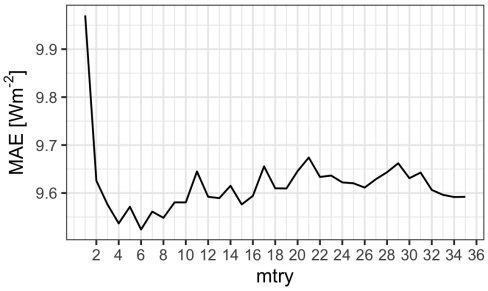

Fig. S2 MAE changing with the number of randomly preselected predictor variables

S2 Monthly anomalies

S2.1 Europe

Number of station Figure S3

Monthly anomalies for one station Figure S4

Monthly anomalies for LSDV+OK Figure S5

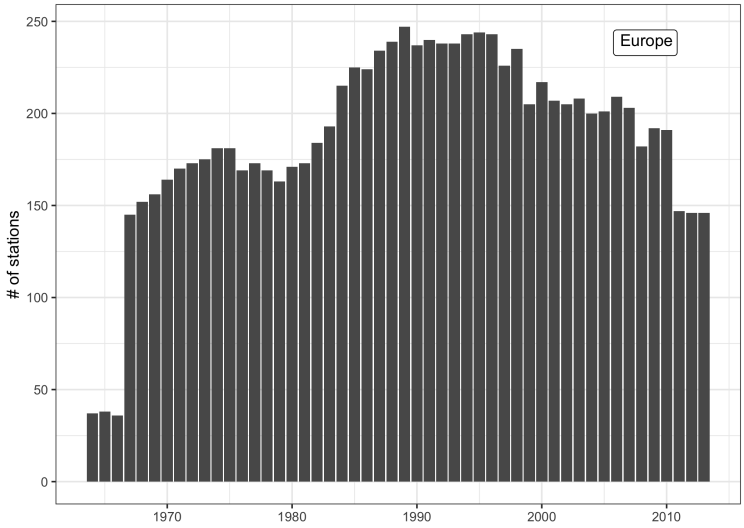

Fig. S3 Number of recorded stations in EU

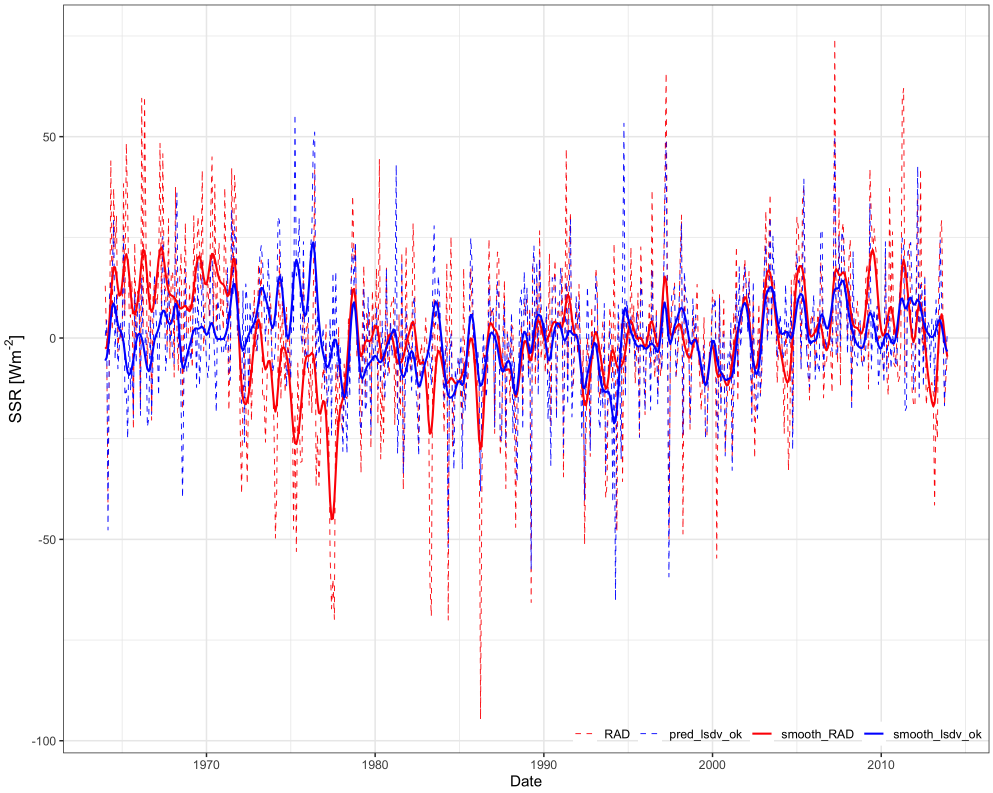

Fig. S4 Monthly SSR anomalies for observations and LSDV+OK simulations for one chosen station in Europe

