scientific data

OPEN

DATA DESCRIPTOR

Using supervised learning to develop BaRAD, a 40-year monthly bias-adjusted global gridded radiation dataset

T. C. Chakraborty ✉ & Xuhui Lee

Diffuse solar radiation is an important, but understudied, component of the Earth's surface radiation budget, with most global climate models not archiving this variable and a dearth of ground-based observations. Here, we describe the development of a global 40-year (1980–2019) monthly database of total shortwave radiation, including its diffuse and direct beam components, called BaRAD (Bias-adjusted RADiation dataset). The dataset is based on a random forest algorithm trained using Global Energy Balance Archive (GEBA) observations and applied to the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) dataset at the native MERRA-2 resolution (0.5° by 0.625°). The dataset preserves seasonal, latitudinal, and long-term trends in the MERRA-2 data, but with reduced biases than MERRA-2. The mean bias error is close to 0 (root mean square error = 10.1 W m^{-2}) for diffuse radiation and −0.2 W m^{-2} (root mean square error = 19.2 W m^{-2}) for the total incoming shortwave radiation at the surface. Studies on atmosphere-biosphere interactions, especially those on the diffuse radiation fertilization effect, can benefit from this dataset.

Background & Summary

The Earth's climate is driven by solar (shortwave) radiation and its interactions with the different components of the Earth system. The shortwave radiation is attenuated by scattering and absorption by atmospheric aerosols, clouds, and gases, with the remaining portion reaching the Earth's surface as direct beam radiation ($K_{\downarrow,b}$). A portion of the scattered radiation also reaches the surface, which deviates from its original path and is known as diffuse radiation ($K_{\downarrow,d}$). The sum of $K_{\downarrow,b}$ and $K_{\downarrow,d}$, or the total incident shortwave radiation at the surface ($K_{\downarrow}$), influences local weather and climate, the hydrological cycle, and the carbon budget. There is also strong scientific interest in $K_{\downarrow,d}$ because a high diffuse fraction can increase agricultural and ecosystem productivity and enhance the terrestrial water flux to the atmosphere through increased photosynthesis in normally shaded parts of the plant canopy, a phenomenon known as the diffuse radiation fertilization effect[1–3].

Current Earth System Models (ESMs) generally overestimate $K_{\downarrow}$ compared to observations, primarily due to errors associated with parameterizations of clouds and aerosols[4–7]. This overestimation would cause artificial surface warming, with undesired consequences on atmosphere-biosphere interactions[8,9]. Although similar evaluations of ESM $K_{\downarrow,d}$ are not available, large differences are reported for $K_{\downarrow,d}$ between reanalysis datasets and observations[10]. The bias in $K_{\downarrow,d}$ in these gridded datasets is not consistent in direction, unlike that for $K_{\downarrow}$. Such biases may contribute to uncertainties in modelling surface energy and carbon budgets and impact optimum placement of concentrating solar power systems[11,12].

Several previous studies have examined the biases in modeled $K_{\downarrow}$ using the clearness index (k_t). This index, defined as the ratio between surface incident and extraterrestrial radiation, captures the combined impact of aerosols, clouds, and gases on atmospheric transmittance on solar radiation[13–15]. These atmospheric constituents attenuate solar radiation as it moves through the atmospheric column. Although k_t, a measure of the total light extinction, directly affects $K_{\downarrow,b}$ and therefore exerts a strong control on $K_{\downarrow}$, it is only tangentially related to $K_{\downarrow,d}$. It is known that $K_{\downarrow,d}$ is primarily controlled by the abundance of scattering agents in the atmosphere, as well as their degree of forward scattering[16]. An atmospheric scattering agent that reduces $K_{\downarrow,b}$ may actually increase $K_{\downarrow,d}$. Thus, a new approach is required to correct biases in $K_{\downarrow,d}$.

School of the Environment, Yale University, New Haven, CT, 06520, USA. ✉e-mail: tc.chakraborty@yale.edu

In recent years, machine learning algorithms have been used to reduce biases in radiation fields derived from reanalysis products or derive the fields from satellite observations[17–22]. By training against observed data, these algorithms can capture previously unknown relationships between actual and gridded variables, generally leading to improvements over traditional parametric and multi-ensemble averaging techniques[22]. However, the majority of these algorithms have been implemented at the regional scale, particularly over China, Europe, and the US, with a focus on the total $K_{\downarrow}$. For reasons briefly described above, it is also important to develop a generalizable bias-correction algorithm for $K_{\downarrow,d}$. Of note, a recent study developed a global hourly $K_{\downarrow,d}$ dataset using a random forest algorithm on satellite retrievals from the Earth Polychromatic Imaging Camera (EPIC)[21], although this focused on a short period from June 2015 to June 2019. A gridded data product after proper bias correction is especially welcome for tropical regions where $K_{\downarrow,d}$ measurements are rare but the diffuse fertilization effect is strong due to high vegetation densities[3,23].

In this paper, we describe the development of a new dataset of monthly gridded radiation fields, including $K_{\downarrow}$, $K_{\downarrow,b}$, and $K_{\downarrow,d}$, from 1980 to 2019, which can be explored through this web application: https://yceo.users.earthengine.app/view/barad. We attempt to improve historical global gridded estimates of $K_{\downarrow,d.}$ through three major steps:

1. Examine the control of k_t on biases in $K_{\downarrow}$, $K_{\downarrow,b,}$ and $K_{\downarrow,d}$ separately
2. Test bias-correction algorithms for $K_{\downarrow}$ and $K_{\downarrow,d}$, including a method based on k_t, a multiple linear regression (MLR) and a random forest (RF) model
3. Implement the best performing bias-correction algorithm to create a global 40-year Bias-adjusted RADiation dataset, or BaRAD.

Methods

Reanalysis data. The gridded data reported here is based on the Modern-Era Retrospective analysis for Research and Applications, version 2 (MERRA-2) global reanalysis dataset[24]. MERRA-2 improves upon the original MERRA dataset in several ways. It adds an extensive aerosol assimilation by using bias-adjusted aerosol optical depth (AOD) from satellite observations[24]. Unlike MERRA, MERRA-2 uses observed precipitation to force the land-surface model[25]. It uses a newer version of the Goddard Earth Observing System (GEOS-5) and assimilates newer satellite observations of aerosols, clouds, and precipitation[26]. MERRA-2 is available from 1980 to present day at a grid resolution of 0.5° latitude and 0.625° longitude. The variables we wish to correct are monthly mean $K_{\downarrow}$ and $K_{\downarrow,d}$ using predictors that physically control transmitted radiation. They include estimates of atmospheric clouds and aerosols, as well information about the position of the Sun, which controls energy input to the atmospheric column.

Ground-Based observations for training and validation. We used the Global Energy Balance Archive (GEBA) for training and validation of bias correction algorithms. GEBA is a comprehensive observational data repository of the components of the Earth's surface energy budget[27]. The latest version of the database has roughly 2500 unique stations[28]. Here, we used the monthly mean $K_{\downarrow}$ and $K_{\downarrow,d}$ stored in the database. The data were screened with several quality control steps. We only selected the monthly mean values lower than 600 W m^{-2} for $K_{\downarrow}$ and 250 W m^{-2} for $K_{\downarrow,d}$. Cases where the ratio of modeled to observed monthly means exceed 5 were ignored. Finally, only sites with all 12 months of available data were selected to avoid biased representation across seasons. After these data screening steps, we obtained 935 unique sites with 134541 site-months of data for $K_{\downarrow}$ and 290 unique sites with 28880 site-months for $K_{\downarrow,d}$ between 1980 and 2017 (Fig. S1). Monthly mean $K_{\downarrow,b}$ was computed as the difference between $K_{\downarrow}$ and $K_{\downarrow,d}$.

