Supplementary Information for:

Quantifying the Impact of Cloud Cover on Solar Irradiance and Environmental Photodegradation

Michelle G. Nevins[a], Jennifer N. Apell[a]*

[a] Department of Civil and Urban Engineering, New York University Tandon School of Engineering, 6 MetroTech Center, Brooklyn, NY 11201, USA

*Email: japell@nyu.edu

Section S1. Data Source

Irradiance and cloud coverage data was provided by the National Renewable Energy Laboratory in Golden, Colorado.[1] Irradiance data was taken from WISER Global and cloud coverage from two sources: ASI-16 Sky Imager and TSI Imager (Figure S1). All data provided a timestamp (year, day of year, and time in MST). EKO WISER global irradiance was given from 290 to 1650 nm but the range of interest was from 290 to 700 nm so anything outside of this range was discarded from consideration for analysis in the main paper.

Under the All Sky Imager (ASI), two algorithms were used to calculate cloud coverage: BRBG and CDOC. The BRBG (blue/red and blue/green) algorithm uses the difference in light scattering between clouds and a clear sky to assign a factor of 0 to 1. This algorithm produces one value to describe the level of cloud coverage in the sky. The CDOC (cloud detection and opacity classification) algorithm identifies clear skies and thin and thick clouds by running across every pixel in the image. CDOC employs a clear sky library to achieve this and removes any pixel effected by sun glare.[2]

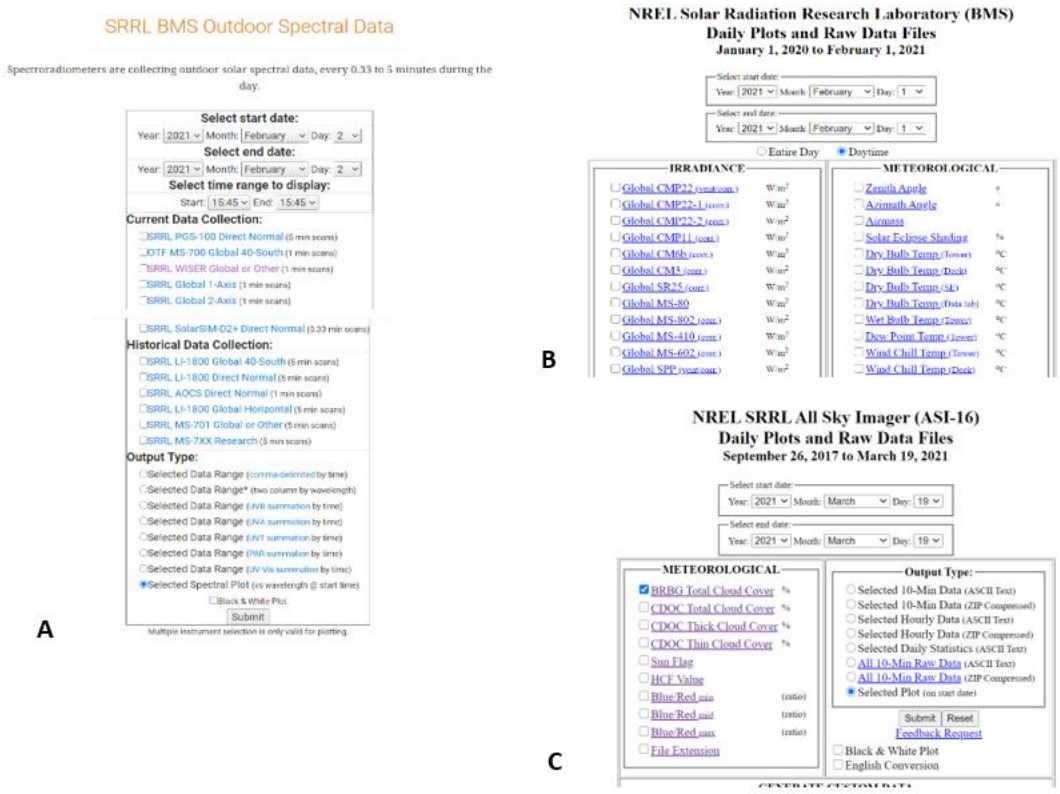

Figure S1. Webpage screenshot for A) WISER irradiance[3], B) Total Sky Imager[4], and C) All Sky Imager data access.[5]

Section S2. Daily Irradiance Calculations

Daily irradiance and cloud coverage values were calculated from the merged data set that had measurements in 10 minute intervals. However, data was not always available every 10 minutes. Therefore, the actual time interval between measurements was calculated using the following equation:

$$\Delta t = \frac{t_{i+1} - t_{i-1}}{2} \quad \text{(S1)}$$

The daily cloud cover was calculated by multiplying each measurment by the length of the time interval, summing the measurements on a given day, and dividing by the length of the daytime measurements using the following equation:

$$\text{daily cloud cover} = \frac{\sum_{day}(\text{cloud cover} \times \Delta t)}{(t_{final} - t_{initial} + 20\ \text{min})} \quad \text{(S2)}$$

In contrast, the daily irradiance was calculated by dividing by the length of the entire day using the following equation:

$$\text{daily irradiance} = \frac{\sum_{day}(\text{irradiance} \times \Delta t)}{24\ \text{hr}} \quad \text{(S3)}$$

The different methods were used to more accurately represent "true" values. There are actual values of cloud cover during the night, but these clouds do not impact values of sunlight irradiance, which is why the cloud cover day was divided by the time interval that measurements were taken. On the other hand, the sunlight irradiance during the night is taken to be zero, but a daily irradiance value should reflect the irradiance received during an entire day.

On some days, the instruments were not all operational and incomplete data was collected. Days where <40 measurements were taken were removed from the data set (n = 47 days). The final data set had values for 1,084 days (Figure S2). The irradiance values are in good agreement for the range of daily irradiance modeled for clear sky days (Table S1).

Table S1. Range of daily irradiance values measured for the studied irradiance regions: UVB, UVA, and PAR.

Irradiance Regions	**Irradiance (W/m^2)**
UVB	0.05-0.6
UVA	5-20
PAR	50-160

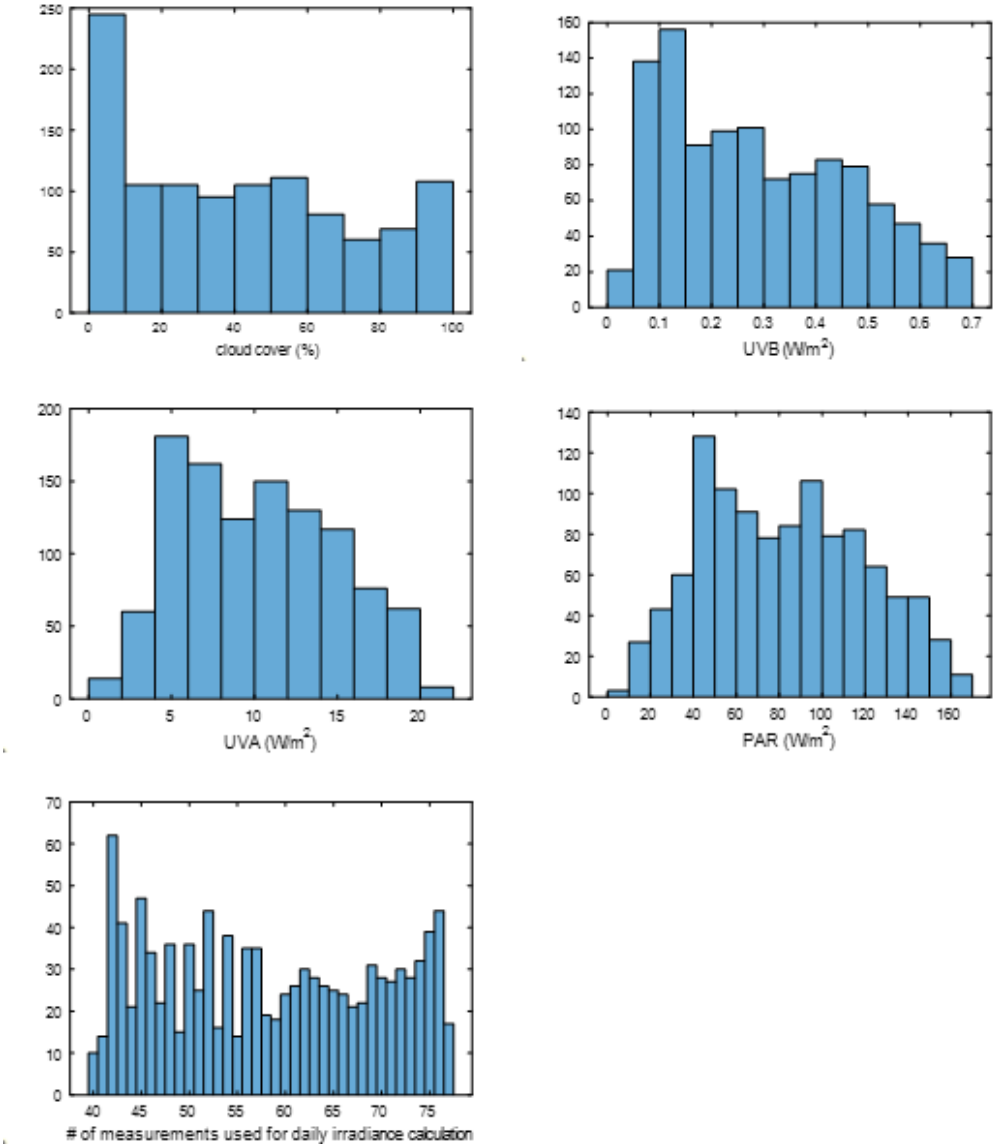

Figure S2. Histograms of daily values calculated after removal of days with <40 measurements.

Section S3. Data Transformation and Normalization

The final data set contained 1) year, 2) day of year, 3) time of day in solar fractional hour, 4) BRBG cloud coverage, 5) CDOC cloud coverage, 6) TSI cloud coverage, and 7) respective daily irradiances (UVB, UVA, and PAR).

Solar Time. The NREL spectroradiometer irradiance data listed output in respect to Mountain Standard Time (MST), which was converted to solar time based on the year, day of year, and watch time. Using the solar time conversion equations from National Oceanic and Atmospheric Administration (Figure S3), the local standard meridian time, variation of the local solar time, and eccentricity of Earth's orbit and axial tilt were all accounted for in respect to the spectroradiometer location in Colorado.[6] The resulting shifts had small deviations from MST because Golden, CO fell closely to the local standard meridian time (105^{o} W verse 105.18^{o}W for the spectroradiometer location). The solar and watch time generally differed by a maximum of ±15 minutes and differed a minimum of 0 minutes.

General Solar Position Calculations
NOAA Global Monitoring Division

First, the fractional year (γ) is calculated, in radians.

$$\gamma = \frac{2\pi}{365} * (\text{day\_of\_year} - 1 + \frac{hour - 12}{24})$$

(For leap years, use 366 instead of 365 in the denominator.)

From γ, we can estimate the equation of time (in minutes) and the solar declination angle (in radians).

$$eqtime = 229.18*(0.000075 + 0.001868\cos(\gamma) - 0.032077\sin(\gamma) - 0.014615\cos(2\gamma) - 0.040849\sin(2\gamma))$$

$$decl = 0.006918 - 0.399912\cos(\gamma) + 0.070257\sin(\gamma) - 0.006758\cos(2\gamma) + 0.000907\sin(2\gamma) - 0.002697\cos(3\gamma) + 0.00148\sin(3\gamma)$$

Next, the true solar time is calculated in the following two equations. First the time offset is found, in minutes, and then the true solar time, in minutes.

$$time\_offset = eqtime + 4*longitude - 60*timezone$$

where eqtime is in minutes, longitude is in degrees (positive to the east of the Prime Meridian), timezone is in hours from UTC (U.S. Mountain Standard Time = –7 hours).

$$tst = hr*60 + mn + sc/60 + time\_offset$$

where hr is the hour (0 - 23), mn is the minute (0 - 59), sc is the second (0 - 59).

The solar hour angle, in degrees, is:

$$ha = (tst / 4) - 180$$

The solar zenith angle (ϕ) can then be found from the hour angle (*ha*), latitude (*lat*) and solar declination (*decl*) using the following equation:

$$\cos(\phi) = \sin(lat)\sin(decl) + \cos(lat)\cos(decl)\cos(ha)$$

And the solar azimuth (θ, degrees clockwise from north) is found from:

$$\cos(180 - \theta) = -\frac{\sin(lat)\cos(\phi) - \sin(decl)}{\cos(lat)\sin(\phi)}$$

Figure S3. Solar time documentation from NOAA used to convert MST observations from NREL to solar time.[6]

Cyclical Variables. A cyclical variable is a feature that repeats after several rotations. In this data set both day of year and time of day were cyclical variables that had to be transformed to improve the accuracy of the model. This was executed by transforming data points for time of day (TOD) and day of year (DOY) into a circle that gave every data point within the day or year their own unique identifier. This transformation separated every time and day point from one to two dimensions by splitting the original value to a cosine and sine value with equations S4 and S5, respectfully. The maximum day was 365 or 366 depending on a normal or leap year and maximum hour was 24.

$$Solar\ Time_{Normalized} = \frac{\sin \text{ or } \cos(2\pi * hour)}{\max(hour)} \quad \text{(S4)}$$

$$Day\ of\ Year_{Normalized} = \frac{\sin \text{ or } \cos(2\pi * (day + 12))}{\max(day)} \quad \text{(S5)}$$

The addition of 12 in DOY equation S5 more accurately aligned the values with solstices and equinoxes. The final transformed data set was between -1 and 1 and gained no further improvements by shifting the data to contain only positive values.

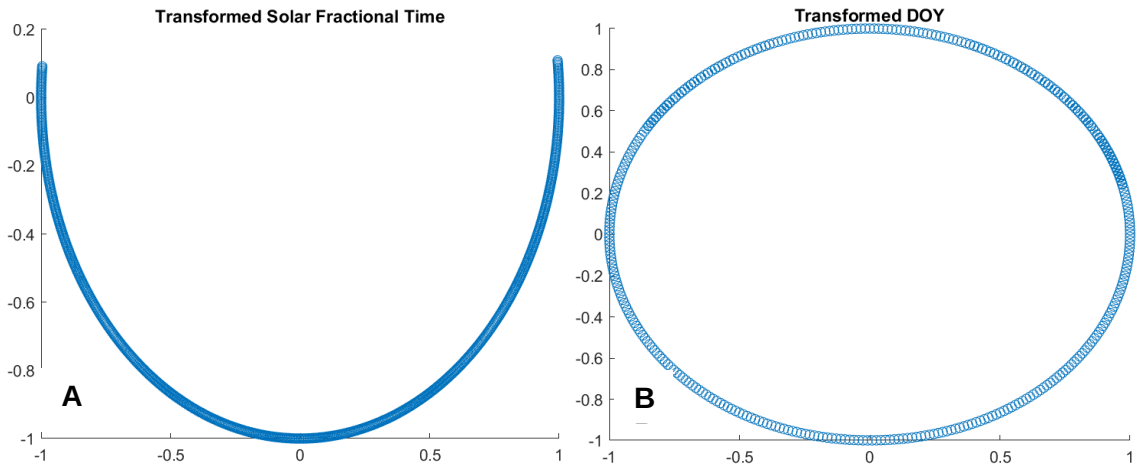

Figure S4. Transformation of cyclical variables a) time of day, and b) day of year (DOY). Time of day is an incomplete circle because measurements were not recorded for irradiance when the sun was low on the horizon or it was dark outside.

Cloud Coverage. The two imagers used at NREL to quantify cloud coverage and for the purpose of this study were the SRRL All-Sky Imager (ASI-16) and the SRRL Total-Sky Imager (TSI-880). Both imagers analyzed clouds between zenith angles of 0 to 80 degrees, and therefore did not consider clouds low on the horizon. Depending on the model, cloud coverage data was normalized as referenced in the main paper.