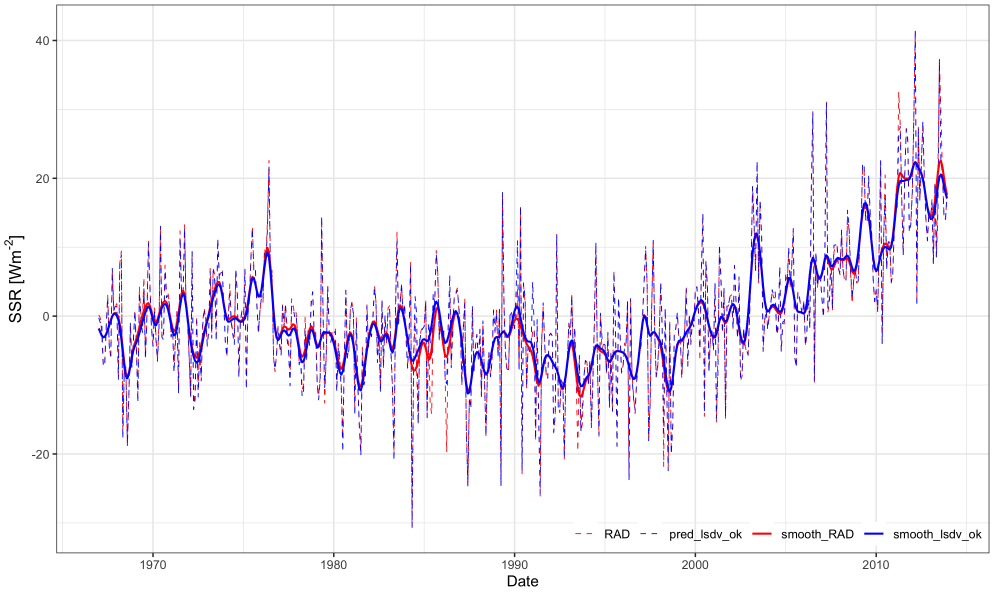

Fig. S5 Monthly SSR anomalies for observations and LSDV+OK simulations for Europe