Bias-Correction algorithms. We tested three bias correction algorithms, including a technique based on clearness index and two data-driven algorithms. Several studies have used clearness index k_t as a threshold for designating sky condition or for estimating $K_{\downarrow}$[13,29,30]. In Zhao *et al.*[29], the bias in $K_{\downarrow}$ (b_m) is related to k_t in a linear fashion:

$$b_{m=}b_0.k_{t+}b_1 \tag{1}$$

Here b_0 is the sensitivity of b_m to k_t, and b_1 is the model bias ratio under completely cloudy conditions. In their study, b_m is given as

$$b_m = \frac{K_R - K_O}{K_R} \tag{2}$$

where K_R and K_O are modeled and observed values, respectively. Clearness index is given by

$$k_t = \frac{K_{\downarrow,o}}{K_{TOA}} \tag{3}$$

where K_{TOA} is the extra-terrestrial radiation at the top of the atmosphere and $K_{\downarrow,O}$ is the observed $K_{\downarrow}$ value. Their method subsequently also accounted for site elevation H through a somewhat arbitrarily chosen quadratic fitting function. Here, we used a multi-linear regression (MLR) model, which the authors[29] note would yield similar results, as a function of k_t, H, and $K_{\downarrow,R}$, the $K_{\downarrow}$ from the reanalysis without correction, to correct $K_{\downarrow}$

$$K_{\downarrow} = \beta_0 K_{\downarrow,R} + \beta_1 k_t + \beta_2 H + \beta_3 \quad (4)$$

where β_0, β_1, β_2, and β_3 are empirical coefficients. A linear model of the same form was also used to correct $K_{\downarrow,d}$. Since k_t involves observed $K_{\downarrow}$ (Eq. 1), Eq. 4 cannot be used to correct biases in gridded data when observations are not available. Thus, we considered two variations of this algorithm, one using observed $K_{\downarrow}$ ($K_{\downarrow,O}$) and site elevation to calculate clearness index, called the $k_{t,O}$ model, and the other using grid-averaged terrain elevation (H_R) and the clearness index calculated from modeled $K_{\downarrow}$ ($K_{\downarrow,R}$), given by:

$$k_t = \frac{K_{\downarrow,R}}{K_{TOA}} \quad (5)$$

which we call the $k_{t,R}$ model.

The second algorithm, another MLR model, expresses the dependent variable as a linear combination of predictors. In the case of $K_{\downarrow}$, it takes the following form

$$K_{\downarrow,O} = \beta_0 K_{\downarrow,R} + \beta_1 \mathrm{SAOD} + \beta_2 \mathrm{AAOD} + \beta_3 \mathrm{COD} + \beta_4 \mathrm{CF} + \beta_5 \theta_z + \beta_6 H_R + \beta_7 \quad (6)$$

where β_0 to β_7 are regression coefficients, $K_{\downarrow,O}$ is the observed (or bias corrected) $K_{\downarrow}$, $K_{\downarrow,R}$ is the $K_{\downarrow}$ from the reanalysis without correction, SAOD is scattering aerosol optical depth (AOD), AAOD is absorption AOD, COD is cloud optical depth, CF is cloud fraction, and θ_z is the monthly mean zenith angle – the angle between the sun and the vertical direction – estimated from the hourly values. The MLR procedure with the same set of predictors was also applied to $K_{\downarrow,d}$. These predictors (summarized in Table S1) provide strong physical constraints on atmospheric radiative transfer[31], with both COD and AOD being direct measures of light extinction along the atmospheric column. Correlation matrices for the features selected show that other than for θ_z and the radiation field (K_R) and AAOD and SAOD, the correlations coefficients between the features are generally smaller than 0.75. Although AAOD and SAOD are strongly correlated, the separation of AOD into SAOD and AAOD is more important for $K_{\downarrow,d}$ than for $K_{\downarrow}$ since while absorption of solar radiation by aerosols would reduce both $K_{\downarrow,d}$ and $K_{\downarrow,b,}$ forward scattering would reduce $K_{\downarrow,b}$ and increase $K_{\downarrow,d}$. With the intent of developing a generalized algorithm, one regression is used for the entire dataset. Since θ_z strongly controls the optical thickness of the atmosphere even for clear-sky conditions[32] and is one of the predictors, seasonal and latitudinal effects are accounted for to some extent. The algorithm was implemented using the stats package on the R programming language.

The third algorithm is a random forest (RF) regression technique[33]. Unlike the MLR models, the RF regression does not assume a standard linear structure of the relationship; instead it derives the relationship from the training data using an ensemble of decision trees. This relationship (for the total incoming radiation) can be expressed in a generic form as:

$$K_{\downarrow,O} = f(K_{\downarrow,R,}\mathrm{SAOD}, \mathrm{AAOD}, \mathrm{COD}, \mathrm{CF}, \theta_z, H_R) \quad (7)$$

This random forest regression was implemented using the R Random Forest package. The default minimum size of terminal nodes (5) was used, but the maximum number of trees to generate was set to 2000. In most folds, the models converged before reaching this limit. As per the default parameters of the package, each tree is trained on 63.2% of the training data with 2 predictor variables chosen at random to split the nodes. Trees were allowed to grow fully rather than be pruned.

We used a 10-fold cross-validation technique to evaluation the performance of these algorithms. The entire GEBA dataset was randomly partitioned into 10 equal subsets. One of the ten subsets was used for validation and the other nine for training. The process was repeated 10 times. The accuracy was quantified using the coefficient of determination (r^2), the root mean square error (RMSE), and the mean bias error (MBE). Cross-validation is desired for the RF algorithm because it is prone to overfitting and using multiple folds allows us to examine the consistency of the results across different training/validation splits. The two linear models (Eqs. 4 and 6) are not prone to overfitting. However, because they are sensitive to outliers, cross-validation was also done to estimate the influence of the training data selection on their performance.

The final data product (BaRAD) consists of monthly $K_{\downarrow,}$ $K_{\downarrow,b,}$ and $K_{\downarrow,d}$ corrected with the best performing algorithm, defined as the one with minimum RMSE and highest r^2 in the consolidated validated data at the native MERRA-2 resolution. Here the algorithm was trained on the whole quality screened GEBA dataset.

Clearness index as a predictor of bias. Zhao *et al.*[29] found systematic overestimation of $K_{\downarrow}$ in two reanalysis datasets. To correct these model biases, they utilized the empirical relationship between the sensitivity of b_m to the observed k_t. Here the sensitivity is the slope of the linear regression between b_m and $k_{t,O}$. To illustrate how this sensitivity varies between $K_{\downarrow}$, $K_{\downarrow,d}$, and $K_{\downarrow,b}$, we separately examined the associations between b_m and $k_{t,O}$.

Unsurprisingly, b_m for $K_{\downarrow}$ and $k_{t,O}$ are negatively correlated, both overall and for the common sites (Fig. S2b and d). Here the common sites are those with simultaneous measurements of $K_{\downarrow}$ and $K_{\downarrow,d}$. The sensitivity of b_m to $k_{t,O}$ is −0.76 for all sites and −0.8 for common sites, which are very close to the value of −0.82 found by Zhao *et al.*[29] for MERRA in North America. Similarly, b_m for $K_{\downarrow,b}$ is also negatively correlated with $k_{t,O}$, with the sensitivity being higher in magnitude (−1.23; Fig. S2c) than that for $K_{\downarrow}$, suggesting that total atmospheric transmittance has a stronger effect on the biases in $K_{\downarrow,b}$ than on the biases in $K_{\downarrow}$. For $K_{\downarrow,d}$, the sensitivity of b_m to $k_{t,O}$ is strong (−0.89; Fig. S2a), but the variability in the bias is not explained well by it ($r^2 = 0.15$). Overall, the coefficient of determination (r^2) is highest for $K_{\downarrow,b}$ and smallest for $K_{\downarrow,d}$, indicating that clearness index is a poor predictor of model bias in $K_{\downarrow,d}$.