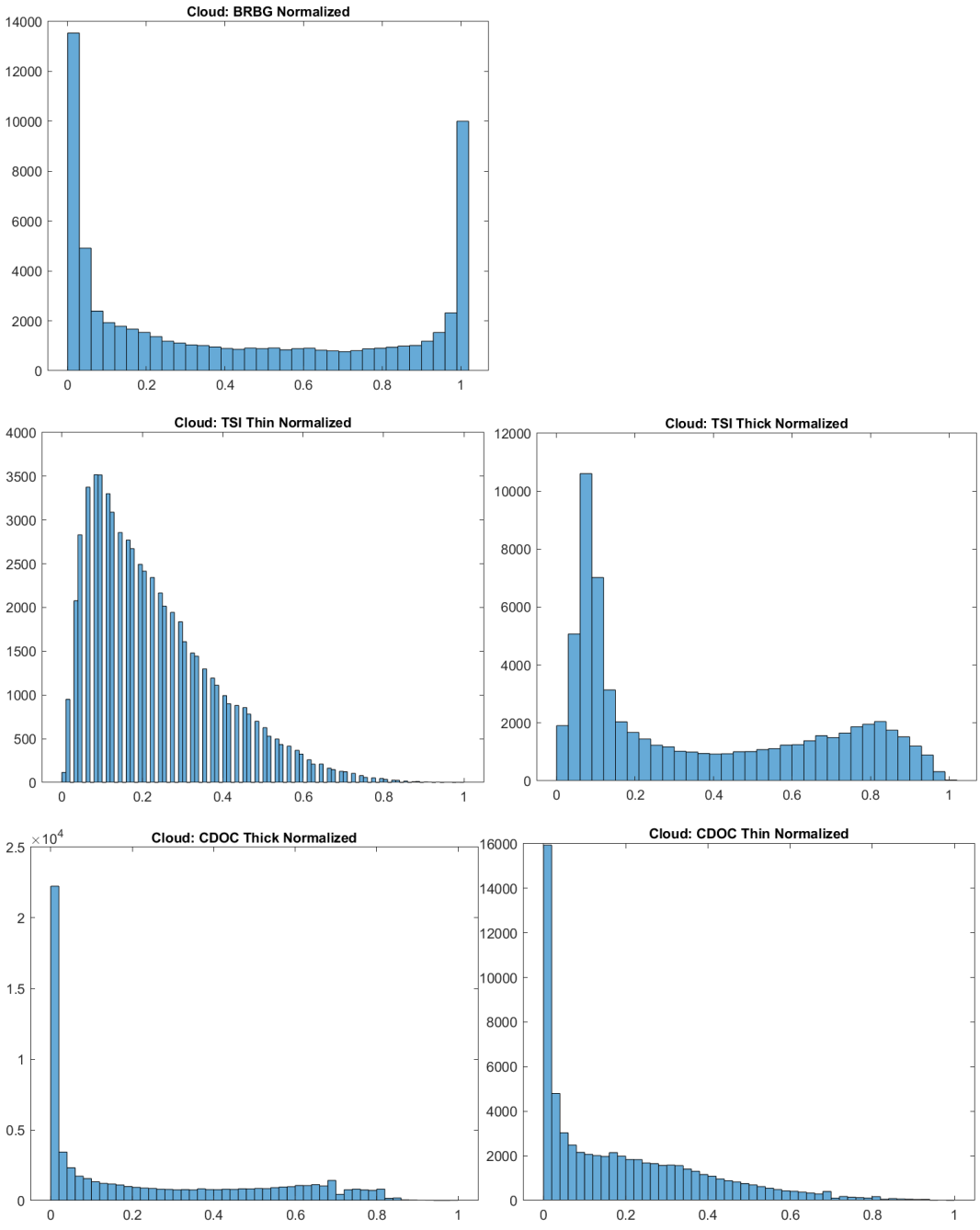

Figure S5. Normalized cloud coverage data histogram statistics.

Irradiance. The three irradiance types analyzed were UVB (ultraviolet B, 290-315 nm), UVA (ultraviolet A, 316-400 nm), and PAR (photosynthetic active radiation, 401 to 700 nm). All irradiance data were log_{10}-transformed, and the data was normalized. The log_{10}-transformation was used if improvement in the model performance was observed.

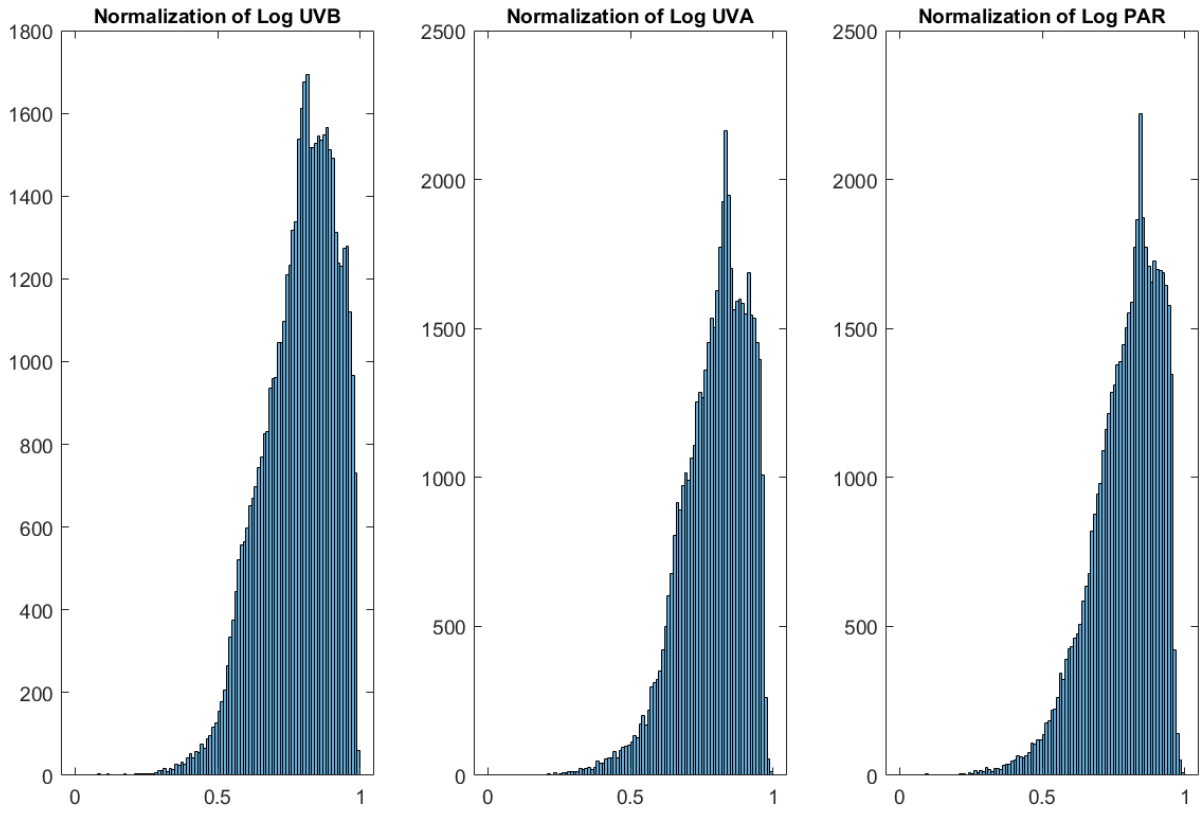

Figure S6. Histogram statistics on irradiance data. Original irradiance data was skewed to the right and after logarithmic transformation irradiance data were more normally distributed but did skew to the left.

Section S4. Multiple Linear Regression Model Development and Testing

Multiple linear regression was performed with MATLAB using transformed and normalized data. Input variables were time of day (sine and cosine), day of year (sine and cosine), and BRBG, CDOC, or TSI cloud data to predict UVB, UVA, and PAR irradiance ($\log_{10}$ transformed). The regression was trained on 68% of the NREL data set (training set), which gave respective coefficients to solve for irradiance. The validation set, 11% of the data, and the remaining 21%, the test set, were then run through the model equations to compare the test model prediction with observed irradiance values. Because multiple linear regression does not need a validation set, unlike the machine learning models, the validation set ultimately served as a second test set. The measured versus the modeled irradiances were plotted for visual comparison and the relative root mean square error (rRMSE) was calculated to quantitatively summarize prediction error from multiple linear regression to random forest and shallow neural networks. The rRMSE values used to compare with the machine learning models came from the test set and not the validation set. The rRMSE results helped to determine how useful normalization and transformations were to model performance.

Section S5. Random Forest Model Development, Algorithm Selection, and Optimization

Initial Optimization. A random forest model consists of a set number of decision trees whose results are averaged to determine a final ensemble modeled value. The ensemble method can either use least-squares boosting (LSBoost) or bootstrap aggregation (Bag). Each decision tree can have a maximum number of branches (number of decision splits) and minimum leaf size (number of data points that must be in the final split groups). When LSBoost is used, the learning rate can also be set. The values of these hyperparameters affect the performance of the model.

To optimize the model, an initial automated hyperparameter optimization was performed in MATLAB to determine a range of reasonable hyperparameter values. Then a grid search was used to determine performance over this range using the values in Table S2.

Table S2. Random forest hyperparameter optimization.

Hyperparameter	Grid Search Values
Ensemble aggregation method	LSBoost or Bag
Minimum leaf size	1, 2, 10, 20, 50
Maximum number of branches (splits)	50, 100, 300, 500, 800
Number of trees in ensemble	10, 30, 60, 100, 150, 200, 300, 500
Learning rate	0.05, 0.1, 0.3, 0.6, 1

Increasing the number of the leaf size, branches, or trees allows for more flexibility in the model fitting but does make the model more susceptible to overfitting. To test the performance at each of these hyperparameter configurations, the root mean square error (RMSE) of the model predicted values and measured values for the validation set were calculated and the R^2 of the linear regression was determined. Specifically, the first 68% of the time series was used to train the model (trained with wavelengths from 290 to 800 nm) and the next 11% was used for model performance validation.

Using LSBoost consistently resulted in better model performance (Figures S7-S27). Predictions for UVB data were always the most accurate (R^2 = 0.72-0.88) compared to UVA (R^2 = 0.57-0.82) and PAR (R^2 = 0.47 to 0.76). Using no cloud data resulted in the worst performances (R^2 = 0.47-0.72). Using the TSI cloud data resulted in the overall best performances (R^2 = 0.76-0.88 and rRMSE 25.5-29.3%). The BRBG cloud data also performed well (R^2 = 0.76-0.88 and rRMSE 26.5-31.0%), and the performance using the CDOC cloud data still improved the model but not as much as the other cloud data sources (R^2 = 0.74-0.88 and rRMSE 27.7-31.5%). The best-performing hyperparameters were similar throughout the model runs but did change slightly when modeling different spectral regions and particularly for models where no cloud data was included. The higher min leaf size and lower number of max branches probably reduced overfitting.

Table S3. Random forest optimal parameters initial grid search.

Optimal Hyperparameters – Initial Grid Search				
Ensemble	Min Leaf Size	Max Branches	Num of Trees	Learning Rate
LSBoost	10, 50	50, 100	30, 60, 100	0.05, 0.1
	cloud data	**UVB**	**UVA**	**PAR**
RMSE	TSI	0.145	5.31	61.9
R^2	TSI	0.88	0.82	0.76

K-Fold Cross-Validation. To further optimize and validate the random forest models for UVB, UVA, and PAR data, K-fold cross-validation with 5 folds was used. The training data set was split into 5 chronologically continuous sets so that the predicted results were not influenced by the occurrence of neighboring data points in the training and validation sets. The models were then re-run with a smaller range of hyperparameters (Figures S7-S27). Cloud data from the CDOC algorithm was not re-run as it had the worst performance in the initial grid search.

Table S4. Random forest optimal hyperparameters.

Optimal Hyperparameters – K-Fold Cross-Validation			
NO CLOUD	**UVB**	**UVA**	**PAR**
Min Leaf Size	60	60	30
Max Branches	50	50	50
Num of Trees	70	70	70
Learning Rate	0.05	0.05	0.05
R^2	0.69	0.56	0.47
RMSE	0.31±0.09	9.78±1.77	108.2±18.7
TSI	**UVB**	**UVA**	**PAR**
Min Leaf Size	60	60	60
Max Branches	125	125	125
Num of Trees	130	100	130
Learning Rate	0.05	0.05	0.05
R^2	0.89	0.84	0.78
RMSE	0.18±0.04	5.84±0.97	68.3±10.5
BRBG	**UVB**	**UVA**	**PAR**
Min Leaf Size	60	60	60
Max Branches	50	75	75
Num of Trees	130	100	100
Learning Rate	0.05	0.05	0.05
R^2	0.87	0.82	0.75
RMSE	0.19±0.05	6.27±1.36	73.2±15.8

Final Optimization of Random Forest Regression. Although using the TSI data (which include values for opaque and non-opaque cloud cover) provided the best performance, the random forest modeling using the BRBG data was chosen because it is easier to understand the results and had similar performance. The final models used the following parameters because they either provided the best performance or 2nd best performance. The final calculation of test statistics was calculated using 7 chronologically continuous folds (i.e., 21% of the data was left out to calculate the model performance and error calculations).

Table S5. Random forest final model optimization.

Parameter	**Parameter Value**
Min Leaf Size	70
Max Branches	100
Num of Trees	120
Learning Rate	0.05

Table S6. Random forest RMSE and R^2 values for irradiance regions (UVB, UVA, and PAR).

Irradiance Region	**RMSE (W/m^2)** (n=7)	**R^2**
UVB	0.192 (0.096 – 0.289)	0.878 (0.839 – 0.911)
UVA	6.23 (3.92 – 8.90)	0.818 (0.766 – 0.866)
PAR	73.0 (49.1 – 100.1)	0.756 (0.705 – 0.802)