S2.2 Africa

Number of station Figure S6

Prediction against observation plot Figure S7

Monthly anomalies trend Figure S8

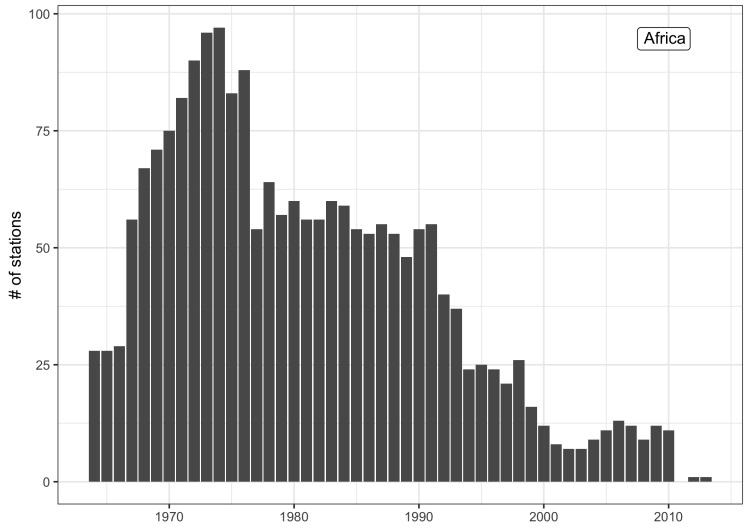

Fig. S6 Number of recorded stations in AF

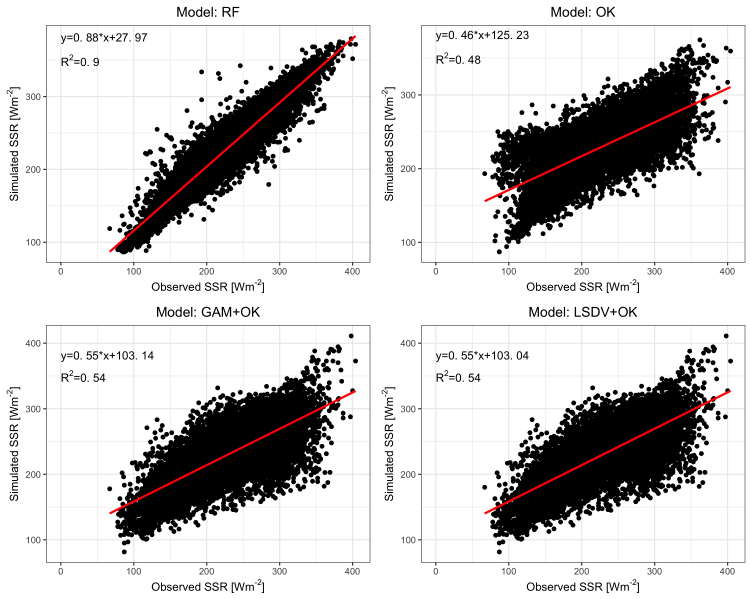

Fig. S7 Simulation v.s. observation values for AF