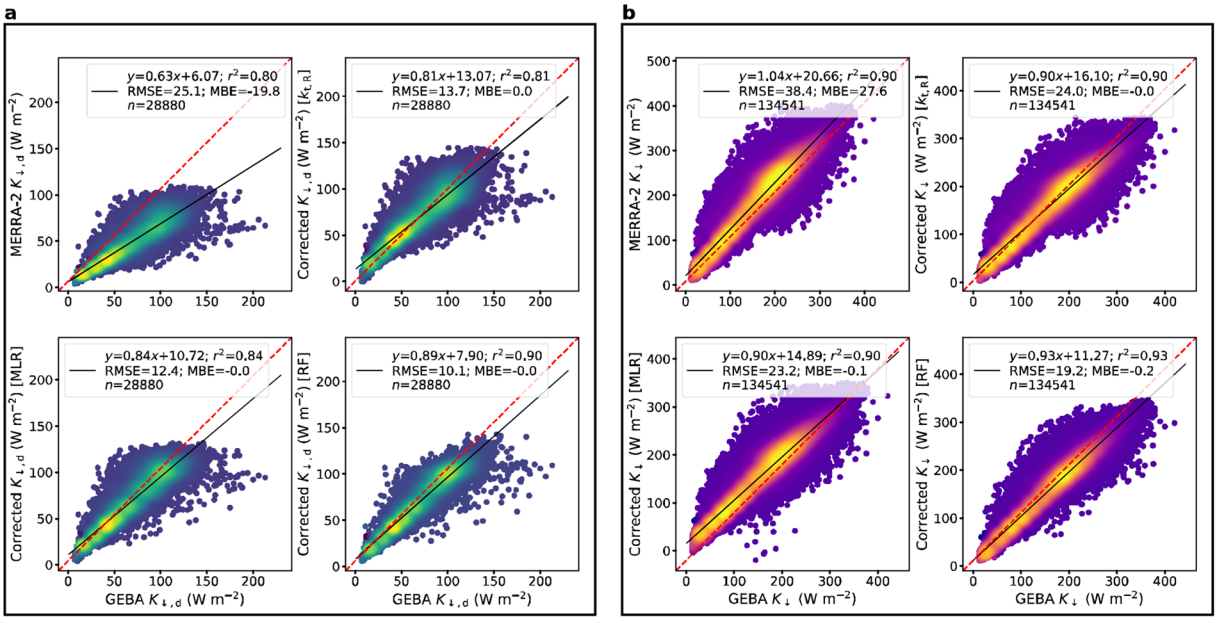

Fig. 1 Comparison of original and bias-adjusted MERRA-2 data with GEBA observations. (**a**) monthly mean diffuse radiation ($K_{\downarrow,d}$) and (**b**) total shortwave radiation ($K_{\downarrow}$) from MERRA-2 as well as the bias-adjusted estimates from the $k_{t,R}$, MLR, and RF models. For the bias-adjusted estimates, the consolidated validation data from all 10 folds are shown. The red dashed lines represent the 1:1 relationship. Color indicates data density and the statistical summaries of the evaluations are noted.

It is also important to note the intercept of the equations shown in Fig. S2. This intercept represents the b_m for a completely non-transmissive atmosphere (i.e. when $k_{t,O} = 0$). For both $K_{\downarrow}$ and $K_{\downarrow,b}$, this value is positive (0.96 for $K_{\downarrow,b}$; 0.5 to 0.53 for $K_{\downarrow}$). This implies that the reanalysis overestimates $K_{\downarrow}$ under non-overcast skies, and its estimates improve for clearer conditions. On the other hand, the intercept for the regression line between the b_m for $K_{\downarrow,d}$ and $k_{t,O}$ is close to zero and the slope is negative, suggesting that MERRA-2 $K_{\downarrow,d}$ is underestimated even under completely clear conditions.

Comparing bias-correction algorithms. Figure 1 shows the comparison of the original MERRA-2 and bias-adjusted values with the GEBA observations. MERRA-2 underestimates $K_{\downarrow,d}$ (MBE = −19.8 W m^{-2}; Fig. 2a) and overestimates $K_{\downarrow}$ (MBE = 27.6 19.8 W m^{-2}; Fig. 2b). Consistent with the $K_{\downarrow}$ overestimation, the modeled clearness index $k_{t,R}$ (0.54 ± 0.11) is higher than the observed index $k_{t,O}$ (0.45 ± 0.12). This increased transmissivity may be caused by underestimation of both clouds and aerosols, although clouds probably play a greater role since MERRA-2 has assimilated observations of AOD. Although an underestimation in clouds would also explain the underestimation in $K_{\downarrow,d}$, the intercept of the equation in Fig. S2a (see previous subsection) suggests that clouds are not the only factor.

All the three algorithms reduce the MBE and RMSE of $K_{\downarrow}$, $K_{\downarrow,b}$, and $K_{\downarrow,d}$ in comparison to the original MERRA-2 values. The RF model performs the best overall, minimizing the RMSE and maximizing r^2 for both $K_{\downarrow}$ (RMSE = 19.2 W m^{-2}; $r^2 = 0.93$) and $K_{\downarrow,d}$ (RMSE = 10.1 W m^{-2}; $r^2 = 0.90$). The Taylor diagrams for the composite validation dataset, along with the results for both the $k_{t,O}$ and $k_{t,R}$ models, are in the supplementary information (Fig. S3). The RF model consistently outperforms the others for every fold (with one exception; see below). For $K_{\downarrow,d}$, the MLR model is not as good as the RF model but is better than the $k_{t,R}$ and $k_{t,O}$ models (Fig. S3a). For $K_{\downarrow}$, the $k_{t,O}$ model performs slightly better than the RF model (Fig. S3b), which makes sense since $k_{t,O}$ includes the observed $K_{\downarrow}$, and thus this model, not useable to correct global datasets, is not shown in Fig. 1. That the $k_{t,O}$ model outperforms the other models for $K_{\downarrow}$ but not for $K_{\downarrow,d}$ confirms our hypothesis that the k_t model is not appropriate to address biases in $K_{\downarrow,d}$. Similar to the $k_{t,R}$ MLR model (Eq. 4), we also train a RF model with k_t derived from MERRA-2 as one of the features. Consistent with the results of the $k_{t,O}$ MLR and $k_{t,R}$ MLR models, the $k_{t,R}$ RF model cannot beat the performance of the RF model using MERRA-2 features (Eq. 7).

Physically, the monthly average radiation components cannot be negative. However, both the $k_{t,R}$ and MLR models predict a small fraction of negative values for $K_{\downarrow}$ (0.15% for $k_{t,R}$ and 0.10% for MLR) and $K_{\downarrow,d}$ (0.24% for $k_{t,R}$ and 0.01% for MLR). One can account for this by setting these negative values to zero after correction. However, this imposed physical constraint is not required for the RF corrected values. We also test whether the distribution of predicted values by the RF model is statistically different from those predicted by the other models using paired Wilcoxon Sign Rank Tests[34]. In all cases, except compared to the $k_{t,O}$ model for $K_{\downarrow}$ (p-value = 0.58; supporting the null hypothesis of no difference), there are statistically significant differences between the tested algorithms. This is further evidence of the usefulness of k_t based models for $K_{\downarrow}$ but not $K_{\downarrow,d}$.