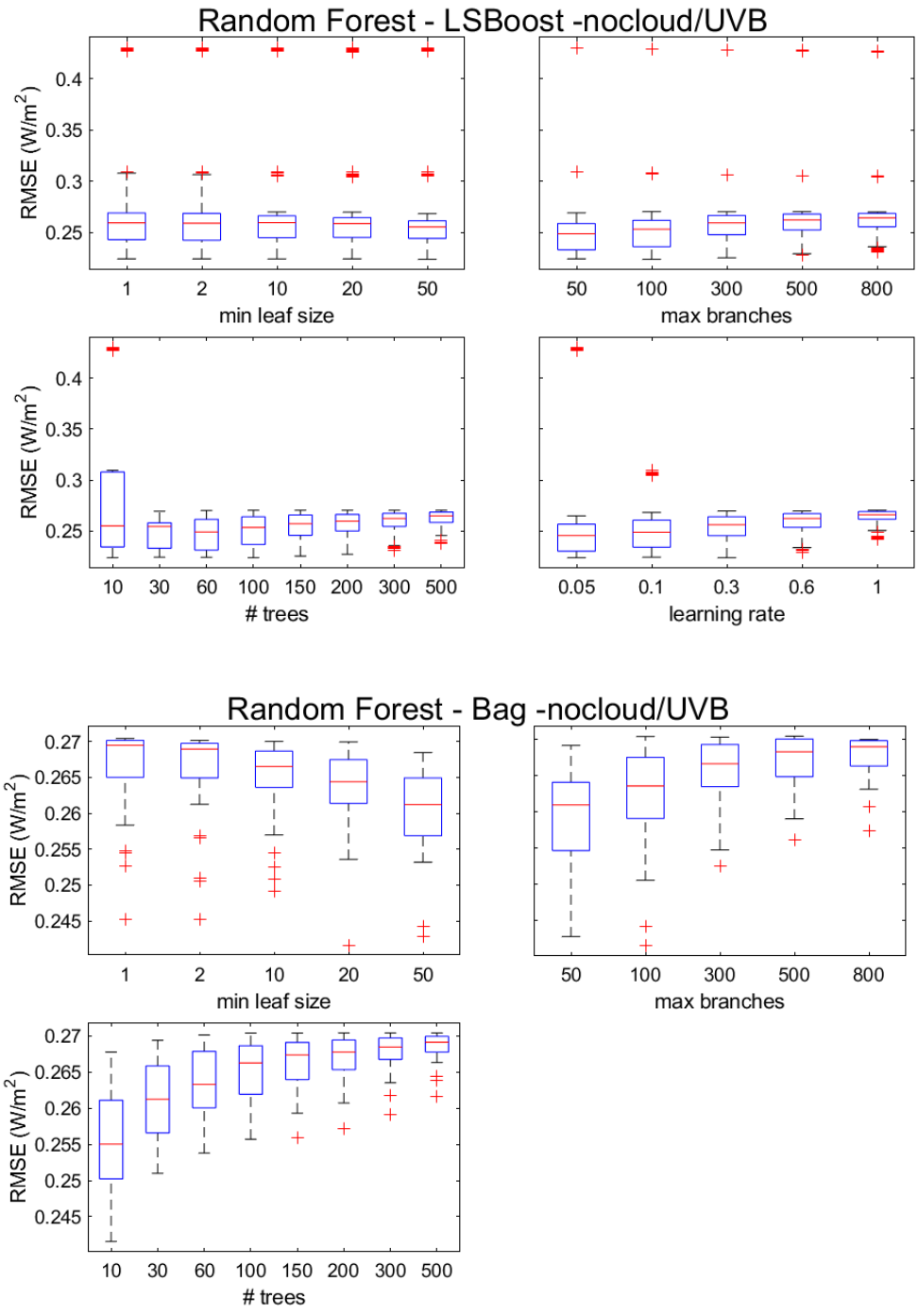

Figure S7. UVB results; no cloud data used

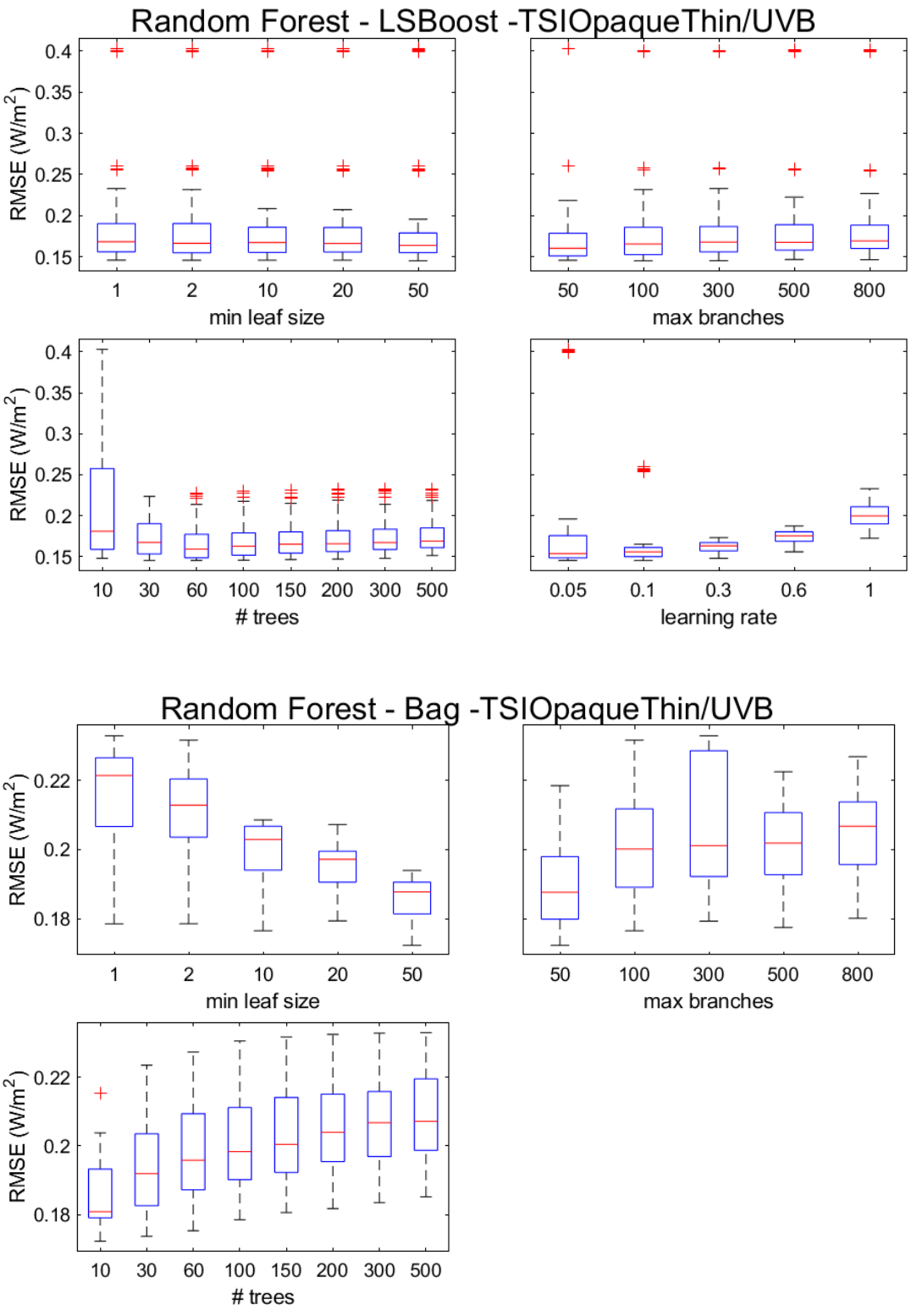

Figure S8. UVB results; TSI cloud data used**.**

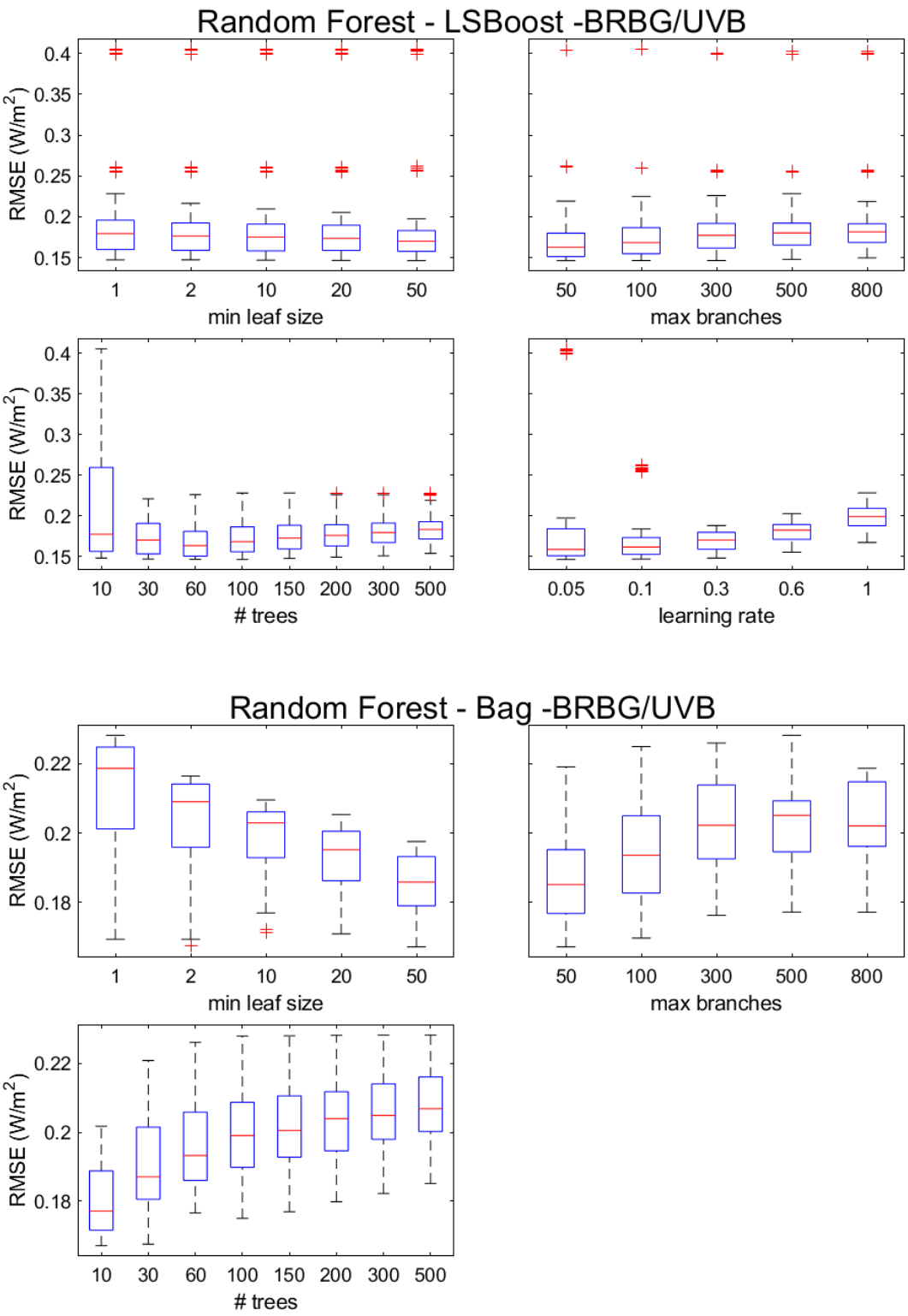

Figure S9. UVB results; BRBG cloud data used**.**

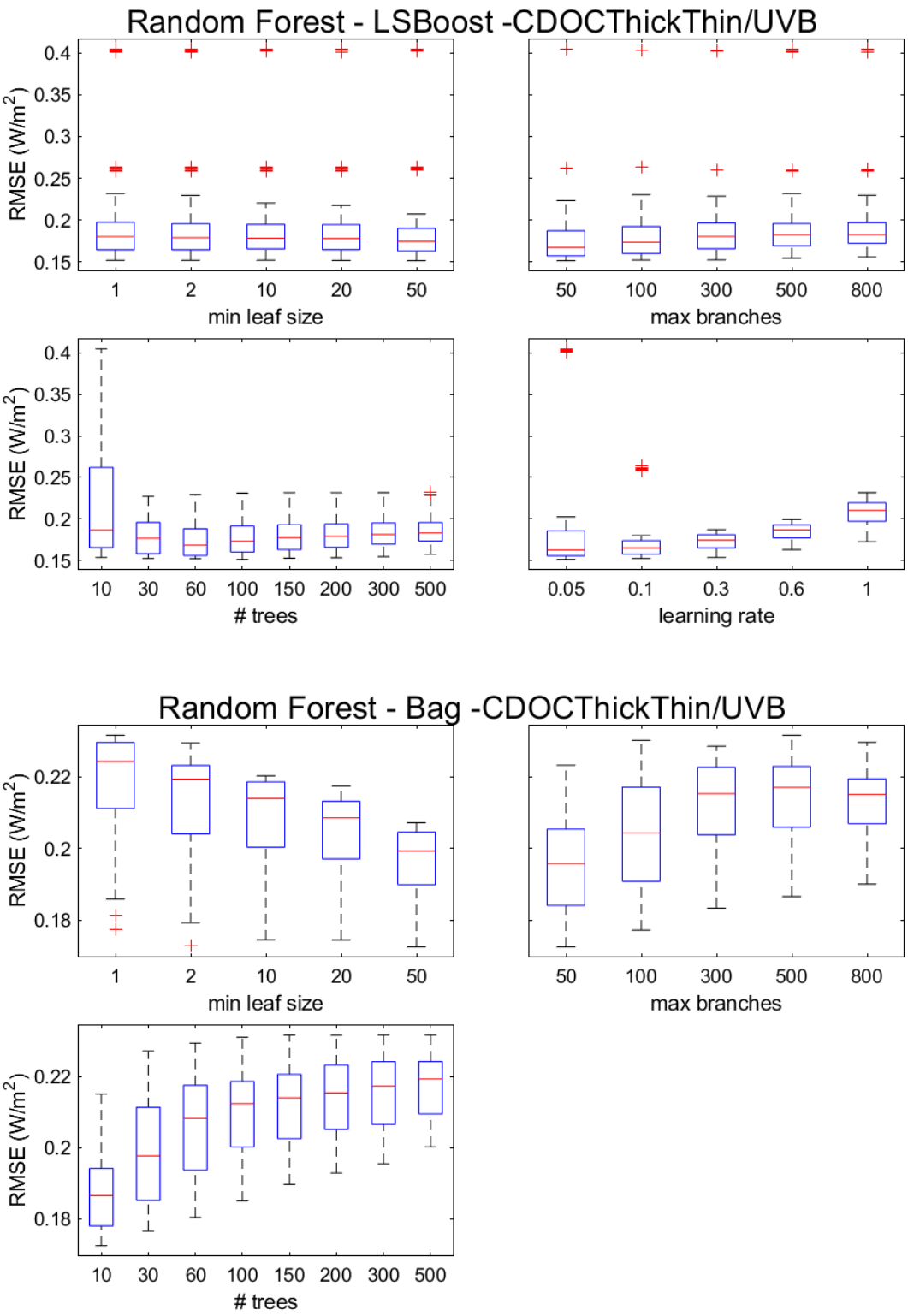

Figure S10. UVB results; CDOC cloud data used.

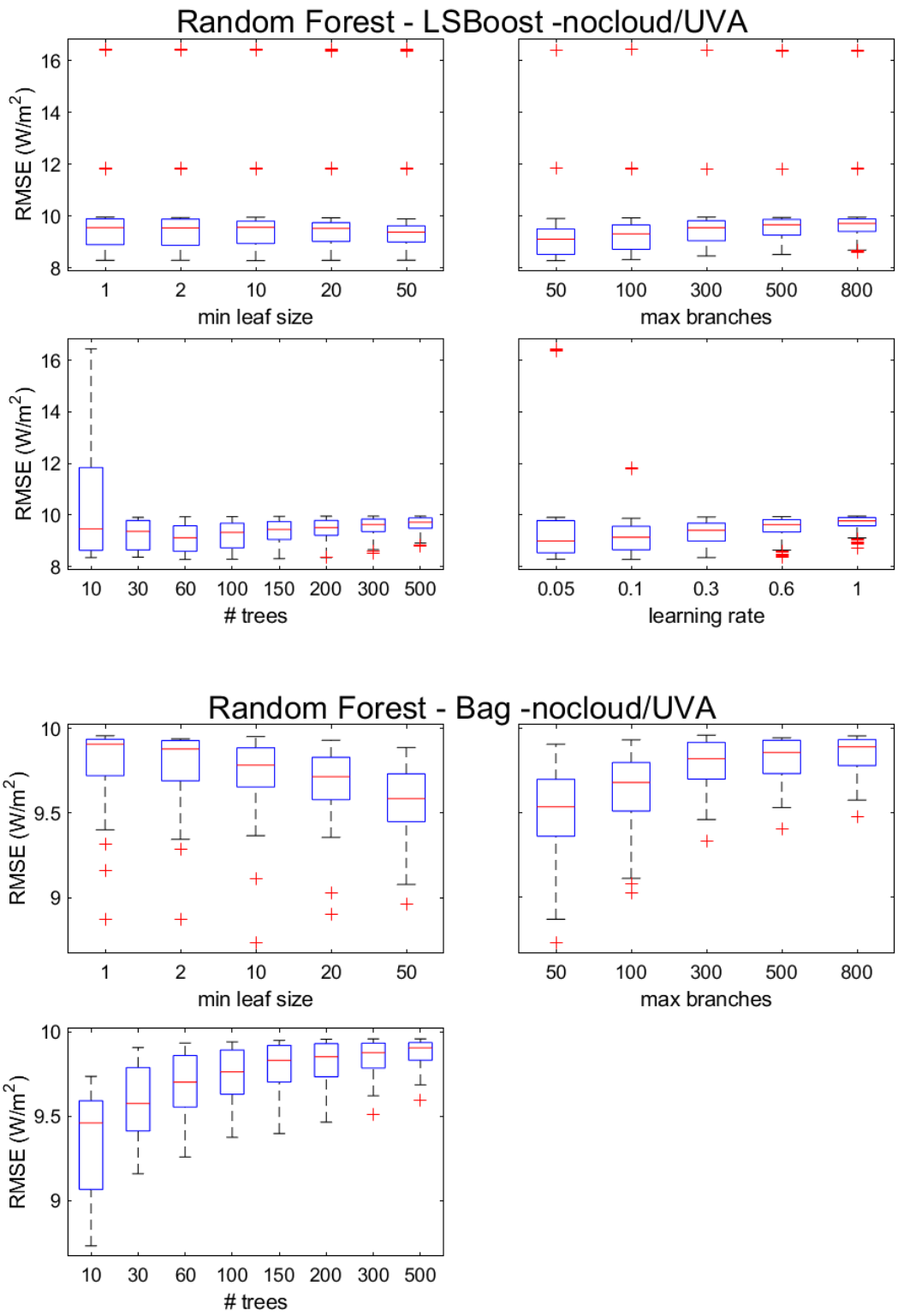

Figure S11. UVA results; no cloud data used.

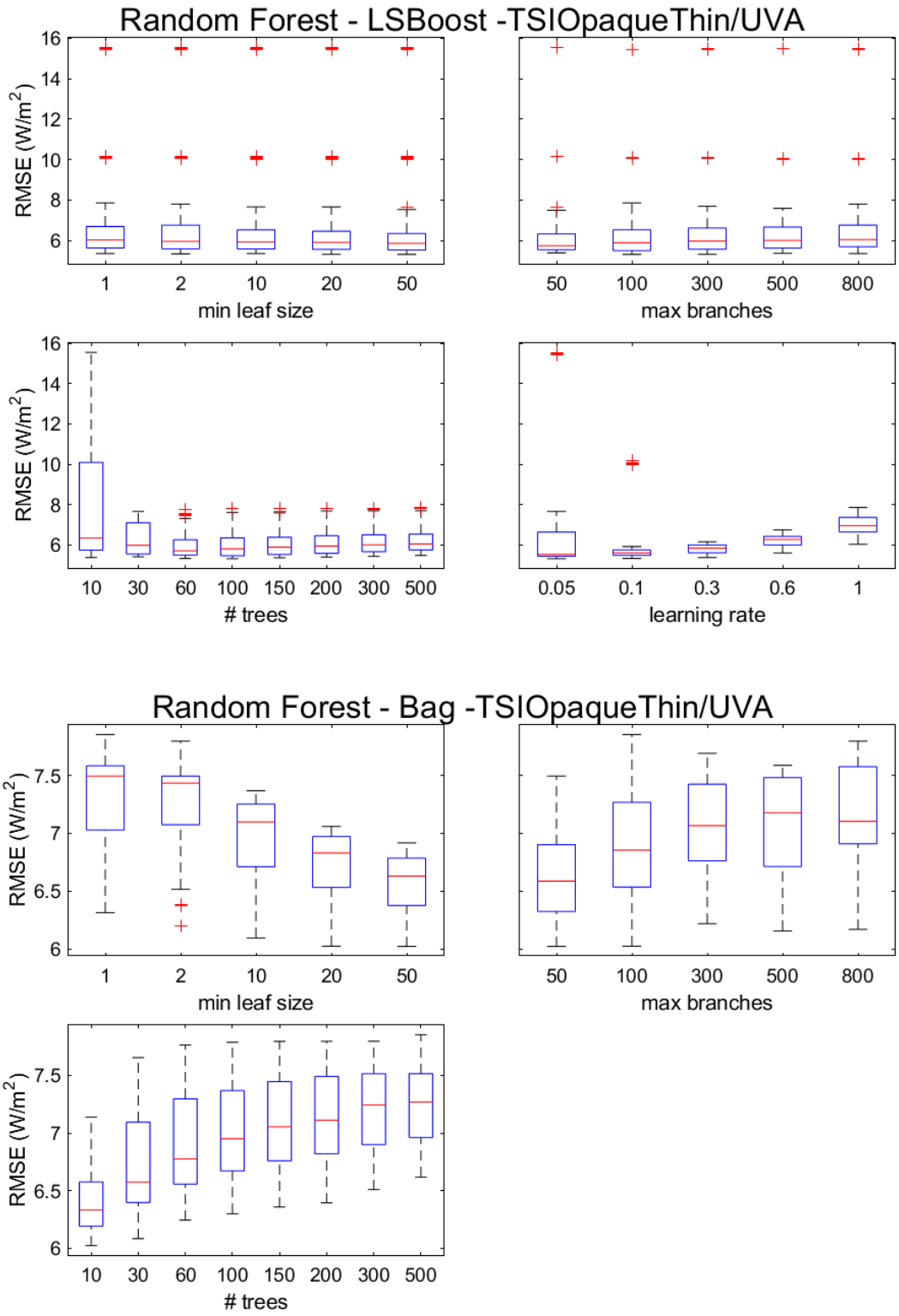

Figure S12. UVA results; TSI cloud data used.