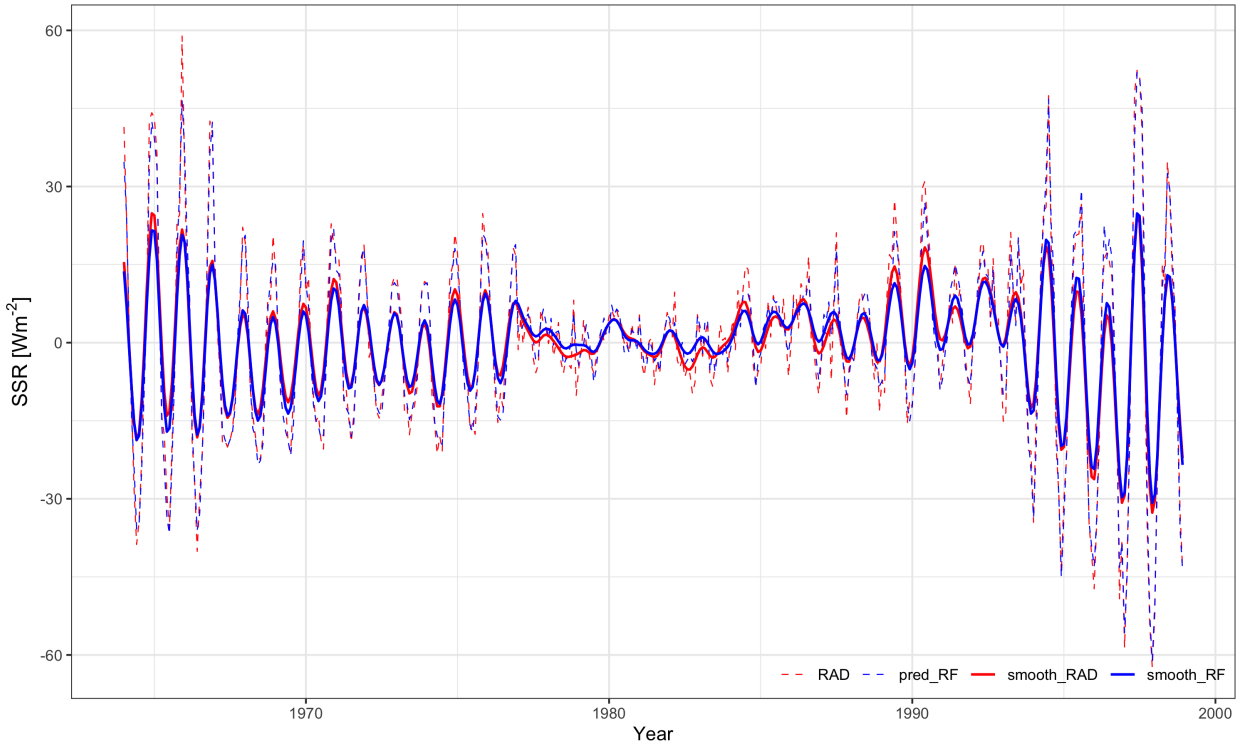

Fig. S8 Monthly anomalies for AF

S2.3 North America

Number of station Figure S9

Prediction against observation plot Figure S10

Monthly anomalies trend Figure S11

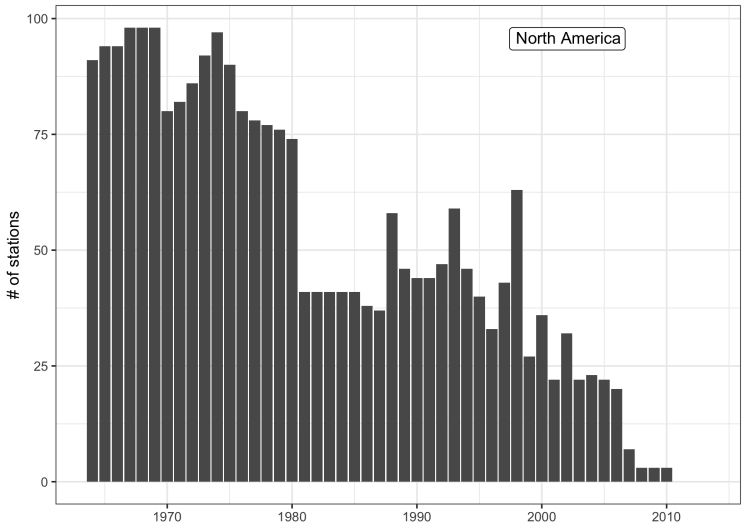

Fig. S9 Number of recorded stations in NA

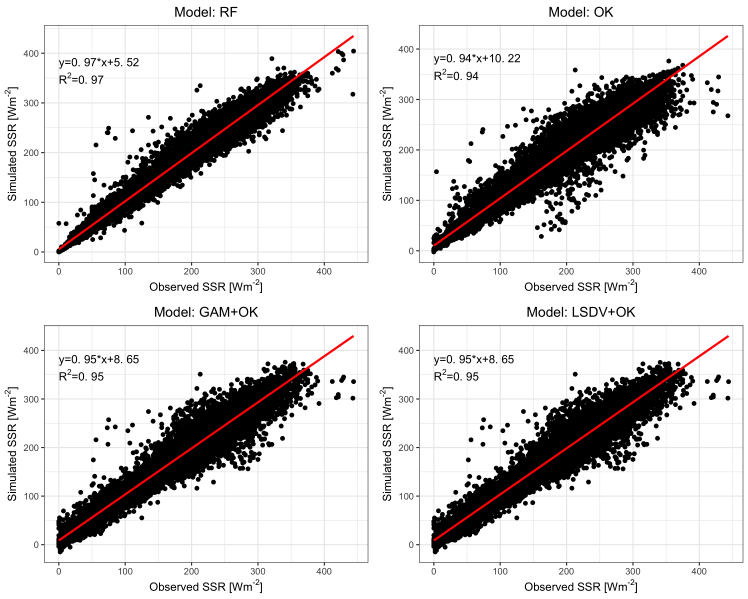