MLR and RF use the same gridded variables as predictors. Fig. S5 presents the feature importance of each variable. For the RF model, a feature importance is the increase in mean square error (MSE) of the predicted values

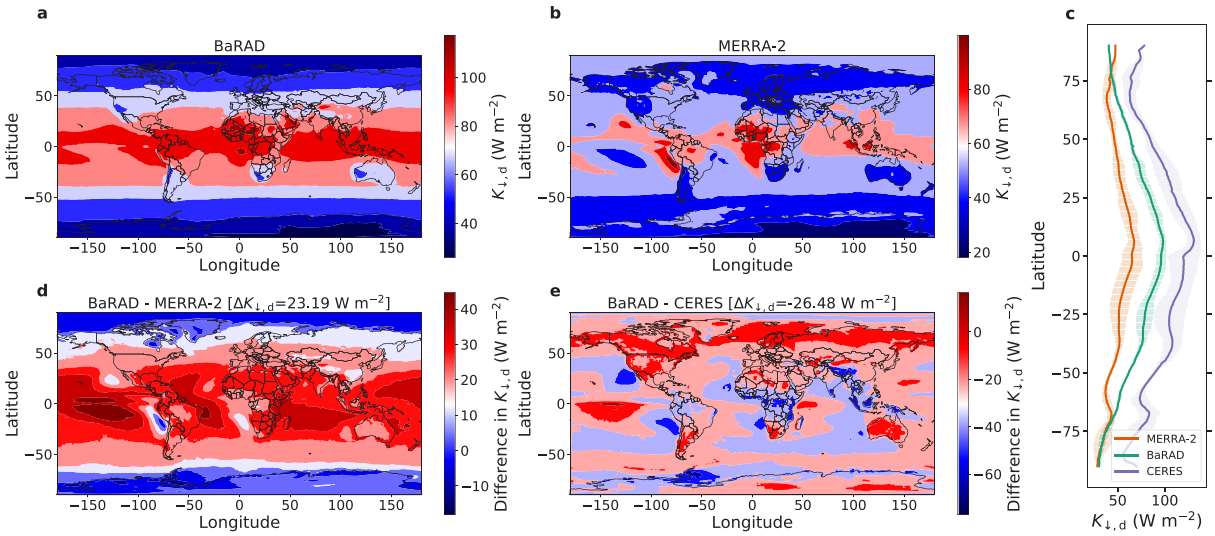

Fig. 2 Spatial and latitudinal variability in diffuse radiation. Global pattern of diffuse radiation ($K_{\downarrow,d}$) in (**a**) the BaRAD product, (**b**) the MERRA-2 dataset, and (**c**) the CERES dataset. The grid-wise difference between BaRAD and (**e**) MERRA-2 and (**f**) CERES are also shown. Sub-figure (**b**) shows the mean latitudinal variability of $K_{\downarrow,d}$ in all three products. The shaded areas represent the standard deviation. The area-weighted mean difference in $K_{\downarrow,d}$ ($\Delta K_{\downarrow,d}$) between the BaRAD data and the MERRA-2 and CERES products, respectively, are shown at the top of sub-figures (**d**) and (**e**), respectively.

if the same variable is removed from the model. A higher increase in MSE indicates that the variable is more important to the performance of the RF model. Although there are several established methods for interpretating the relative importance of variables for MLR models, for ease of comparison, we use a model-agnostic permutation method similar to the one used for the RF model using the iml package for the R programming language. For the RF model, the two best predictors are different for $K_{\downarrow,d}$ and for $K_{\downarrow}$. For $K_{\downarrow,d}$, COD and CF have the highest importance scores ($224.4 \pm 3.3\%$ for COD; $138.2 \pm 3.1\%$ for CF; Fig. S5c), and for $K_{\downarrow}$, AAOD and SAOD have the highest importance scores ($223.6 \pm 16.5\%$ for AAOD; $206.5 \pm 7.5\%$ for SAOD; Fig. S5d). In contrast, the radiation field is the most important variable in the MLR models for both $K_{\downarrow}$ ($451.2 \pm 1.7\%$) and $K_{\downarrow,d}$ ($276.5 \pm 1.1\%$; Figs. S5a and S5b). These differences are expected since the model architectures are also different, with the MLR model assuming linear relationships between the output and the input features and the RF model also accounting for non-linear interactions.

The BaRAD dataset. Based on our cross-validation results, we choose the RF model to adjust the biases in the MERRA-2 $K_{\downarrow}$ and $K_{\downarrow,d}$. We re-trained the model twice, one for $K_{\downarrow}$ and the other for $K_{\downarrow,d}$, using the same predictors and all available quality screened GEBA observations (instead of random training subsets of it as done during the cross-validation phase). The trained model was used to bias-adjust the corresponding gridded monthly MERRA-2 fields from 1980 to 2019. The bias-adjusted dataset is referred to as BaRAD. The final BaRAD data deposited in the public archive has gone through two additional post-correction adjustments. First, because of lack of training data in polar regions, there exist a few positive values at some polar grids during polar nights; these positive values constitute 6.5% of the entire dataset for $K_{\downarrow}$. Here we have forced the bias-adjusted $K_{\downarrow}$ and $K_{\downarrow,d}$ to zero when the corresponding MERRA-2 values are zero in those grids. Second, since $K_{\downarrow,d}$ and $K_{\downarrow}$ were trained separately, there is a small fraction of gridded data (less than 0.5%) where $K_{\downarrow,d}$ exceeds $K_{\downarrow}$, which is physically impossible. For these cases, we have set the $K_{\downarrow,d}$ value equal to $K_{\downarrow}$.

Data Records

The BaRAD dataset is available in netCDF format and includes the monthly values of $K_{\downarrow}$ (variable name: K_down), $K_{\downarrow,d}$ (variable name: K_diff), and $K_{\downarrow,b}$ (variable name: K_dir) starting from January, 1980[35]. Separate netCDF files are generated for each year from 1980 to 2019. All variables have the unit of W m^{-2} and are available at the MERRA-2 native resolution of 0.5° by 0.625°. The BaRAD dataset generated in this study is available in this GitHub repository: https://github.com/TC25/BaRAD/tree/main/BaRAD_Dataset and also through PANGAEA (https://doi.org/10.1594/PANGAEA.932924)[35]. The training data are also available in the main GitHub repository.

Technical Validation

Comparison of BaRAD dataset with other gridded data products. In Figs. 2, S6, and 3, we compare the spatial, zonal, and seasonal patterns in the BaRAD dataset with the original MERRA-2 dataset. We also compare these patterns with the latest version of the Clouds and the Earth's Radiant Energy System (CERES) surface radiation product[36]. The CERES dataset provides satellite-based estimates of the Earth's radiative budget (from the surface to the top of the atmosphere) and clouds. The data are available globally at 1° by 1° resolution from 2000 onwards. The latest version (CERES_SYN1deg_Ed4.1) of the dataset includes monthly estimates of both $K_{\downarrow,d}$ and $K_{\downarrow}$.