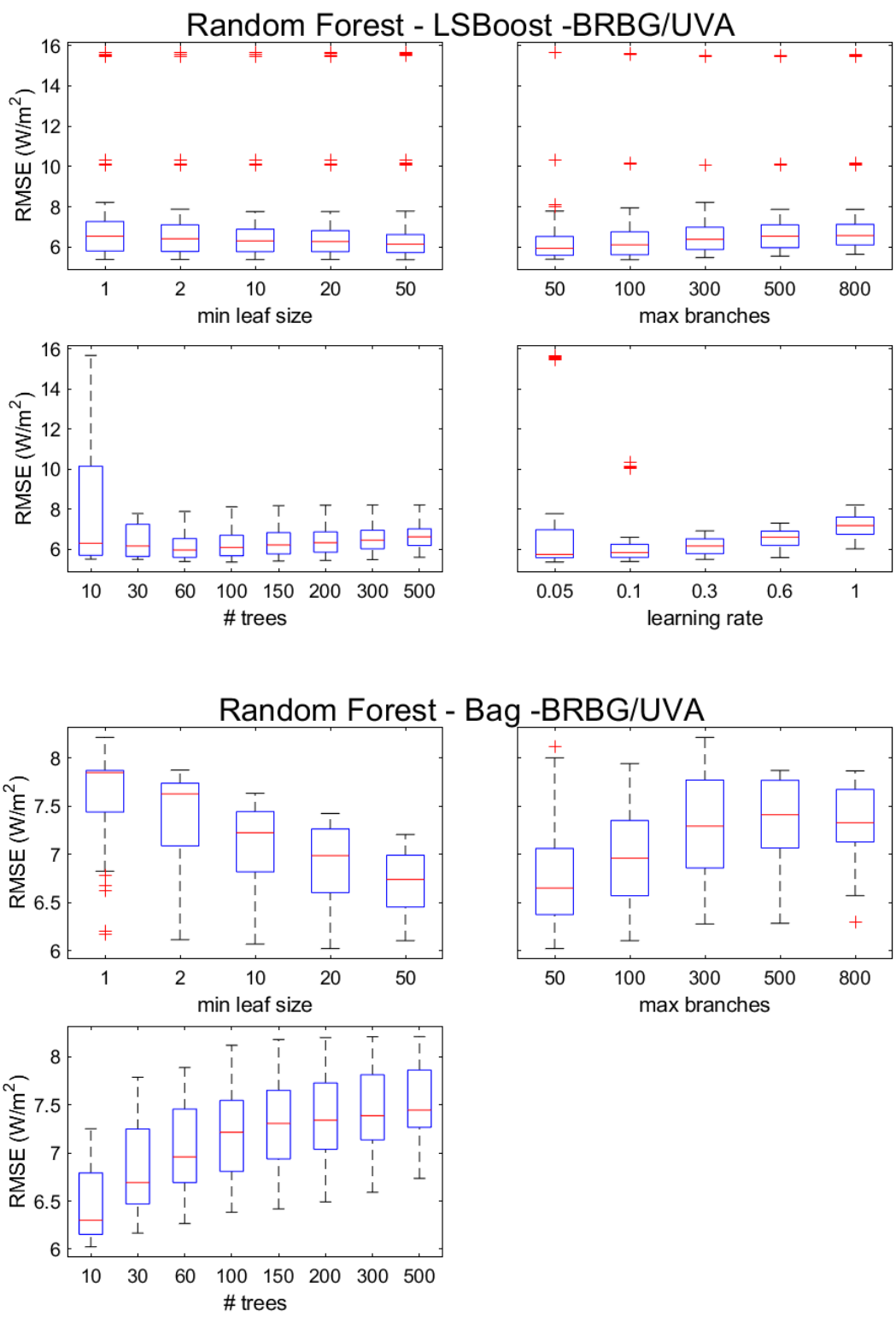

Figure S13. UVA results; BRBG cloud data used.

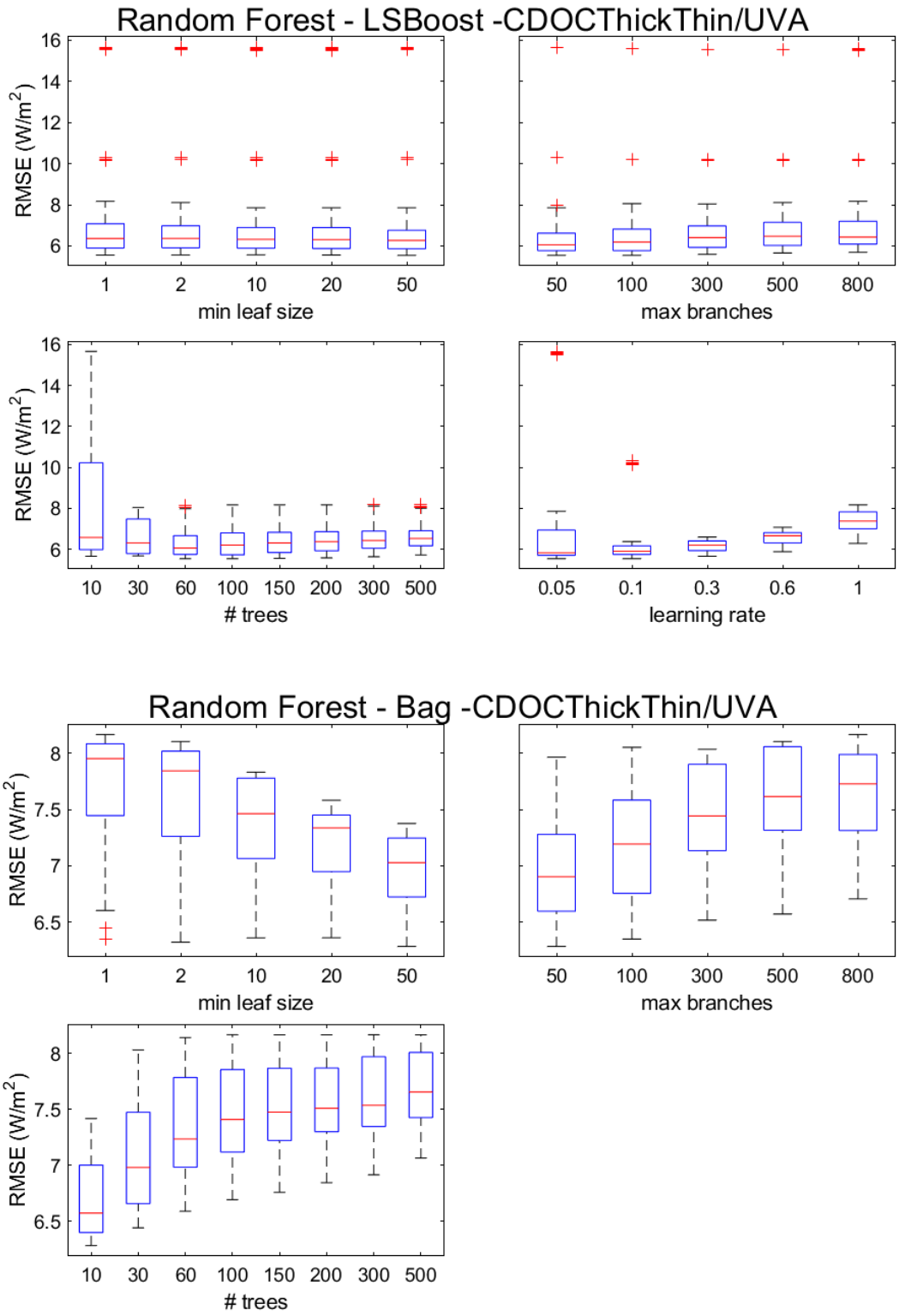

Figure S14. UVA results; CDOC cloud data used.

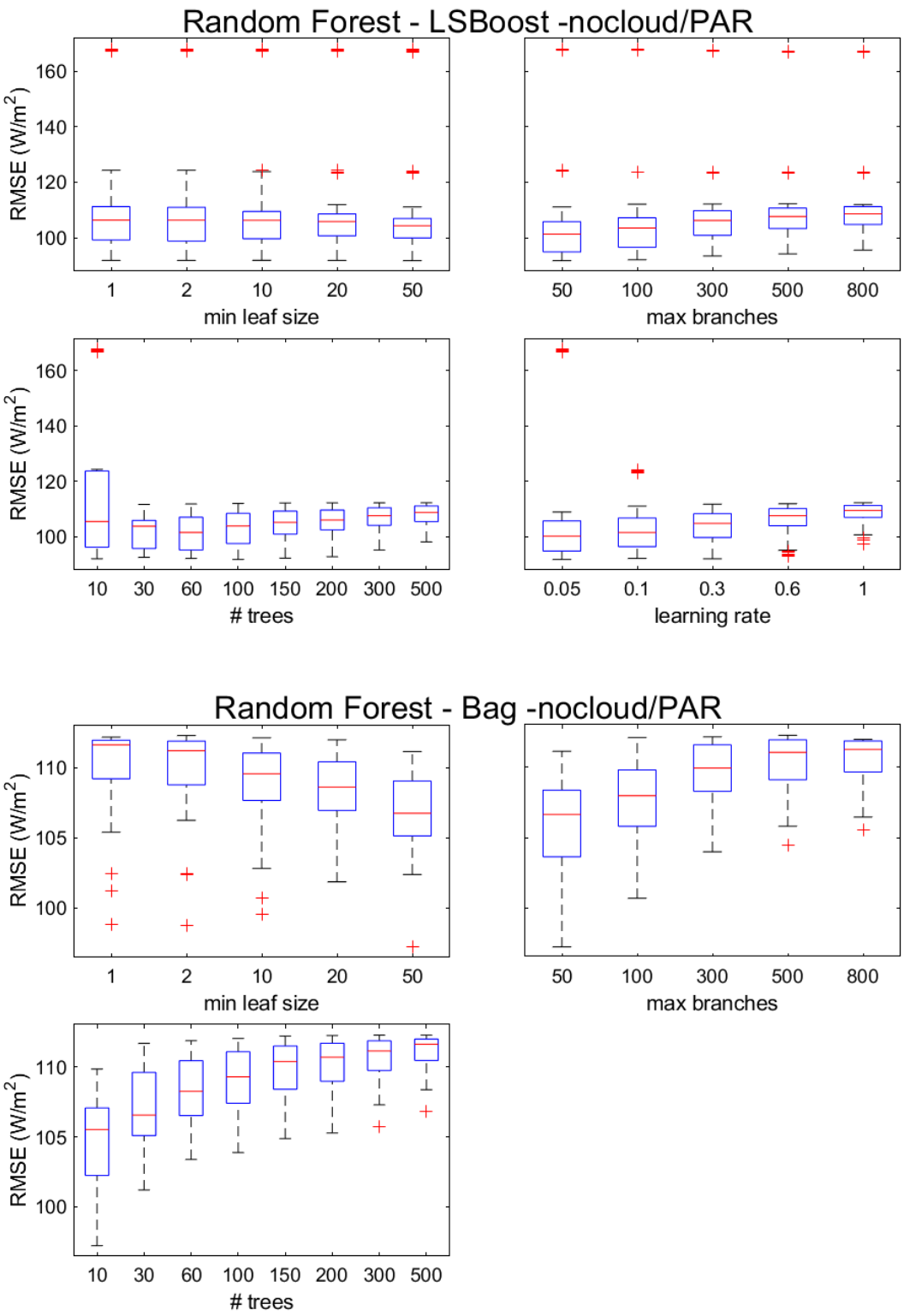

Figure S15. PAR results; no cloud data used.

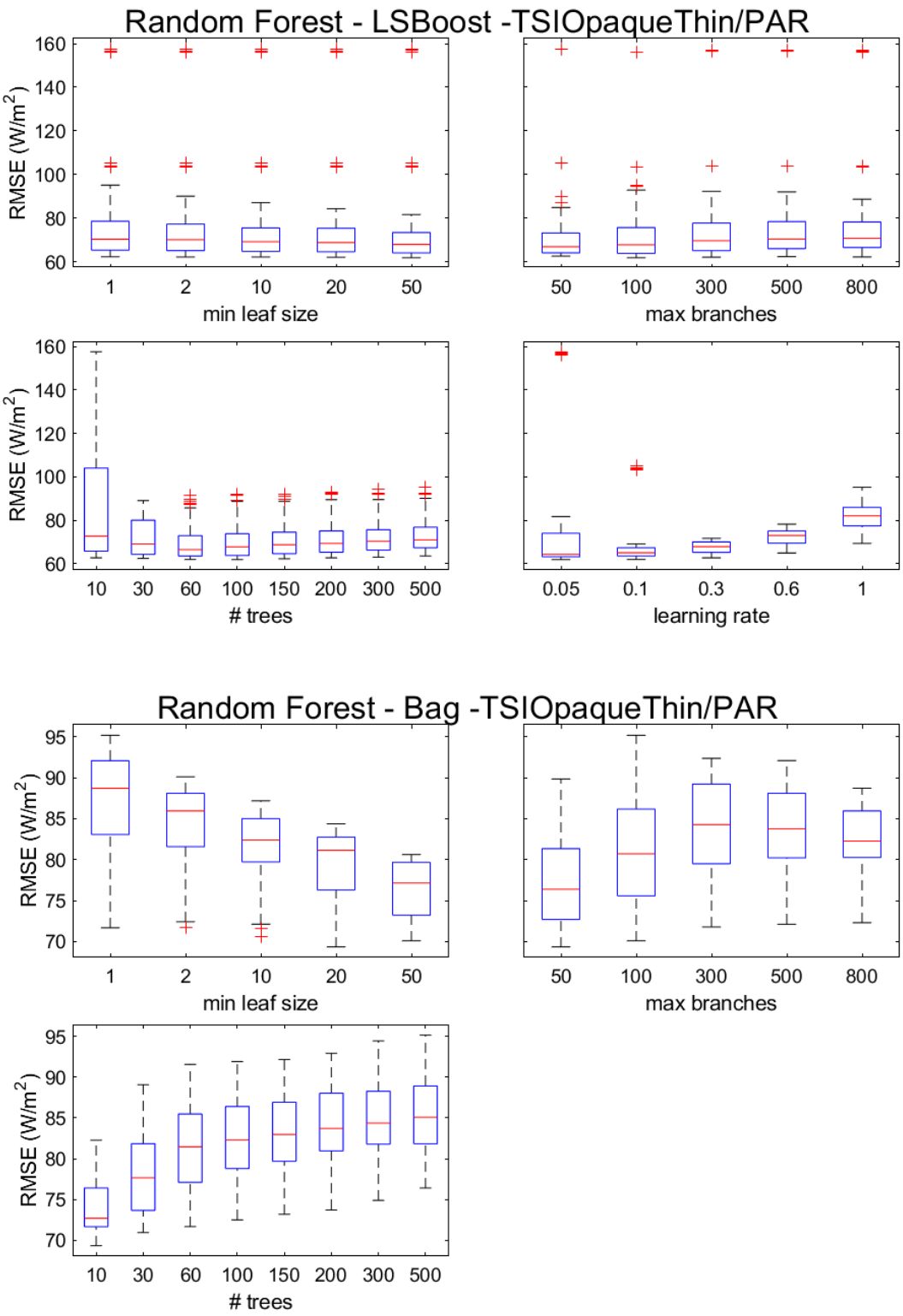

Figure S16. PAR results; TSI cloud data used.