Fig. S10 Simulation v.s. observation values for NA

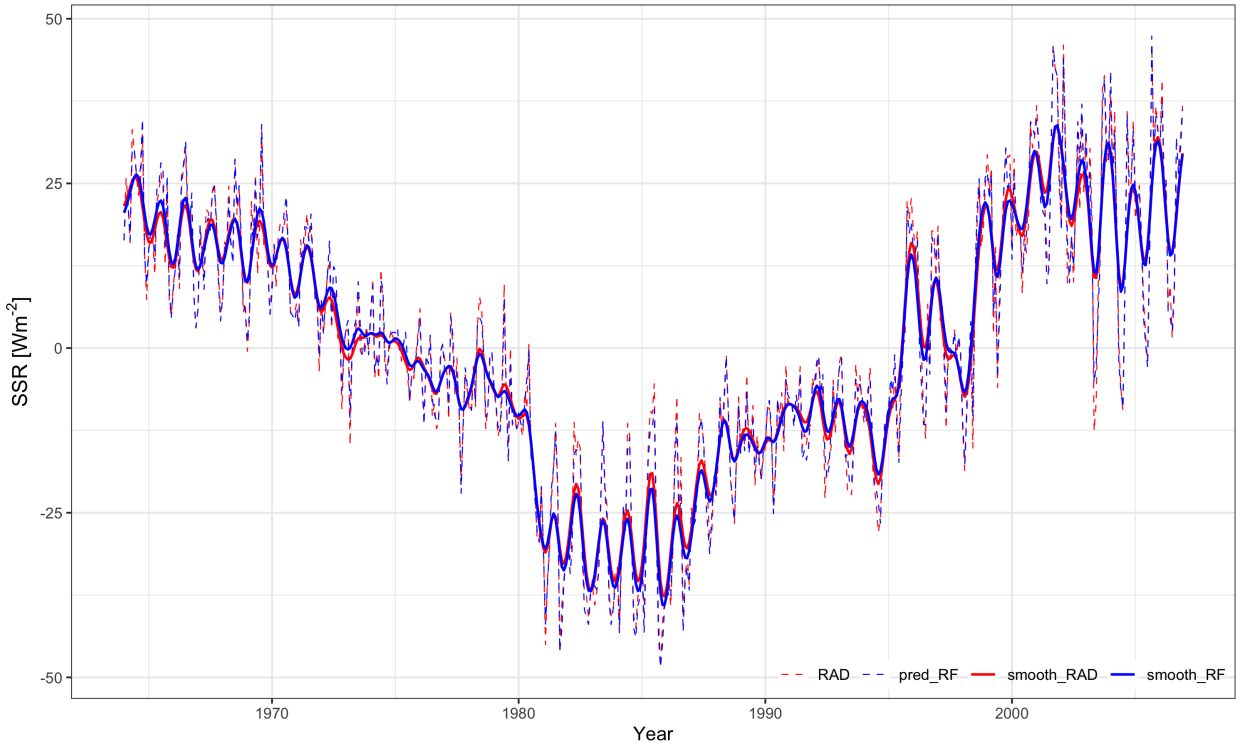

Fig. S11 Monthly anomalies for NA

S2.4 South America

Number of station Figure S12

Prediction against observation plot Figure S13

Monthly anomalies trend Figure S14

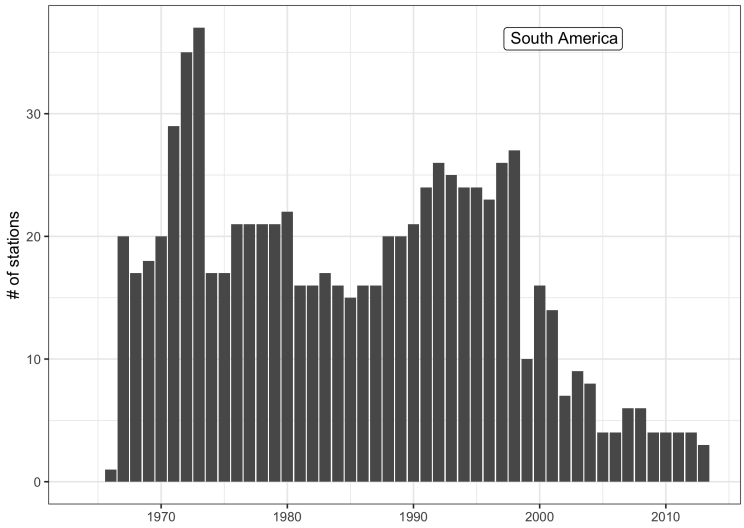