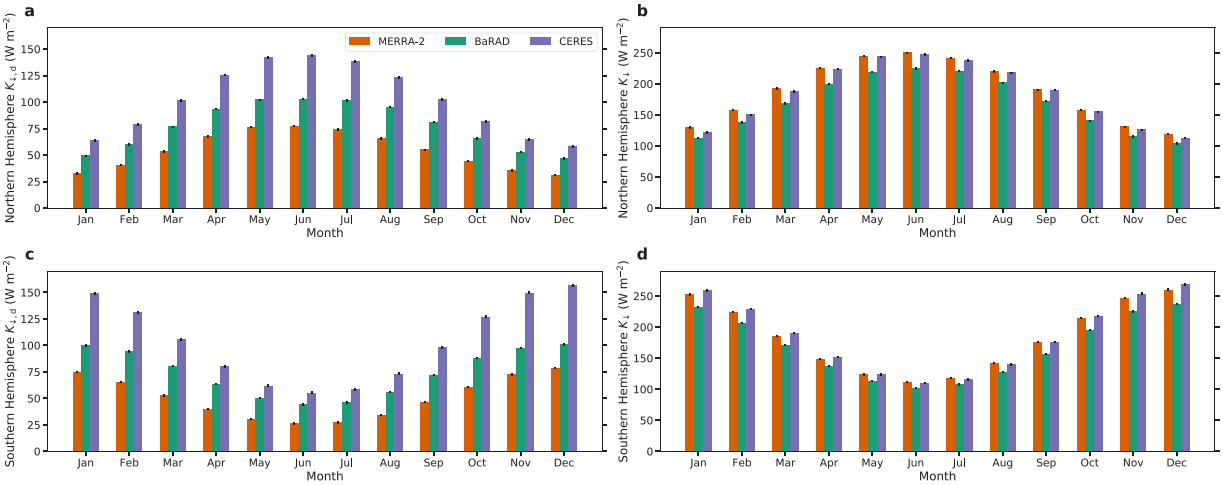

Fig. 3 Seasonal variability in all products. Monthly variability in diffuse radiation ($K_{\downarrow,d}$) in MERRA-2, BaRAD, and CERES for (**a**) the northern hemisphere and (**c**) the southern hemisphere. Sub-figures (**b**) and (**d**) are the same, but for total shortwave radiation ($K_{\downarrow}$). The error bars show the standard errors for each month.

Although the three datasets show broadly similar latitudinal (Figs. 2c and S6c) and spatial patterns (Figs. 2d,e, S6d, and S6e), $K_{\downarrow,d}$ in the BaRAD dataset is higher than in MERRA-2 over the Sahara and India and higher than the CERES data over Australia. For $K_{\downarrow}$, BaRAD shows a lower value than both MERRA-2 and CERES over the tropical region. Figures 2 and S6 also show the mean area-weighted difference ($\Delta K_{\downarrow}$ and $\Delta K_{\downarrow,d}$) between the BaRAD data and the MERRA-2 and CERES products, respectively. The global mean $K_{\downarrow}$ and $K_{\downarrow,d}$ are 167.9 and 75.8 W m^{-2}, respectively, according to BaRAD. In comparison, the global mean $K_{\downarrow}$ is higher at 185.4 W m^{-2} and 185.9 W m^{-2} according to MERRA-2 and CERES respectively, and the global mean $K_{\downarrow,d}$ is lower at 52.6 W m^{-2} according to MERRA-2 and higher at 102.4 W m^{-2} according to CERES.

We calculate the seasonal trends of $K_{\downarrow,d}$ and $K_{\downarrow}$ in the northern and southern hemisphere grids (Fig. 3). Although there are large differences in the magnitude of the three datasets, the seasonal variation is captured by the BaRAD dataset (when compared to the other two). For instance, the highest northern hemisphere averages are during the boreal summer and the lowest values are during the winter; vice versa for the southern hemisphere. These patterns are evident in all the datasets.

We also compare the BaRAD dataset with the newly developed $K_{\downarrow}$ and $K_{\downarrow,d}$ datasets from the EPIC measurements between 2016 and 2019[21]. The EPIC instrument housed on the Deep Space Climate Observatory (DSCOVR) satellite, takes narrow band spectral images of the sunlit face of Earth for 10 channels every 60 to 100 min. The dataset generated by Hao *et al.*[21] is available at 0.1° by 0.1° resolution and is based on a random forest algorithm trained using *in situ* observations and the EPIC-derived variables[37]. Here, we compare the available observations with the BaRAD data for the same period. Although the EPIC-based dataset has several advantages over many existing global estimates of $K_{\downarrow,d}$, namely the much higher spatial and temporal (up to hourly) resolution, it is not ideal for studying climatological trends. The EPIC instrument is affected by cloud cover and downtime. Thus, the EPIC data are interrupted by data gaps, with 5.1% of days missing between 2016 and 2019. Moreover, the product is only available over land. We regridded the EPIC-derived data to the native MERRA-2 resolution using a nearest neighbor interpolation and compared the spatial and latitudinal trends in the $K_{\downarrow,d}$ and $K_{\downarrow}$ with the BaRAD values (Fig. 4). Overall, the global mean $K_{\downarrow,d}$ in BaRAD is very close to the EPIC-derived values, with a mean difference of only −0.72 W m^{-2}. Greater differences are seen for $K_{\downarrow}$ with BaRAD underestimating it by 22.55 W m^{-2}. Many of the differences between the two products occur over Africa, as also seen from the latitudinal trends (Fig. 4b,d). It is important to note that the *in situ* observations used in Hao *et al.*[21] to evaluate the product lacks spatial representation over central Africa, while the GEBA observations are much more frequent here, at least for $K_{\downarrow}$ (Fig. S1). For $K_{\downarrow,d}$, both GEBA and the datasets used in Hao *et al.*[21] are sparse, which could explain the low $\Delta K_{\downarrow,d}$ for this variable.

Validation against baseline surface radiation sites in the tropics. Given the lack of observations in tropical regions and in the southern hemisphere, we examined how the lack of data in those regions affect the BaRAD results. To do so, we processed minute-level observations from the Baseline Surface Radiation Network (BSRN)[38] and found two sites with sufficient (more than 8650 h in a year) observations in these data-scarce regions to evaluate the gridded products (namely MERRA-2, BaRAD, and CERES). The observations are from the GOB (Gobabeb, Namib Desert, Namibia at 23.56° S, 15.04° E) and PTR (Petrolina, Brazil at 9.07° S, 40.32° W) stations, shown as black stars in Fig. S1b. For the GOB site, 2013, 2014, and 2015 are the years with sufficient observations, while for the PT site, the years 2010, 2011, and 2014 are chosen. Note that although the GEBA dataset includes several BSRN sites, these two sites (and some others) are not included since they do not have enough observations to reliably compute monthly means.

Figure 5 shows the seasonal trends of $K_{\downarrow,d}$ and $K_{\downarrow}$ from the available BSRN observations, as well as the corresponding monthly composites from MERRA-2, BaRAD, and CERES. For both the stations, the BaRAD data shows less bias (MBE = 10.49 W m^{-2} for GOB; 2.31 W m^{-2} for PTR) than both MERRA-2 (MBE = −12.54 W m^{-2}