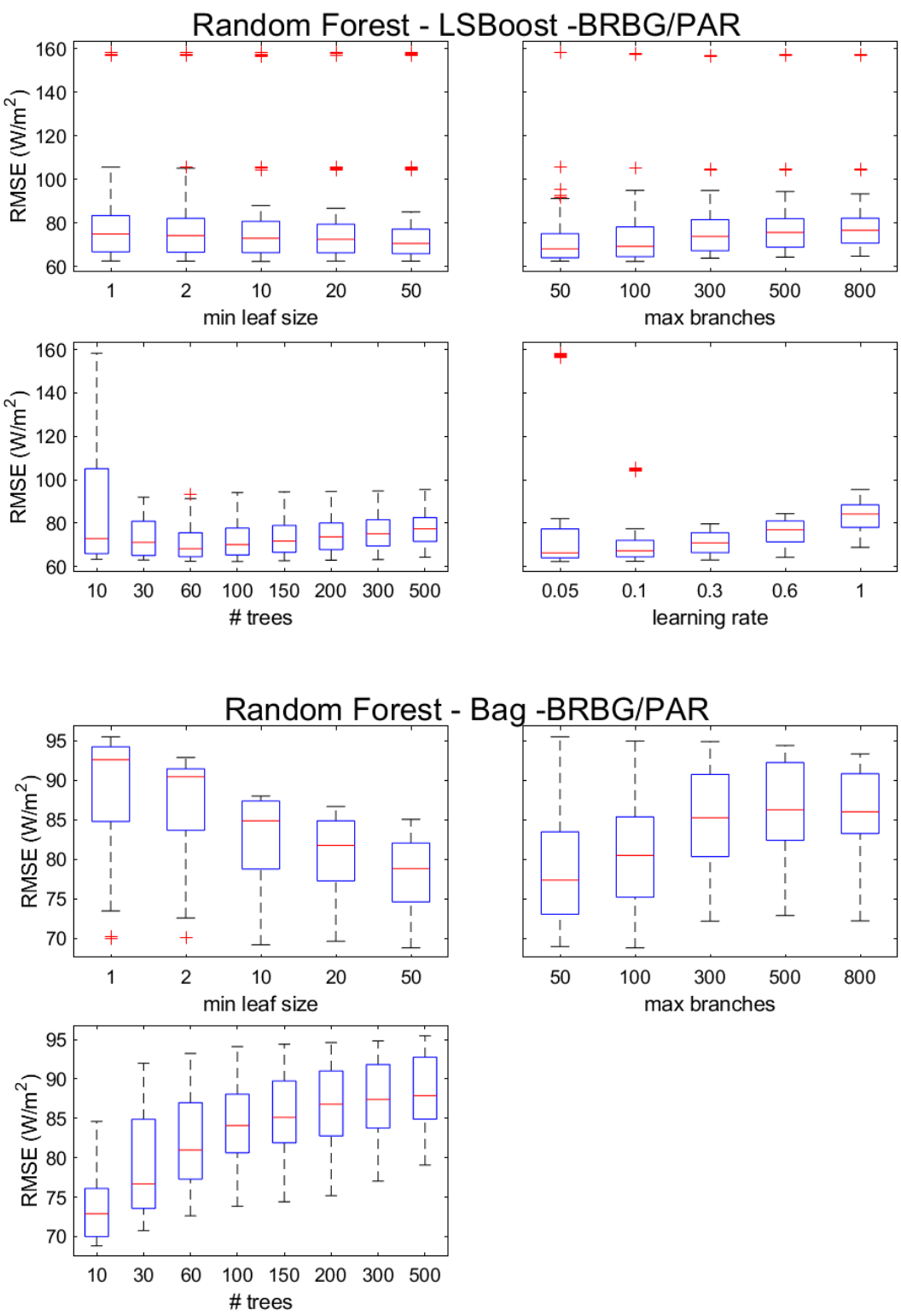

Figure S17. PAR results; BRBG cloud data used.

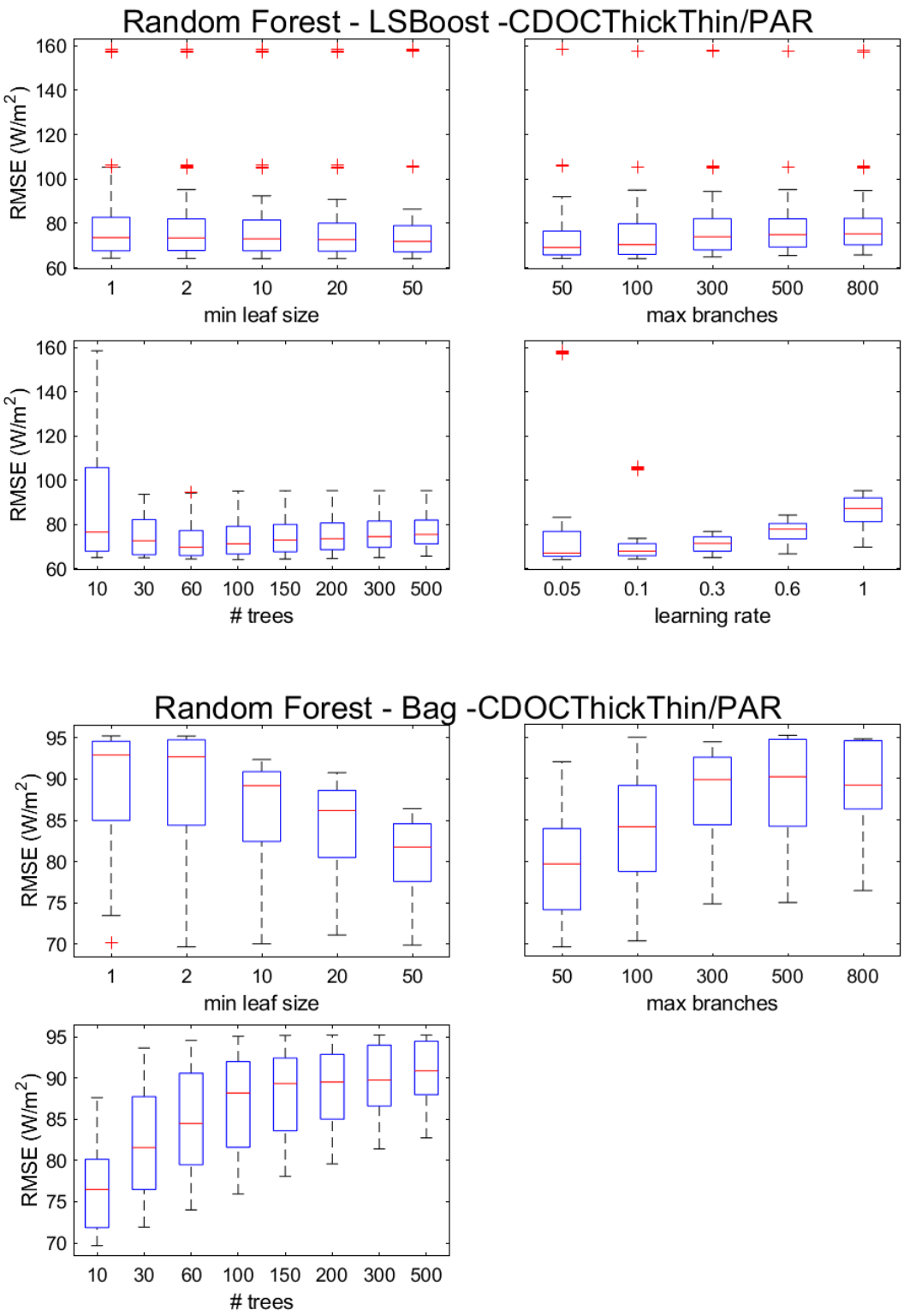

Figure S18. PAR results; CDOC cloud data used.

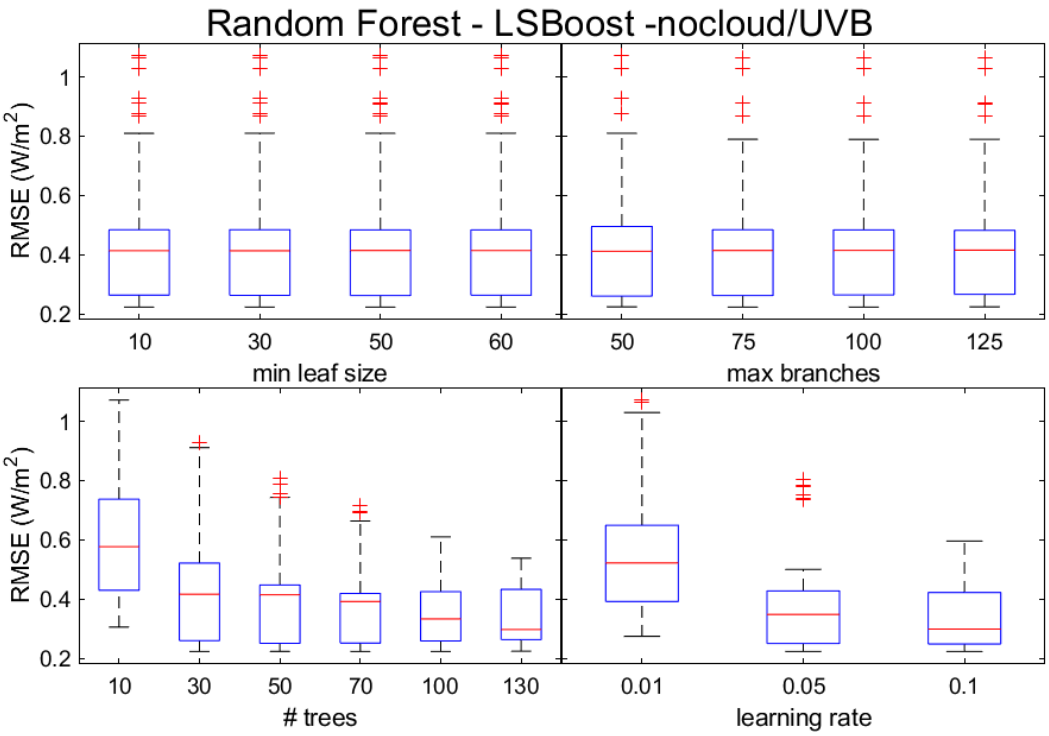

Figure S19. UVB results; no cloud data used.

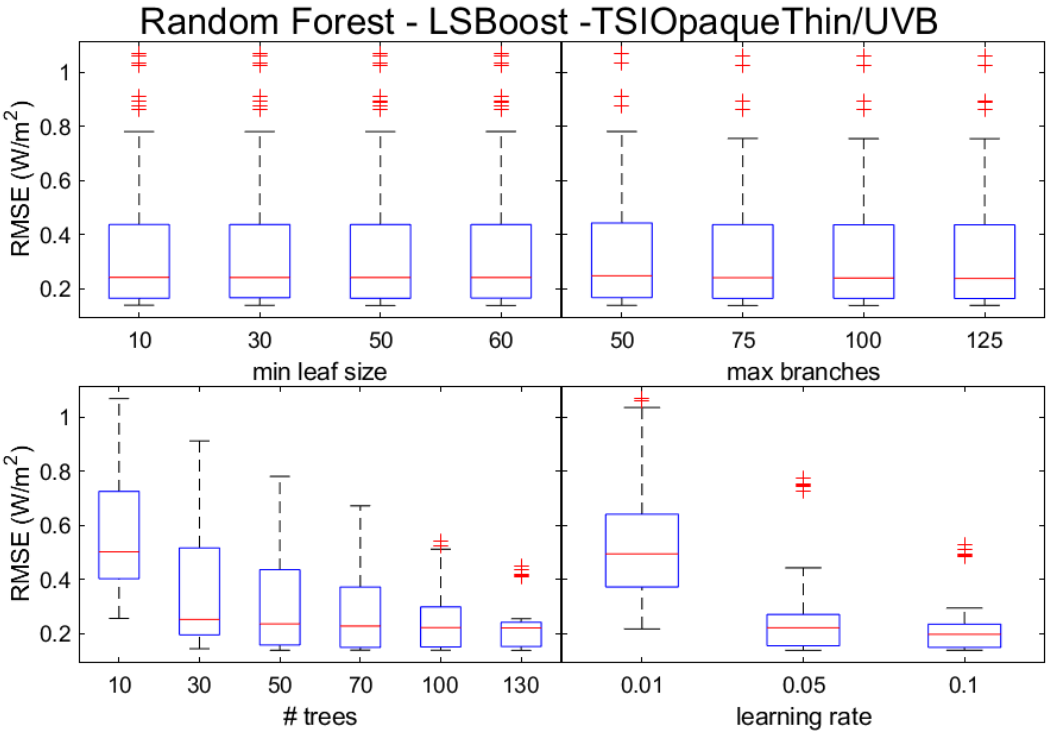

Figure S20. UVB results; TSI cloud data used.

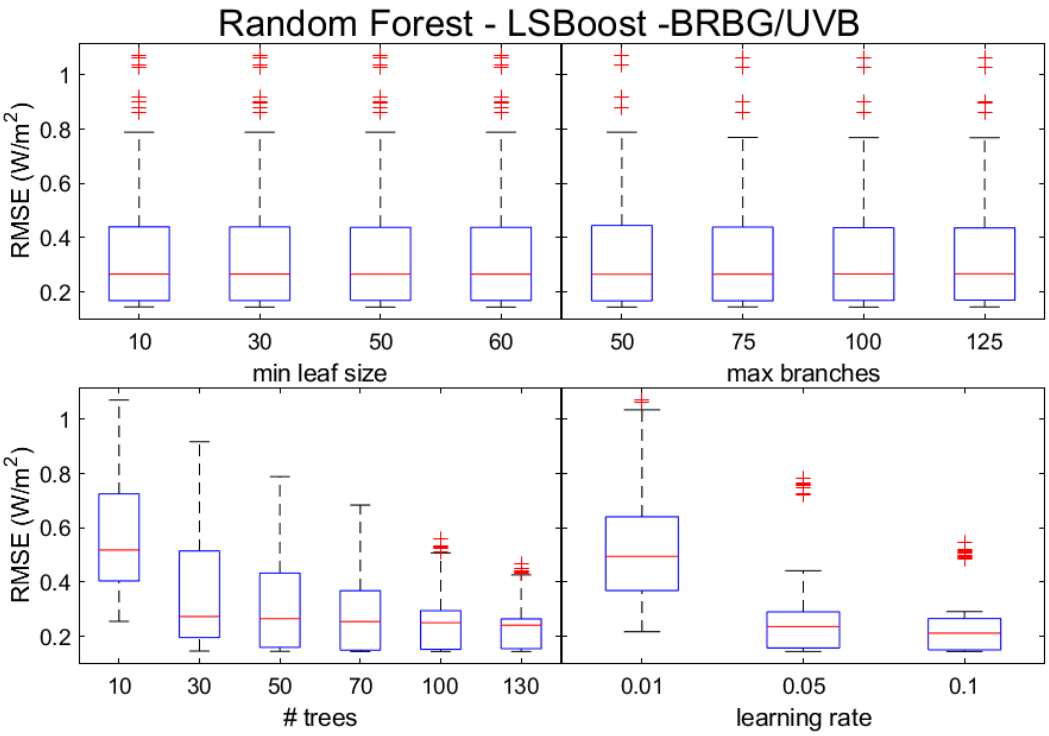

Figure S21. UVB results; BRBG cloud data used.

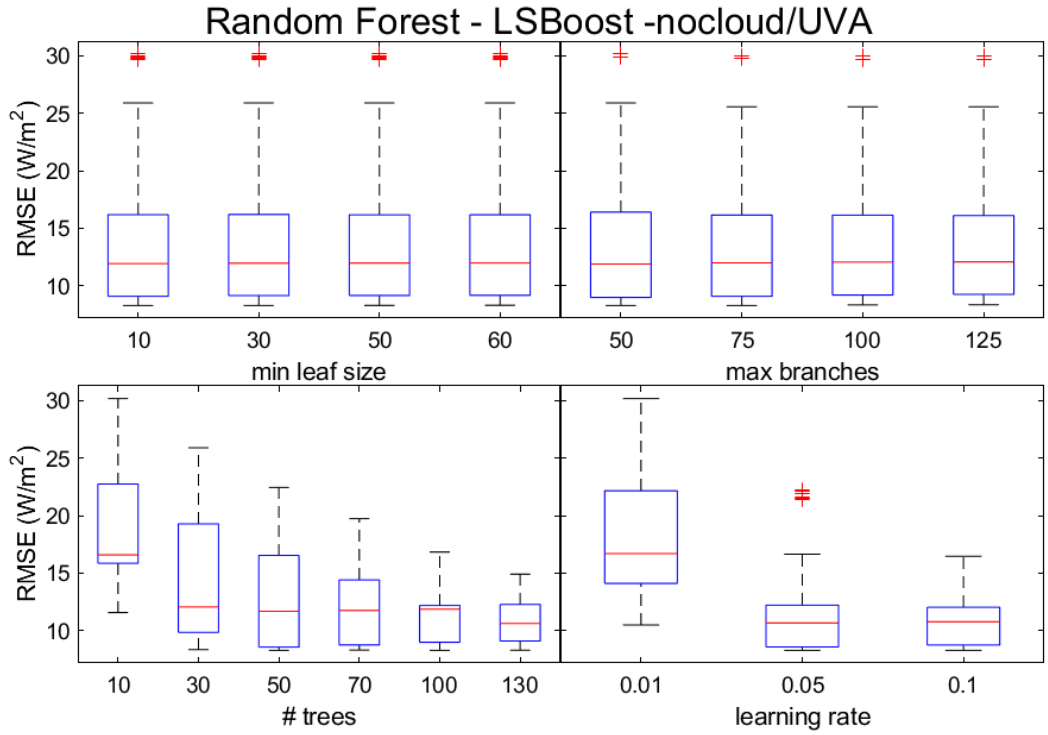

Figure S22. UVA results; no cloud data used.