Fig. S12 Number of recorded stations in SA

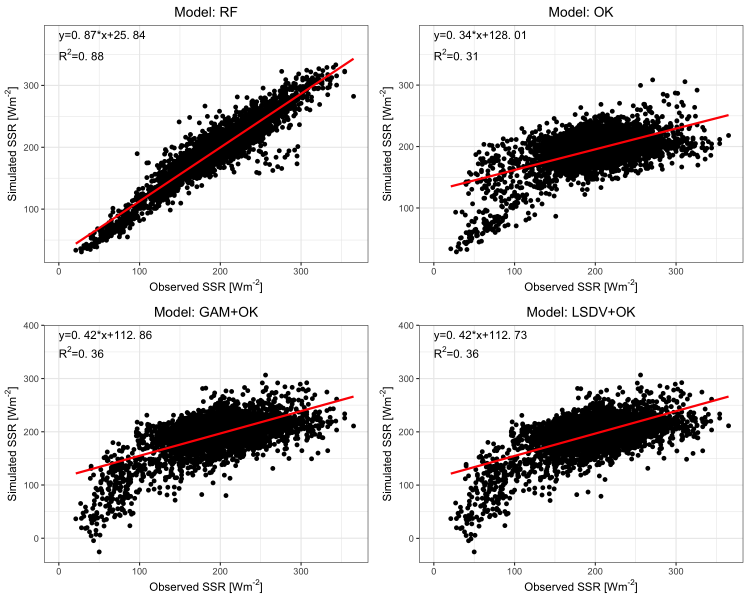

Fig. S13 Simulation v.s. observation values for SA

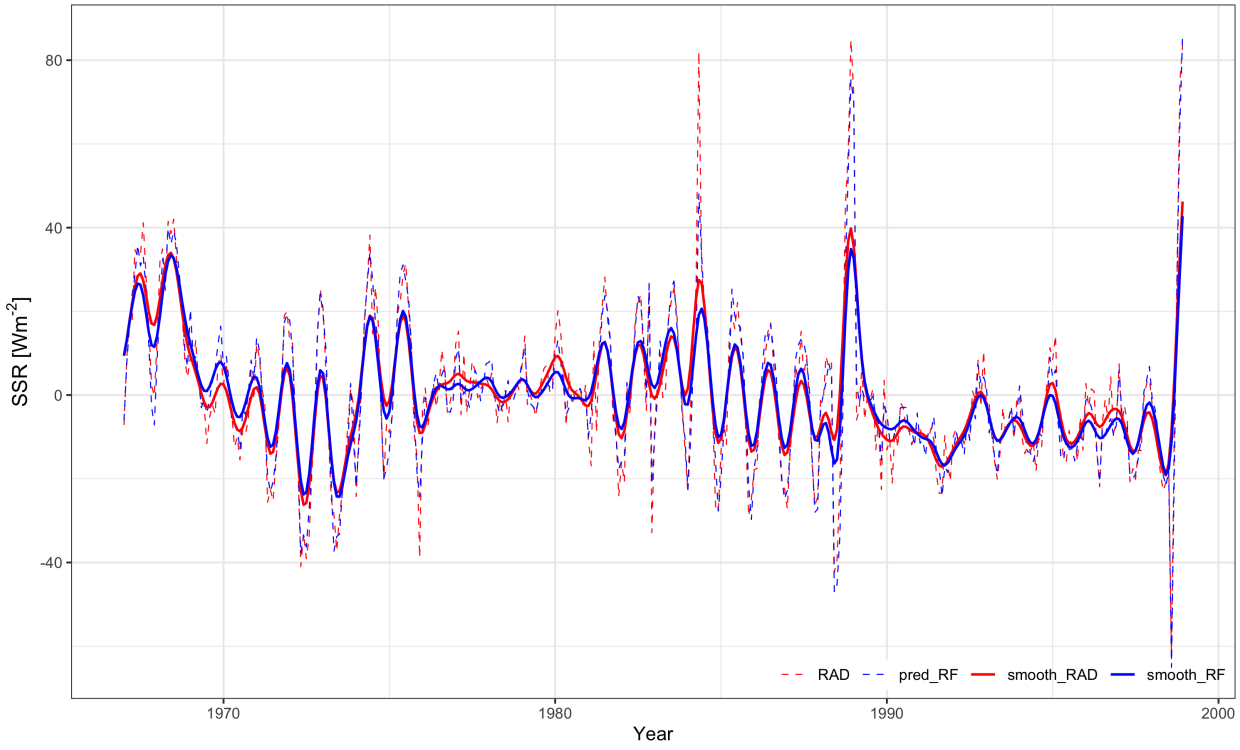

Fig. S14 Monthly anomalies for SA

S2.5 Asia

Number of station Figure S15

Prediction against observation plot Figure S16

Monthly anomalies trend Figure S17