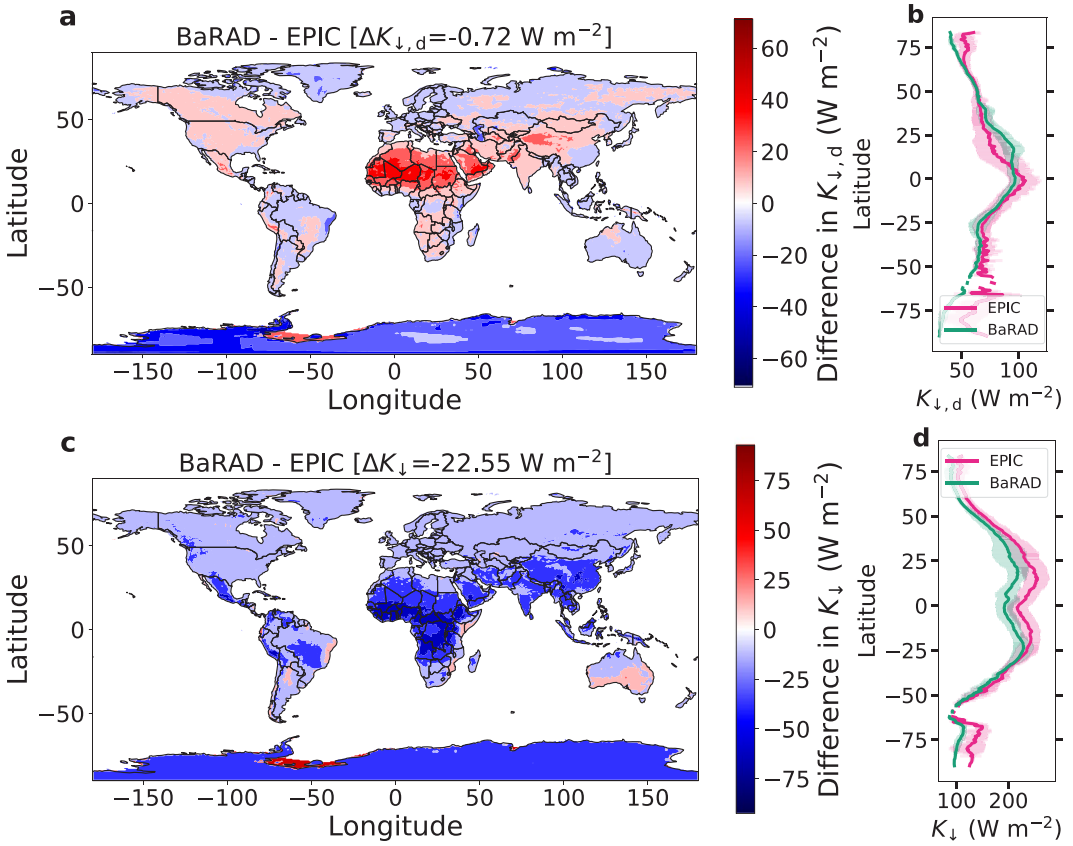

Fig. 4 Comparison of spatial and latitudinal variability in total shortwave radiation and diffuse radiation between the BaRAD product and EPIC-derived estimates. Spatial patterns of the grid-wise difference in (**a**) diffuse radiation ($K_{\downarrow,d}$) and (**b**) total shortwave radiation ($K_{\downarrow}$) over land. Sub-figure (**b**) and (**d**) show the mean latitudinal variability of $K_{\downarrow,d}$ and $K_{\downarrow}$ over land for the two products. The shaded areas represent the standard deviation. The area-weighted difference in $K_{\downarrow,d}$ ($\Delta K_{\downarrow,d}$) and $K_{\downarrow}$ ($\Delta K_{\downarrow}$) between the BaRAD product and the EPIC-derived dataset are shown at the top of sub-figures (**a**) and (**c**), respectively.

for GOB; −37.63 W m^{-2} for PTR) and CERES (MBE = 61.76 W m^{-2} for GOB; 36.33 W m^{-2} for PTR) for $K_{\downarrow,d}$. CERES overestimates $K_{\downarrow,d}$ and MERRA-2 underestimates it compared to observations, which is consistent with the hemispherical results in Fig. 3 and previous estimates. For $K_{\downarrow}$, the results are mixed, with BaRAD performing better than CERES but worse than MERRA-2 (MBE = 6.95, −25.17, and −62.97 W m^{-2} for MERRA-2, BaRAD, and CERES, respectively) at the GOB site and better than MERRS-2 and comparable to CERES (MBE = 48.16, 13.21, and −13.17 W m^{-2}) at PTR. Note that for the GOB site, there is frequently more missing data at night or early morning than during daytime, which would lead to artificially higher annual $K_{\downarrow}$ values than true annual composites, making the comparison with BaRAD seem worse (and vice versa for MERRA-2).

Long-term trends. Figure S7a–d show the 40-year trend in $K_{\downarrow}$ and $K_{\downarrow,d}$ in the MERRA-2 and the BaRAD dataset for the two hemispheres. The two datasets show similar trends for $K_{\downarrow}$ and $K_{\downarrow,d}$, but they are offset by about 20 W m^{-2} for both $K_{\downarrow}$ and $K_{\downarrow,d}$. More importantly, the BaRAD dataset captures the impacts of the two large volcanic eruptions, El Chichón in 1982 and Mount Pinatubo in 1991, on $K_{\downarrow,d}$, particularly in the northern hemisphere (Fig. S7). This is probably because the aerosol, cloud, and radiation fields from the MERRA-2 reanalysis, which is known to capture large volcanic activity[39], are used to create the BaRAD dataset. For the northern hemisphere, the anomaly in $K_{\downarrow}$ from the mean of the previous and subsequent years (1981 and 1983) due to the El Chichón eruption was −1.95 W m^{-2} in MERRA-2 versus −2.81 W m^{-2} in the BaRAD dataset. For the Mount Pinatubo eruption, the $K_{\downarrow}$ anomaly was −1.28 W m^{-2} in MERRA-2 versus −1.39 W m^{-2} in the BaRAD dataset. For northern hemisphere $K_{\downarrow,d}$, there was an increase by 2.67 W m^{-2} in 1982 compared to the average of the values in 1981 and 1983 in MERRA-2 and 2.13 W m^{-2} for BaRAD. Similarly, in 1991, the northern hemisphere $K_{\downarrow,d}$ was higher by 1.75 W m^{-2} compared to 1990 and 1992 in MERRA-2 versus 1.14 W m^{-2} in BaRAD.

Figure 6a, b are two examples of site-level comparison with observations made at Sapporo, Japan (43.05° N, 141.33° E for $K_{\downarrow}$) and Würzburg, Germany (49.77° N, 9.97° E for $K_{\downarrow,d}$). These two sites are chosen because they have the longest data availability. The BaRAD dataset replicates both the magnitude and long-term variability of the site observations (r^2 = 0.99 and MBE = −3.65 W m^{-2} for $K_{\downarrow,d}$; r^2 = 0.97 and MBE = −8.64 W m^{-2} for $K_{\downarrow}$). On the other hand, MERRA-2 captures the variability (r^2 = 0.98 for $K_{\downarrow,d}$; 0.97 for $K_{\downarrow}$), but has larger biases for both $K_{\downarrow,d}$ (MBE = −22.95 W m^{-2}) and $K_{\downarrow}$ (MBE = 16.85 W m^{-2}).