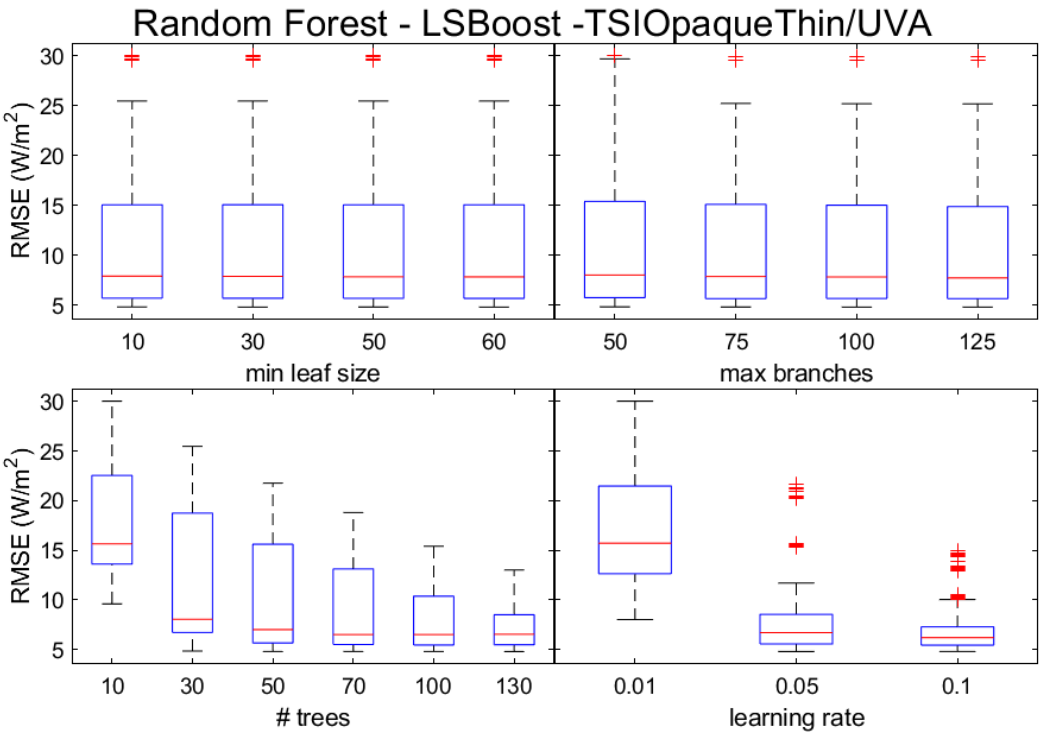

Figure S23. UVA results; TSI cloud data used.

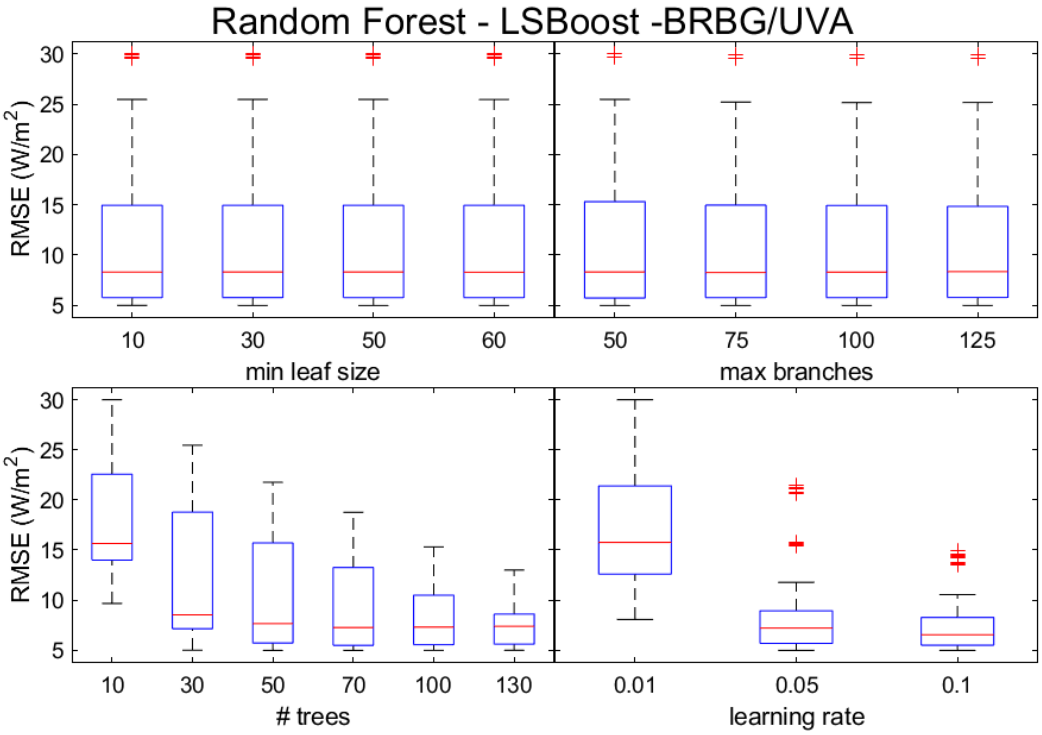

Figure S24. UVA results; BRBG cloud data used.

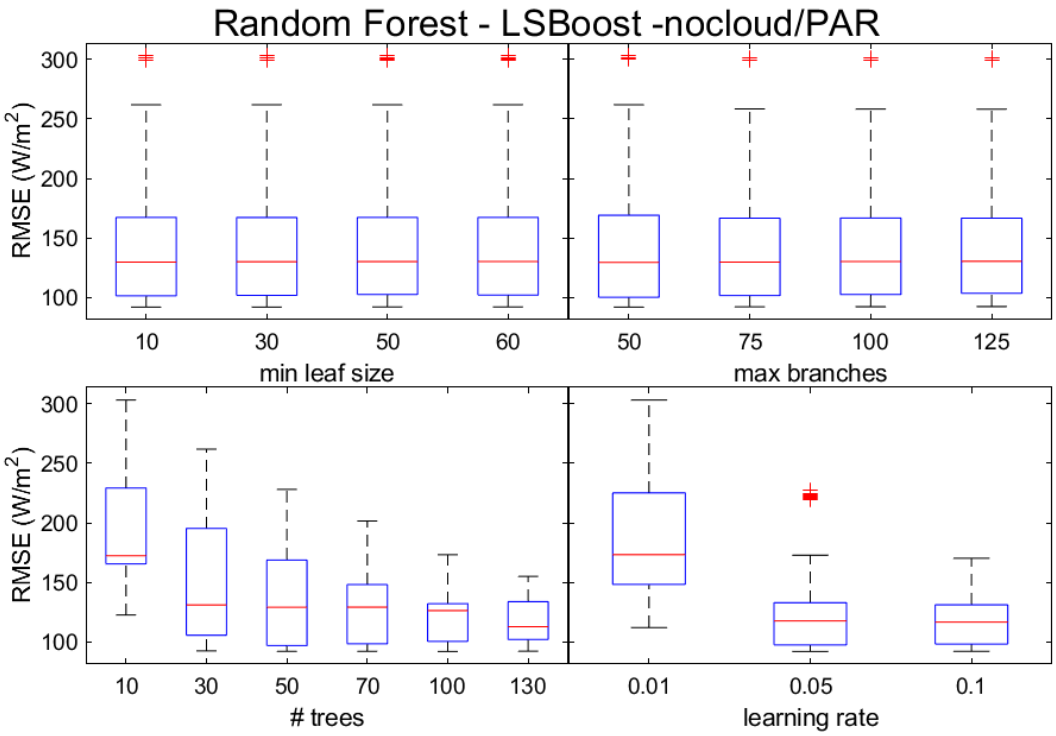

Figure S25. PAR results; no cloud data used

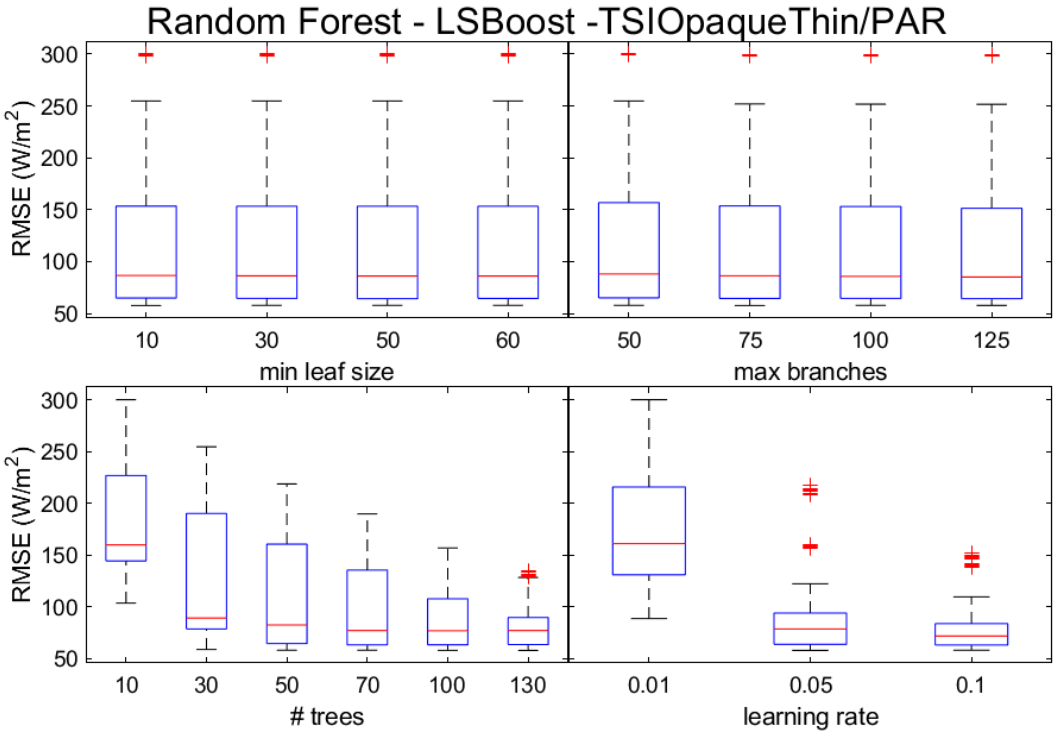

Figure S26. PAR results; TSI cloud data used.

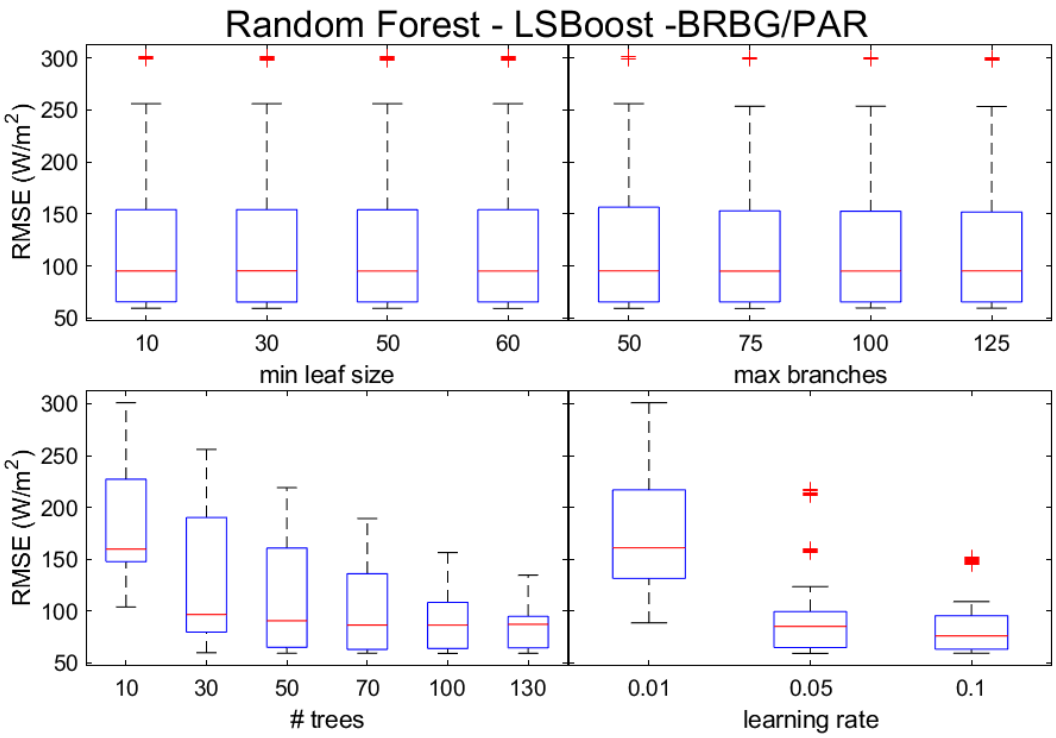

Figure S27. PAR results; BRBG cloud data used.

Section S6. Shallow Artificial Neural Network Model Development, Algorithm Selection, and Optimization

Fundamentals of Shallow Neural Network. A shallow neural network, as opposed to a deep neural network, was chosen because (1) the number of data points (n≈50,000 for training and validation) and number of features (n=3-5) was relatively low, and (2) the relationship between day of year, time of day, cloud cover, and irradiance was expected to be nonlinear but not too complex. The data set was split into chronologically continuous blocks for training (first $6/7^{th}$ of data) and validation (last $1/7^{th}$ of data) to avoid overfitting. The neural network was always trained with normalized data to assist in model convergence, and the model was trained for as many epochs as required to achieve a consistently low mean squared error. The model was trained on irradiance from 290-700 nm. The results from the "best performing" epoch was selected automatically as the final training result.

Initial Optimization. To begin the optimization of the model hyperparameters, five features were used as input where four of the features were the sine and cosine components for day of year and time of day (see Section S3 for more details) and the last feature was the cloud cover. Up to 3 layers were included in the neural network. Each layer had an equivalent number of nodes (aka neurons) that varied between 3 and 30 nodes/layer. The same transfer function was used between each layer except the last layer which used the typical linear transfer function. Five learning functions were tested to determine any potential differences.

With the initial grid search, both 1 and 2 layers performed well but the inclusion of a 3rd layer seemed to lead to overfitting (Figure S28). Between 15 and 18 nodes/layer performed the best when considering all the results from UVB, UVA, and PAR irradiance models. Lastly, the Levenberg-Maraquardt learning function resulted in by far the most consistent model performance with low RMSE (Figure S28).

Table S7. SNN model hyperparameters for corresponding initial grid search

Hyperparameter	Grid search values
number of layers	1, 2, 3
number of nodes	3, 6, 9, 12, 15, 18, 21, 24, 27, 30
learning function	1. Levenberg-Marquardt
	2. BFGS Quasi-Newton
	3. Fletch-Powell Conjugate Gradient
	4. One Step Secant
	5. Scaled Conjugate Gradient
transfer function	3. positive linear (ReLu)

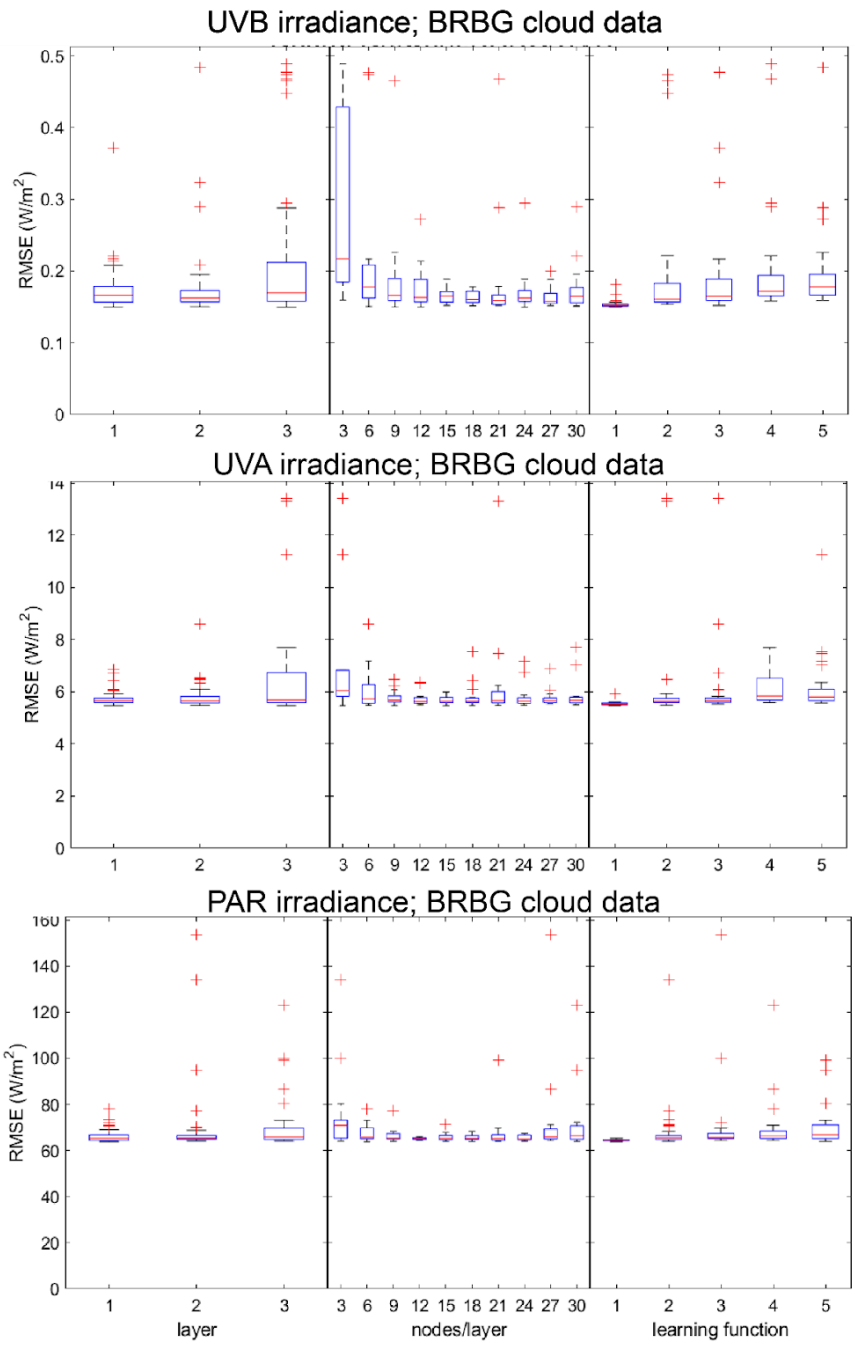

Figure S28. Root mean squared errors for the validation set (1/7th data set from the ~50,000 observations for training/validation) from the initial grid search to optimize hyperparameters. Input data for day of year and time of day was transformed into its sine and cosine components. All data was normalized.