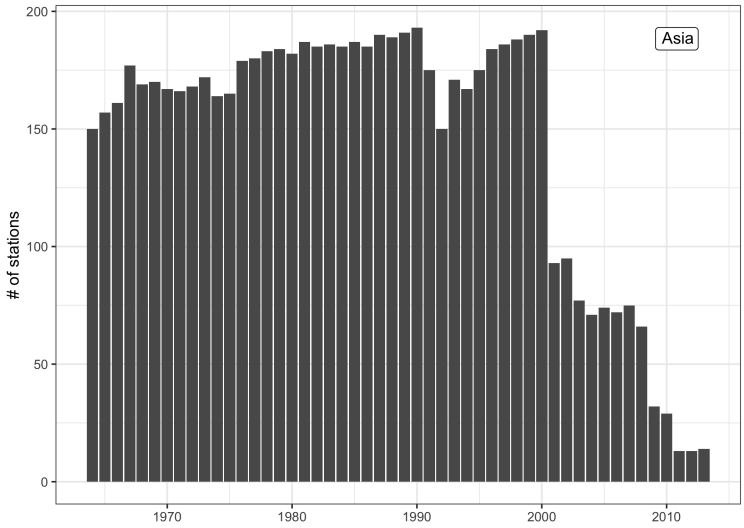

Fig. S15 Number of recorded stations in AS

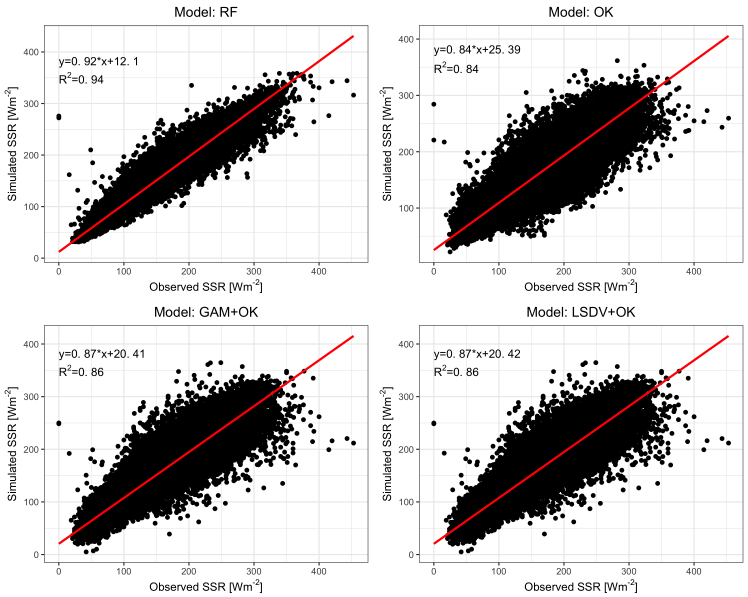

Fig. S16 Simulation v.s. observation values for AS

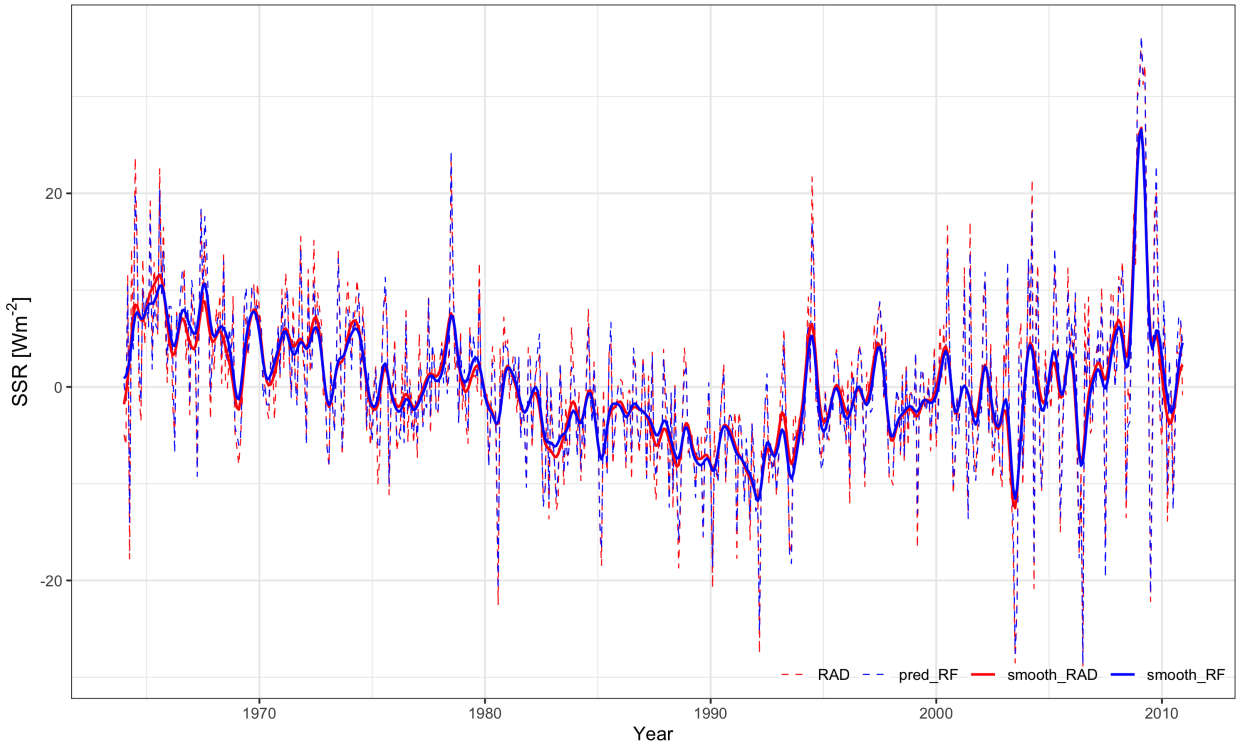

Fig. S17 Monthly anomalies for AS

S2.6 Oceania

Number of station Figure S18