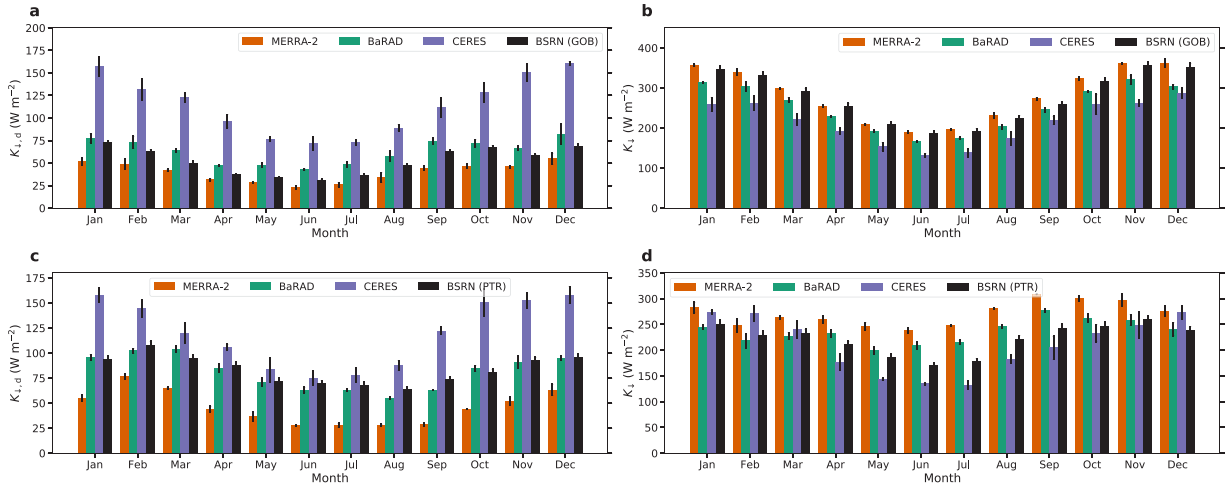

Fig. 5 Validation of BaRAD against BSRN observations. Monthly variability in gridded values in MERRA-2, BaRAD, and CERES, and values observed at the GOB (Gobabeb, Namib Desert, Namibia at 23.56° S, 15.04° E) BSRN station for (**a**) diffuse radiation ($K_{\downarrow,d}$) and (**b**) total shortwave radiation ($K_{\downarrow}$). Sub-figures (**c**) and (**d**) are the same, but for the PTR (Petrolina, Brazil at 9.07° S, 40.32° W) BSRN site. The error bars show the standard errors for each month.

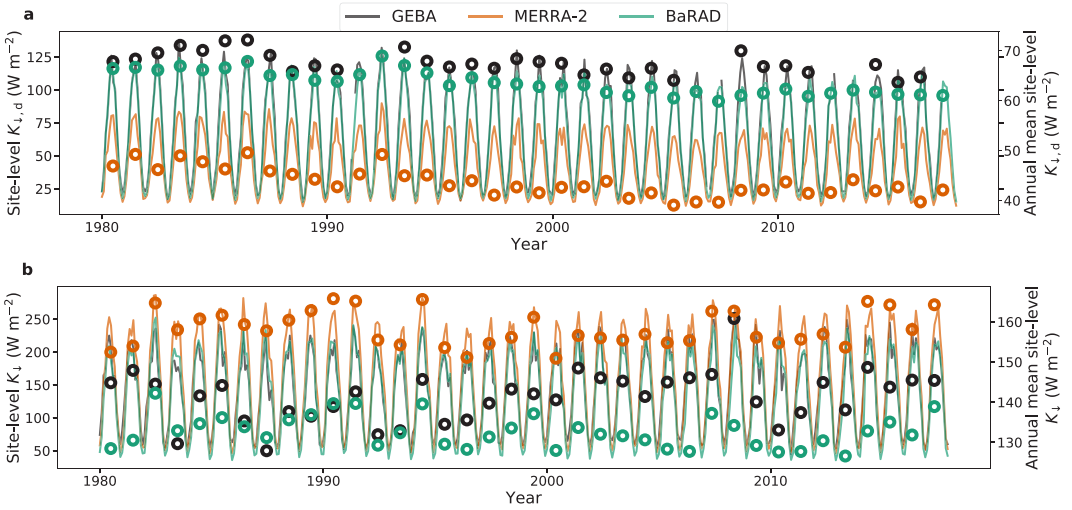

Fig. 6 Long-term trends at site scale. Long-term trends in (**a**) diffuse radiation ($K_{\downarrow,d}$) and (**b**) total shortwave radiation ($K_{\downarrow}$) for GEBA sites with longest archival history, along with corresponding gridded values from MERRA-2 and BaRAD. For $K_{\downarrow,d}$, the site with the longest archival history is located in Würzburg, Germany (49.77° N, 9.97° E) and the site with the longest archival history of $K_{\downarrow}$ is in Sapporo, Japan (43.05° N, 141.33° E). The monthly values are plotted on the left y-axes as lines and the annual averages (plotted as circles) are on the right y-axes.

Limitations and Future work. In the present study, our objective was to compare a conceptual k_t model, a linear model, and an RF model that explicitly considers non-linear interactions to bias-adjust the MERRA-2 $K_{\downarrow}$ and $K_{\downarrow,d}$ fields. Based on the cross-validation, the RF model was used to develop the BaRAD dataset. It should be noted that there are many machine learning architectures that can capture non-linear interactions. A comprehensive cross-validation of all such models is beyond the scope of the present study but should be undertaken in future work. Compared to many of these other architectures, RF models are easier to train, less sensitive to hyperparameters, simple to interpret, and have been used in similar supervised learning problems with similar sample sizes[40]. Although we expect the improvements in bias-adjusted radiation fields to be minor when moving to more complicated machine learning models for the current training data, architectures like deep neural networks are expected to perform better as the training sample size increases. For larger sample sizes, feature selection would also be much more important to optimize training time and further improve accuracy.

Here we focus on monthly means since we have the most comprehensive geographic distribution of radiation observations at this temporal scale through GEBA. As more data are incorporated in this archive, we plan to update the BaRAD dataset. It is possible to generate datasets similar to BaRAD at sub-monthly and even sub-daily

scales, though this requires more comprehensive training data than currently available. Observation networks like BSRN can help in this regard, but it is critical to set up new observation sites to continuously observe $K_{\downarrow,d}$ to reduce sampling biases, especially in tropical regions where $K_{\downarrow,d}$ would have a stronger influence on the terrestrial carbon, energy, and water cycles[23].

Usage Notes

The BaRAD dataset[35] developed here performs well when compared to the GEBA dataset and captures the seasonal, latitudinal, and long-term trends in $K_{\downarrow}$ and $K_{\downarrow,d}$. However, the dataset can be affected by biased sampling in the GEBA dataset. The GEBA dataset is overrepresented in the northern hemisphere, especially in Europe and China[10,28]. A second source of bias is associated with the lack of training data over ocean surfaces. Finally, polar regions are under-sampled by GEBA as noted above. We urge caution when using this dataset over polar regions and ocean surfaces. For land grids in the southern hemisphere, although there are many observations for $K_{\downarrow}$, there are fewer stations with $K_{\downarrow,d}$ measurements. Even though Fig. 5 suggests that the BaRAD $K_{\downarrow,d}$ has less bias than the MERRA-2 dataset for sites not 'seen' by the bias-correction algorithm, when possible, we suggest independent validation of the BaRAD $K_{\downarrow,d}$ data before its applications for southern hemisphere land grids. For basic visualization, we have also developed a Google Earth Engine[41] web application (https://yceo.users.earthengine.app/view/barad), that will allow one to download the time series of monthly diffuse and direct beam radiation for any grid. A summary of the datasets compared in the present study are given in Table S2. For a comprehensive comparison of reanalysis datasets that archive $K_{\downarrow,d}$, see Chakraborty & Lee[10].

Code availability

The scripts used to generate the BaRAD dataset are available in this GitHub repository: https://github.com/TC25/BaRAD/tree/main/Scripts.