Further Optimization with Transformed and Non-Transformed Inputs. In further optimization efforts, the use of transformed and non-transformed input features was explored as well as other transfer functions (Figures S29-S30). With the transformed data as inputs, model performance was similar over all grid search parameters. Arguably, the model performance using 1 layer outperformed using 2 layers (Figure S30). However, when non-transformed data were used as inputs, having 2 layers and a higher number of neurons helped the model learn the nonlinear relationship present (Figure S30). Because minimizing data manipulation by the user will likely minimize potential errors, the final model was trained with non-transformed data using 2 layers. Lastly, the positive linear transfer function, which was initially chosen, was found to perform worse when using non-transformed data. Therefore, the final model used the hyperbolic tangent sigmoid function, which is the default selection in MATLAB.

Table S8. SNN model hyperparameters for further optimization.

hyperparameter	**grid search values**
number of layers	1, 2
number of nodes	8, 10, 12, 14, 16, 18, 20, 22, 24
learning function	1. Levenberg-Marquardt
transfer function	1. log-sigmoid
	2. hyperbolic tangent sigmoid
	3. positive linear (i.e, ReLU)
	4. saturating linear

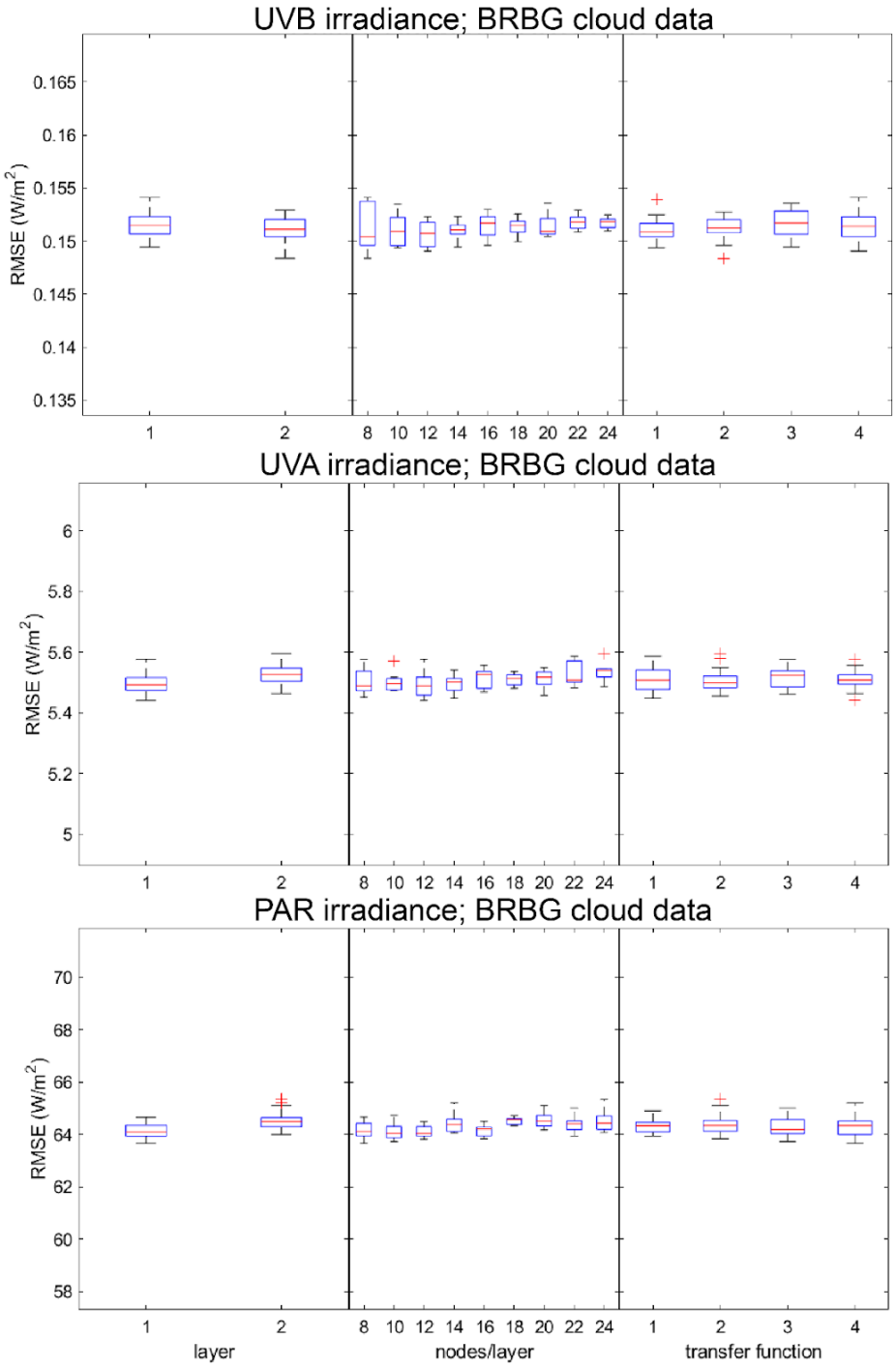

Figure S29. Root mean squared errors for the validation set (1/7th data set from the ~50,000 observations for training/validation) from further optimization of the hyperparameters. Input data for day of year and time of day was transformed into its sine and cosine components. All data was normalized.

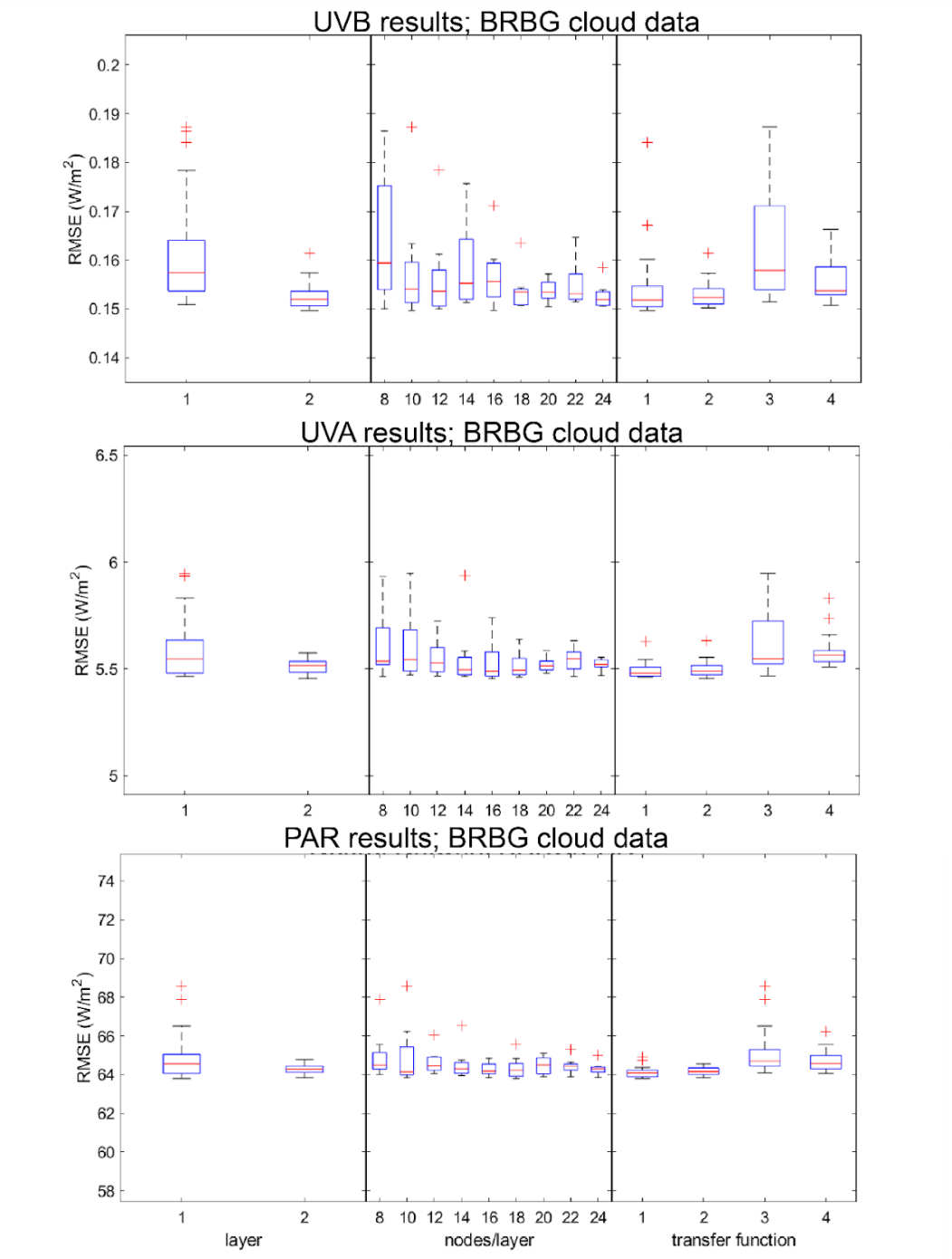

Figure S30. Root mean squared errors for the validation set (1/7th data set from the ~50,000 observations for training/validation) from further optimization of the hyperparameters. Input data for day of year and time of day was not transformed. All data was normalized.

Final Model Parameters. To quantify the expected performance for the validation set, the model was trained 50 times. Differences in performance are due to the randomized initial weights and data divisions used to train the model. The range of RMSE values calculated is narrower than for the random forest model because these values were only calculated on one validation set (the last1/7th of the data set from the ~50,000 observations for training/validation).

Table S9. SNN model hyperparameters for further optimization.

Hyperparameter	**Grid Search Values**
number of layers	2
number of nodes	16
learning function	1. Levenberg-Marquardt
transfer function	2. hyperbolic tangent sigmoid

Table S10. SNN model RMSE achieved during further optimization.

Irradiance Region	**RMSE (W/m^2)** (n=50)
UVB	0.151 (0.149 – 0.153)
UVA	5.50 (5.45 – 5.56)
PAR	64.2 (63.7 – 64.7)

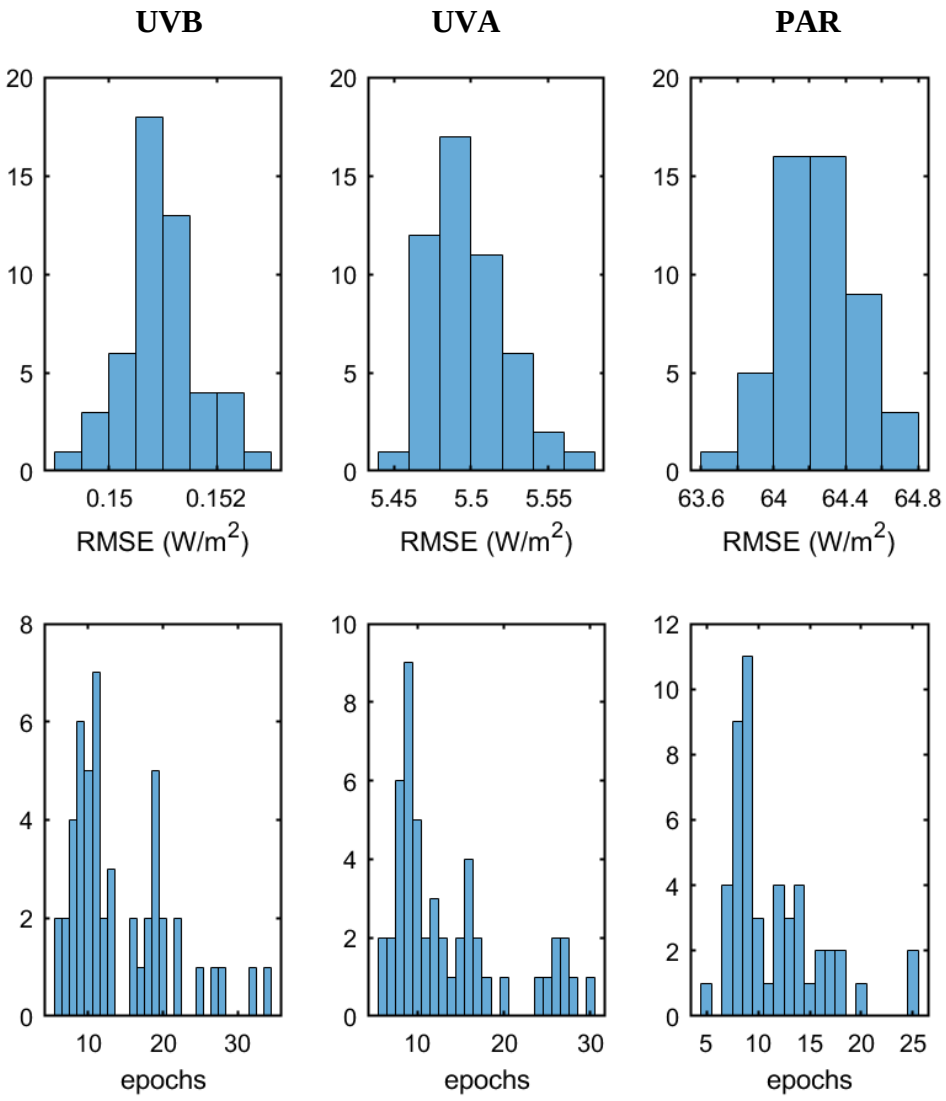

Figure S31. Histograms for 50 iterations of final model training for the shallow neural network (number of layers = 2; 16 nodes/layer; Levenberg-Marquardt learning function; hyperbolic tangent sigmoid transfer function for both layers). The top row shows the RMSE for the validation set ($1/7^{th}$ of the data set from the ~50,000 observations for training/validation) and the bottom row shows the number of epochs taken to get to the best performance.

Section S7. Multiple Linear Regression Model Results

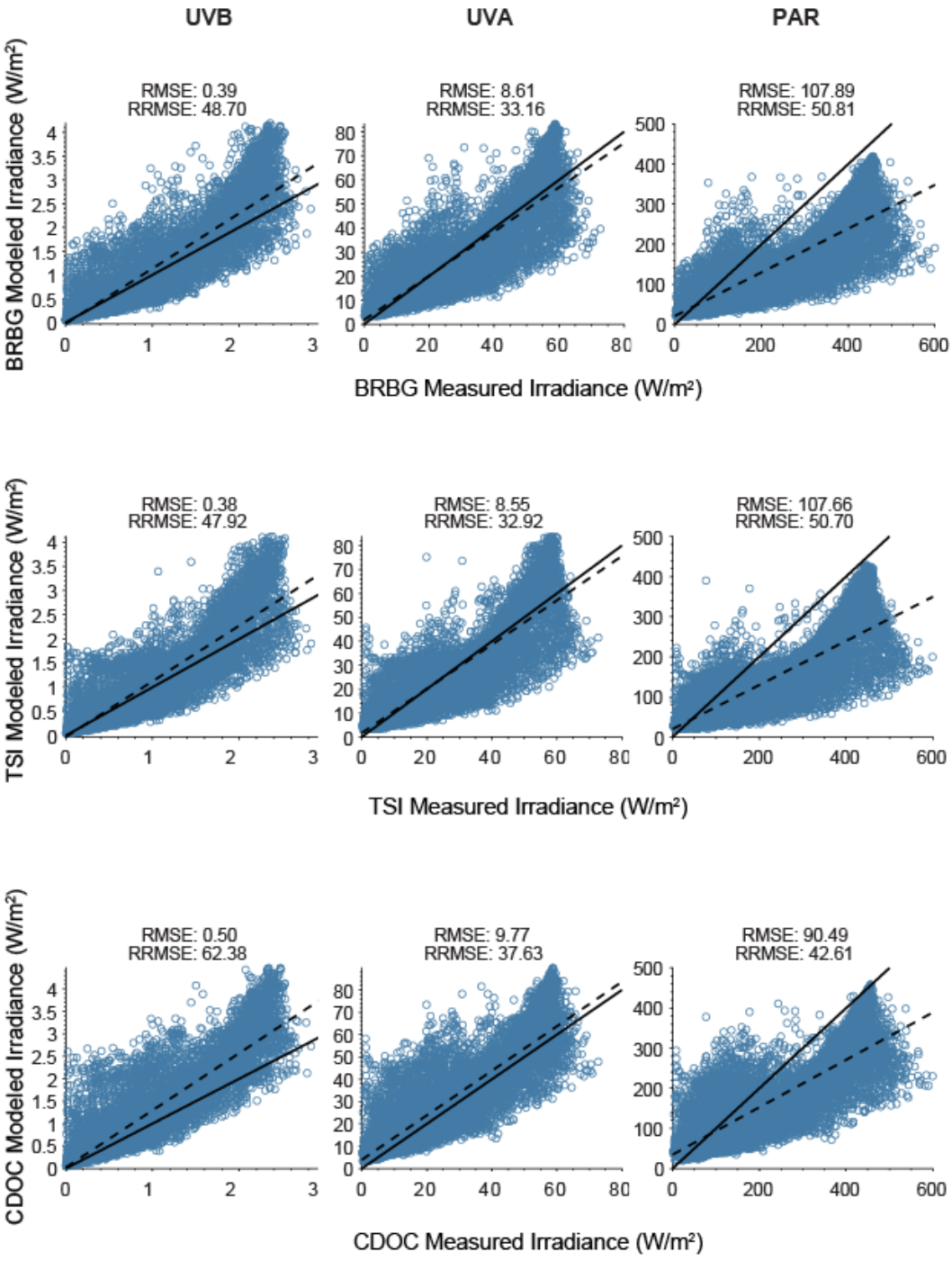

Figure S32. MLR test data set results for measured verse modeled irradiance. The slashed line represents the linear regression trend and the solid line is a one to one line (x=y).