Prediction against observation plot Figure S19

Monthly anomalies trend Figure S20

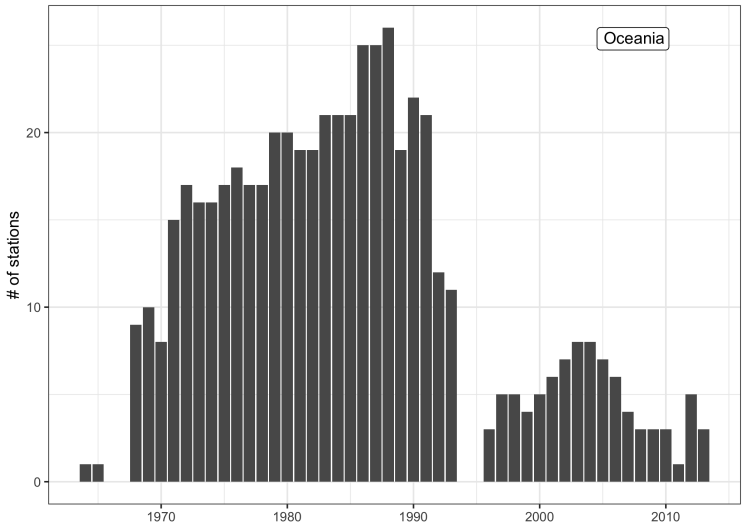

Fig. S18 Number of recorded stations in OC

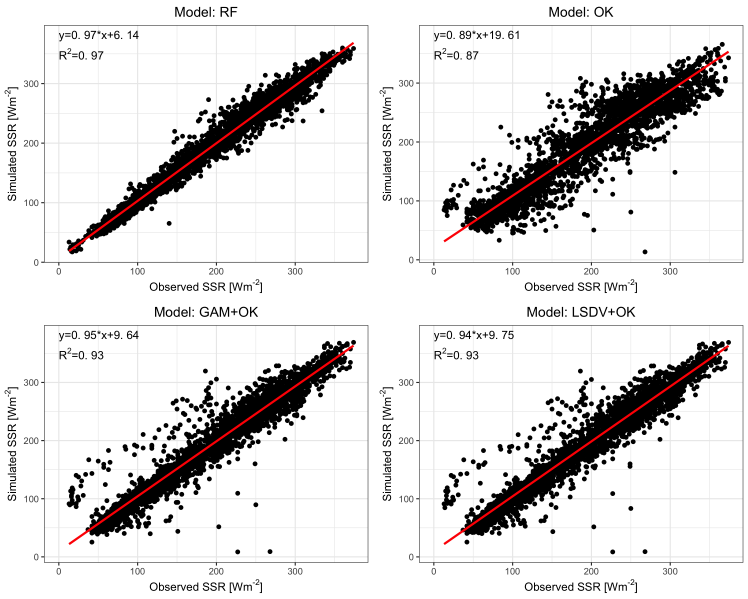

Fig. S19 Simulation v.s. observation values for OC

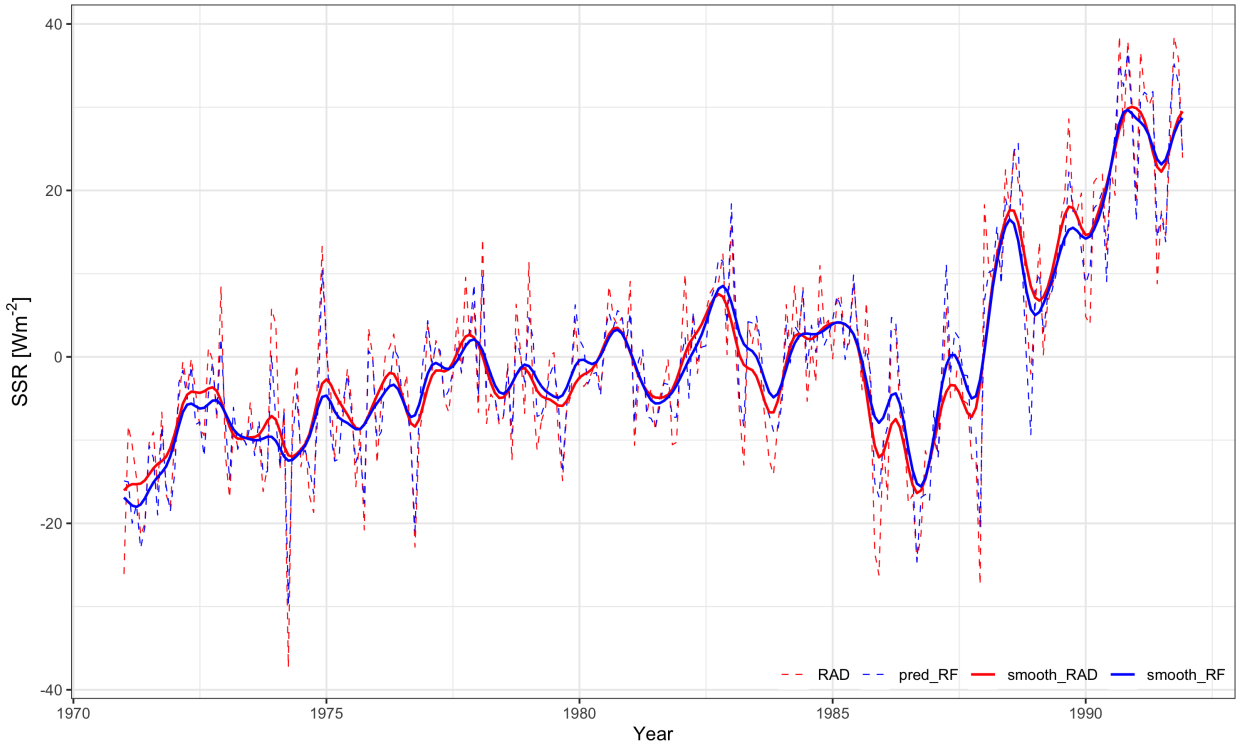

Fig. S20 Monthly anomalies for OC