Received: 12 April 2021; Accepted: 9 August 2021;
Published online: 15 September 2021

References

1. Gu, L. Response of a Deciduous Forest to the Mount Pinatubo Eruption: Enhanced Photosynthesis. *Science* **299**, 2035 (2003).
2. Mercado, L. *et al.* Impact of changes in diffuse radiation on the global land carbon sink. en. *Nature* **458**, 1014–1017 (2009).
3. Rap, A. *et al.* Enhanced global primary production by biogenic aerosol via diffuse radiation fertilization. *Nature Geoscience* **11**, 640–644 (2018).
4. Markovic, M., Jones, C. G., Winger, K. & Paquin, D. The surface radiation budget over North America: gridded data assessment and evaluation of regional climate models. *International Journal of Climatology: A Journal of the Royal Meteorological Society* **29**, 2226–2240 (2009).
5. Bosilovich, M. G., Robertson, F. R. & Chen, J. Global energy and water budgets in MERRA. *Journal of Climate* **24**, 5721–5739 (2011).
6. Kennedy, A. D. *et al.* A comparison of MERRA and NARR reanalyses with the DOE ARM SGP data. *Journal of Climate* **24**, 4541–4557 (2011).
7. Zhang, X. *et al.* Evaluation of the reanalysis surface incident shortwave radiation products from NCEP, ECMWF, GSFC, and JMA using satellite and surface observations. *Remote Sensing* **8**, 225 (2016).
8. Wild, M. Decadal changes in radiative fluxes at land and ocean surfaces and their relevance for global warming. *Wiley Interdisciplinary Reviews: Climate Change* **7**, 91–107 (2016).
9. Chakraborty, T. & Lee, X. Land Cover Regulates the Spatial Variability of Temperature Response to the Direct Radiative Effect of Aerosols. *Geophysical Research Letters* **46**, 8995–9003 (2019).
10. Chakraborty, T. & Lee, X. Large Differences in Diffuse Solar Radiation Among Current-Generation Reanalysis and Satellite-Derived Products. *Journal of Climate*, **34**, 6635–6650 (2021).
11. Oliveira, P. J. C., Davin, E. L., Levis, S. & Seneviratne, S. I. Vegetation-mediated impacts of trends in global radiation on land hydrology: a global sensitivity study. *Global Change Biology* **17**, 3453 (2011).
12. Lee, K.-T. *et al.* Concentrator photovoltaic module architectures with capabilities for capture and conversion of full global solar radiation. *Proceedings of the National Academy of Sciences* **113**, E8210–E8218 (2016).
13. Zhao, L., Lee, X. & Liu, S. Correcting surface solar radiation of two data assimilation systems against FLUXNET observations in North America. *Journal of Geophysical Research: Atmospheres* **118**, 9552–9564 (2013).
14. Boilley, A. & Wald, L. Comparison between meteorological re-analyses from ERA-Interim and MERRA and measurements of daily solar irradiation at surface. *Renewable Energy* **75**, 135–143 (2015).
15. Trolliet, M. *et al.* Downwelling surface solar irradiance in the tropical Atlantic Ocean: a comparison of re-analyses and satellite-derived data sets to PIRATA measurements. *Ocean Science* **14**, 1021–1056 (2018).
16. Plass, G. N. & Kattawar, G. W. Monte Carlo calculations of light scattering from clouds. *Applied optics* **7**, 415–419 (1968).
17. Zhou, Q., Flores, A., Glenn, N. F., Walters, R. & Han, B. 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**, e0180239 (2017).
18. Frank, C. W. *et al.* Bias correction of a novel European reanalysis data set for solar energy applications. *Solar Energy* **164**, 12–24 (2018).
19. Yang, L. *et al.* Estimating surface downward shortwave radiation over china based on the gradient boosting decision tree method. *Remote Sensing* **10**, 185 (2018).
20. Wei, Y. *et al.* Estimation of surface downward shortwave radiation over China from AVHRR data based on four machine learning methods. *Solar Energy* **177**, 32–46 (2019).
21. Hao, D. *et al.* DSCOVR/EPIC-derived global hourly and daily downward shortwave and photosynthetically active radiation data at 0.1°× 0.1° resolution. *Earth System Science Data* **12**, 2209–2221 (2020).
22. Peng, L. *et al.* Reducing Solar Radiation Forcing Uncertainty and Its Impact on Surface Energy and Water Fluxes. *Journal of Hydrometeorology* **22**, 813–829 (2021).
23. Chakraborty, T., Lee, X. & Lawrence, D. M. Strong local evaporative cooling over land due to atmospheric aerosols. *Journal of Advances in Modeling Earth Systems*, **13**, e2021MS002491 (2021).
24. Randles, C. *et al.* The MERRA-2 aerosol reanalysis, 1980 onward. Part I: System description and data assimilation evaluation. *Journal of Climate* **30**, 6823–6850 (2017).
25. Reichle, R. H. & Liu, Q. Observation-corrected precipitation estimates in GEOS-5 (2014).
26. Reichle, R. H. *et al.* Assessment of MERRA-2 land surface hydrology estimates. *Journal of Climate* **30**, 2937–2960 (2017).
27. Gilgen, H. & Ohmura, A. The global energy balance archive. *Bulletin of the American Meteorological Society* **80**, 831–850 (1999).

28. Wild, M. *et al*. The Global Energy Balance Archive (GEBA) version 2017: a database for worldwide measured surface energy fluxes. *Earth System Science Data* **9**, 601–613 (2017).
29. Iziomon, M., Mayer, H. & Matzarakis, A. Empirical models for estimating net radiative flux: A case study for three mid-latitude sites with orographic variability. *Astrophysics and Space Science* **273**, 313–330 (2000).
30. Jiang, B. *et al*. Empirical estimation of daytime net radiation from shortwave radiation and ancillary information. *Agricultural and Forest Meteorology* **211**, 23–36 (2015).
31. Schwarz, M., Folini, D., Yang, S., Allan, R. P. & Wild, M. Changes in atmospheric shortwave absorption as important driver of dimming and brightening. *Nature Geoscience* **13**, 110–115 (2020).
32. Cronin, T. W. On the choice of average solar zenith angle. *Journal of the Atmospheric Sciences* **71**, 2994–3003 (2014).
33. Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).
34. Wilcoxon, F. in *Breakthroughs in statistics* 196–202 (Springer, 1992).
35. Chakraborty, T. & Lee, X. BaRAD: Bias-Adjusted RADiation Dataset. *PANGAEA* https://doi.org/10.1594/PANGAEA.932924 (2021).
36. Kato, S. *et al*. Surface irradiances of edition 4.0 clouds and the earth's radiant energy system (CERES) energy balanced and filled (EBAF) data product. *Journal of Climate* **31**, 4501–4527 (2018).
37. Hao, D. *et al*. Estimating hourly land surface downward shortwave and photosynthetically active radiation from DSCOVR/EPIC observations. *Remote Sensing of Environment* **232**, 111320 (2019).
38. Driemel, A. *et al*. Baseline Surface Radiation Network (BSRN): structure and data description (1992–2017). *Earth System Science Data* **10**, 1491–1501 (2018).
39. Buchard, V. *et al*. The MERRA-2 aerosol reanalysis, 1980 onward. Part II: Evaluation and case studies. *Journal of Climate* **30**, 6851–6872 (2017).
40. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: data mining, inference, and prediction* (Springer Science & Business Media, 2009).
41. Gorelick, N. *et al*. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment* **202**, 18–27 (2017).

Acknowledgements

This research is supported in part by the US National Science Foundation (grant AGS1933630) and the Yale Institute for Biospheric Studies. We also acknowledge Computational & Information Systems Lab at the National Center for Atmospheric Research (NCAR), the Yale Center for Earth Observation (YCEO) and Microsoft (through an AI for Earth grant) for providing computational resources.

Author contributions

T.C.C. and X.L. designed the research. T.C.C. performed the data analysis and wrote the first draft of the manuscript. X.L. contributed to manuscript revision.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-021-01016-4.

Correspondence and requests for materials should be addressed to T.C.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.