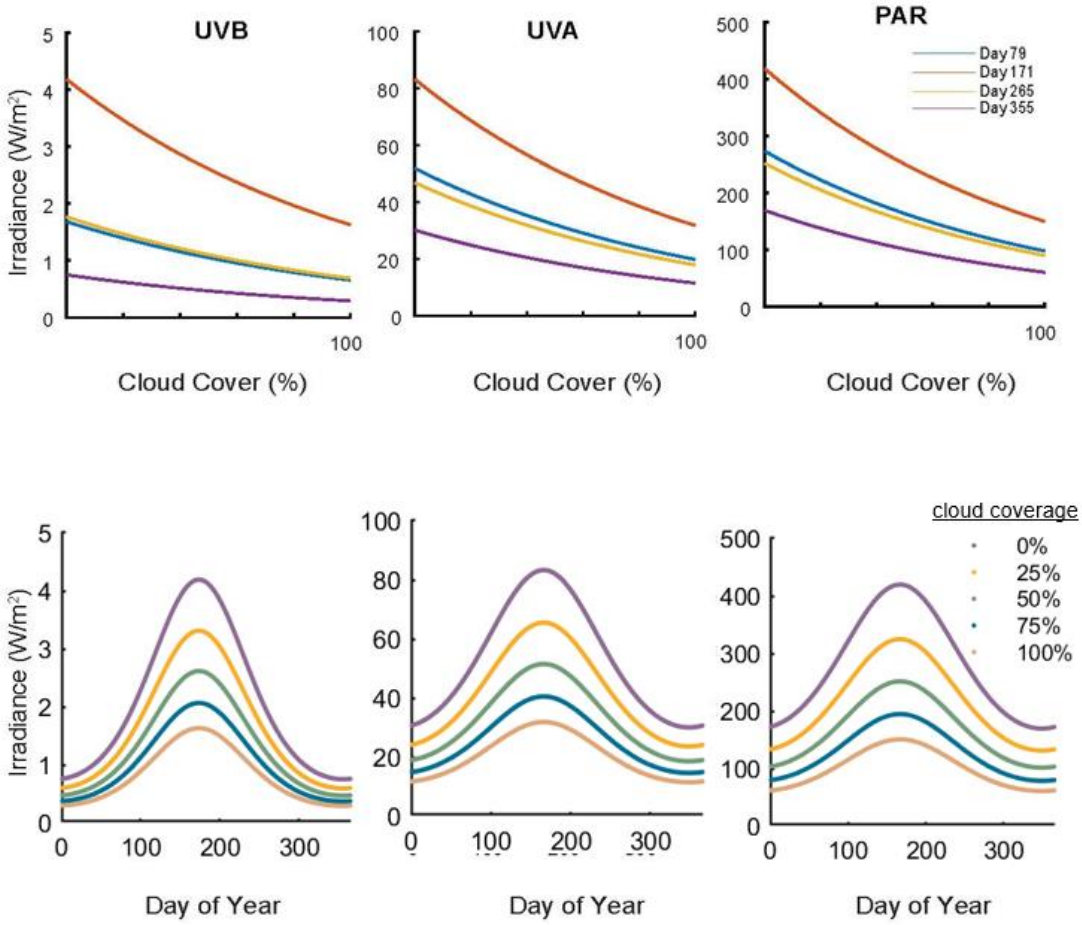

Figure S33. MLR results for varying conditions using BRBG cloud coverage. Day 79, 171, 265, and 355 are representative of the equinoxes/solstices (Spring, Summer, Fall, and Winter, respectively).

Section S8. Random Forest Model Results

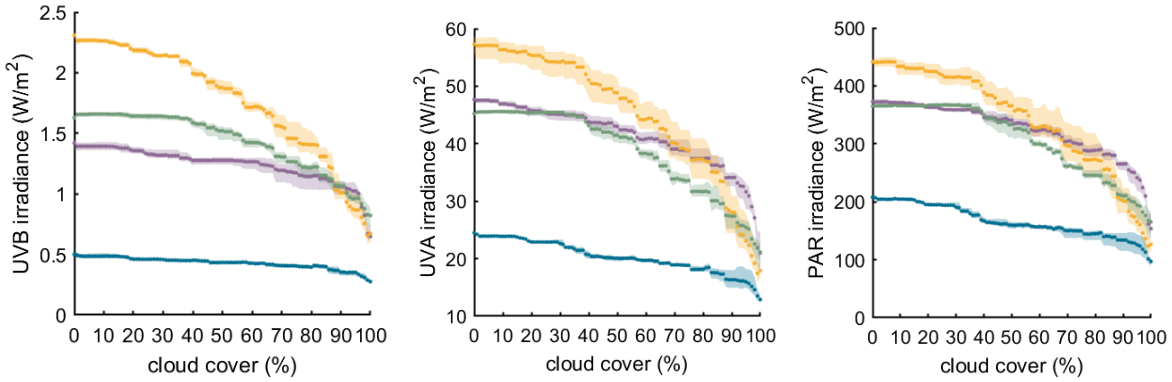

Figure S34. RF result summary for BRBG cloud cover. The lines are for the seasons; yellow=summer, green=fall, blue=winter, and purple=spring. The shaded regions indicate model predictions and the average of these is shown with the solid point.

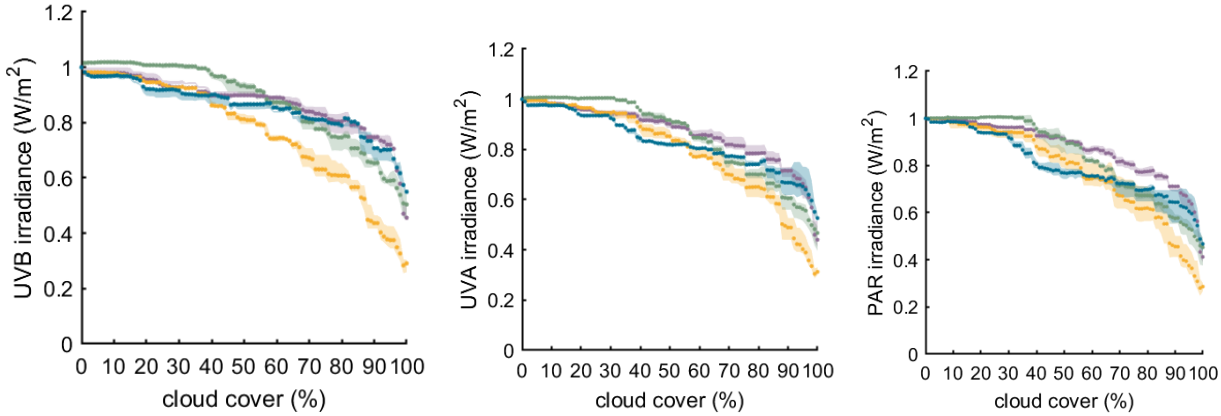

Figure S35. RF result ratios for BRBG cloud cover with clear sky day normalization. The lines are for the seasons; yellow=summer, green=fall, blue=winter, and purple=spring. The shaded regions indicate model predictions and the average of these is shown with the solid point.

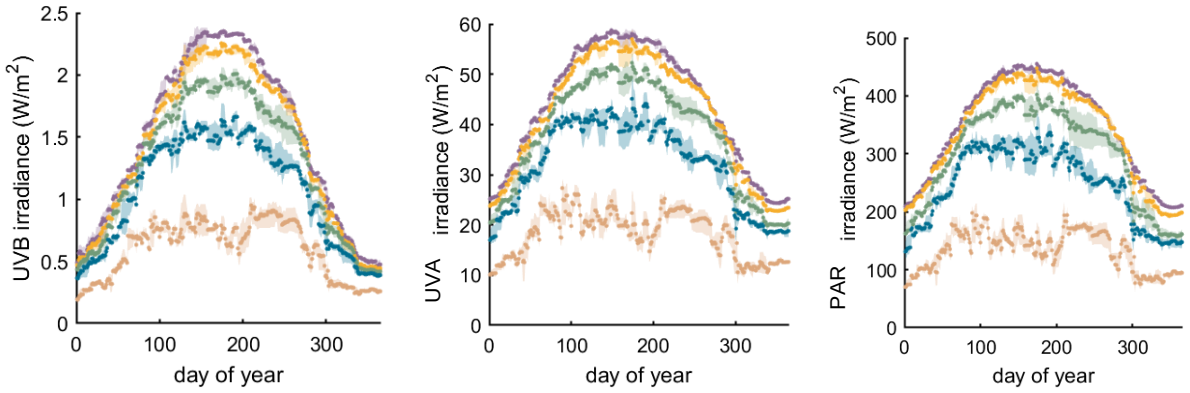

Figure S36. RF result summary for BRBG day of year. The lines are for the cloud cover; purple=0%, yellow=25%, green=50%, blue=75%, and orange=100%. The shaded regions indicate model predictions and the average of these is shown with the solid point.

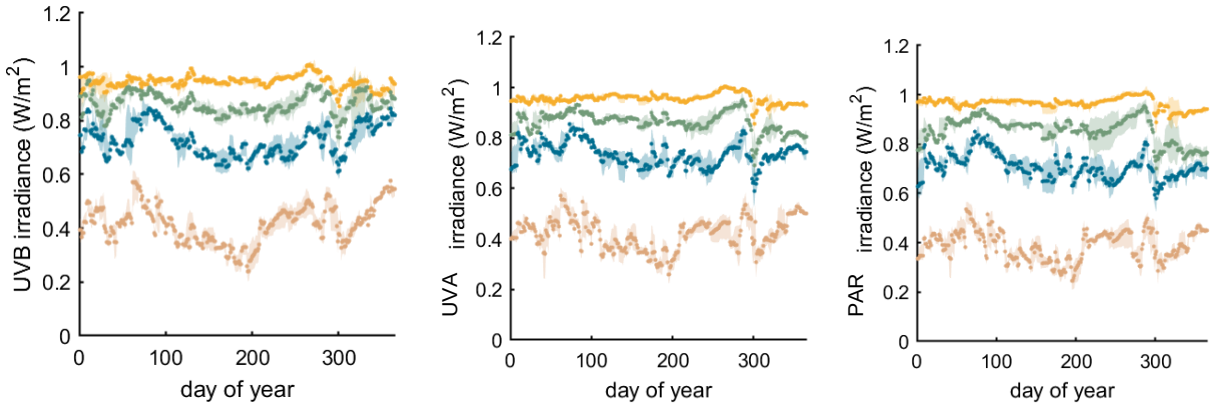

Figure S37. RF result ratios for BRBG day of year. The lines are for the cloud cover; yellow=25%, green=50%, blue=75%, and orange=100%. The shaded regions indicate model predictions and the average of these is shown with the solid point.

Section S9. Neural Network Model Results

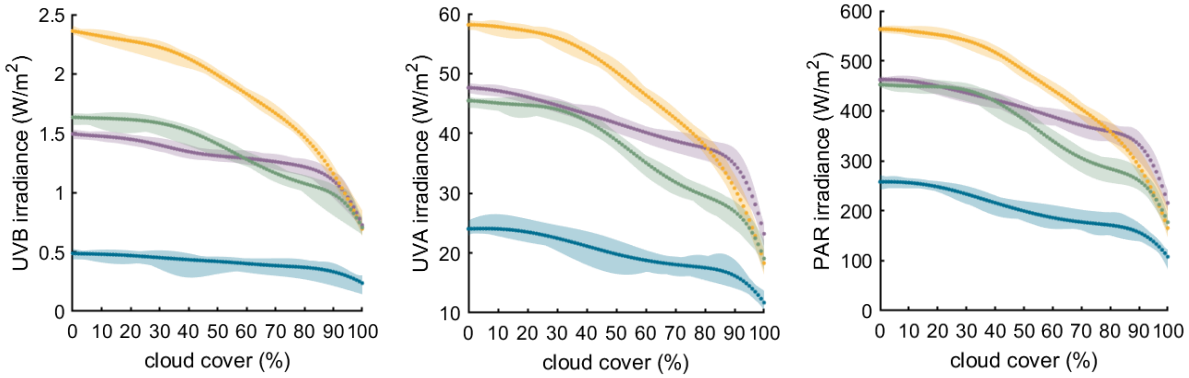

Figure S38. NN result summary for BRBG cloud cover. The lines are for the seasons; yellow=summer, green=fall, blue=winter, and purple=spring. The shaded regions indicate model predictions and the average of these is shown with the solid point.

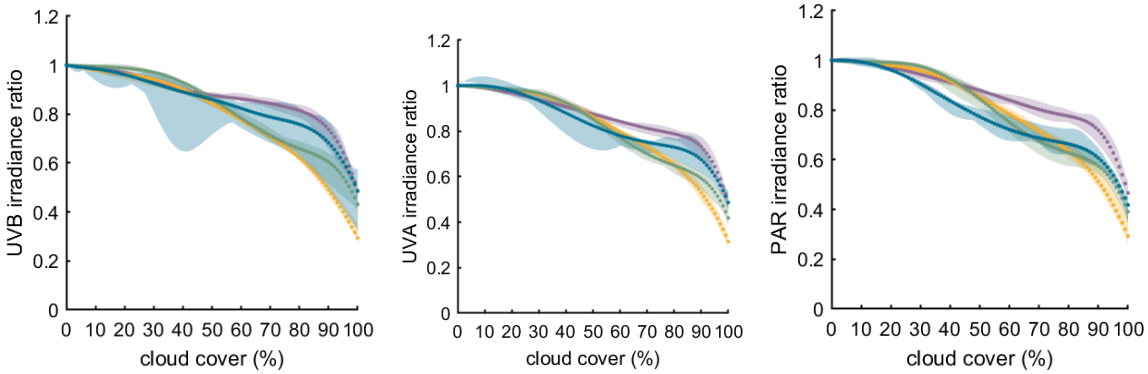

Figure S39. NN result summary for BRBG cloud cover. The lines are for the seasons; yellow=summer, green=fall, blue=winter, and purple=spring. The shaded regions indicate model predictions and the average of these is shown with the solid point.

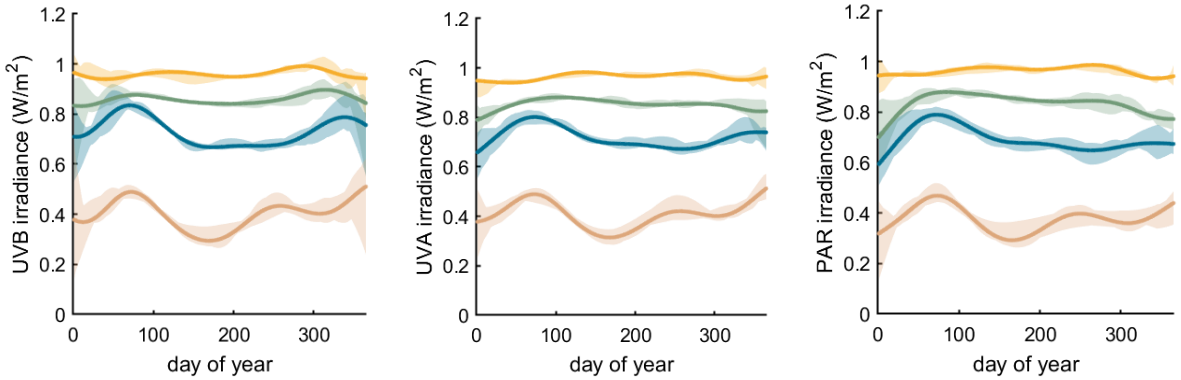

Figure S40. RF result ratios for BRBG day of year. The lines are for the cloud cover; yellow=25%, green=50%, blue=75%, and orange=100%. The shaded regions indicate model predictions and the average of these is shown with the solid point.

Section S10. Modeled Results on Sun-Cloud Position for Irradiance Regions

The neural network model was additionally run with the position of the sun in respect to the clouds on top of cloud cover, time of day, and day of year. This variable, referred to as sun flag or solar disc, was reported by all sky imager (ASI) software. The software summarized whether the sun was not visible, visible on a clear sky day, partly covered, behind the clouds but a bright dot was visible, or outside of view due to the solar zenith angle or horizon.[7]

Irradiance was modeled over the year at different levels of cloud cover (25, 50, and 75%) for varied sun flag summaries (Figure S41). As expected, where the sun was least obstructed by clouds was where irradiance was most intense. Addition of the sun flag resulted in moderate improvement of model performance. From the original BRBG data, there was a 21-24% improvement in model results (UVB, UVA, and PAR). For example, the UVB rRMSE was originally 28.3% and with the sun flag it decreased to 22.3%. The sun flag for the TSI data improved the rRMSE by 19-22% with the original UVB rRMSE of 27.8% and with the sun flag, 19.4%.

Although the sun flag improved the model, the sun flag was ultimately not used for final irradiance model prediction as it was not a strong irradiance indicator. This is likely attributed to other atmospheric components such as gaseous species and particulate matter unaccounted for[8] and previous studies have shown that factors like cloud altitude are important. [9]Additionally, cloud cover percent innately extracts the likelihood that the sun is covered by clouds alone such as if it is mostly cloudy, the more likely the sun is to be cloud-covered. Lastly, there were inaccurate sun flag definitions; glare on a clear sky day was often interpreted as cloud cover (Figure S42).

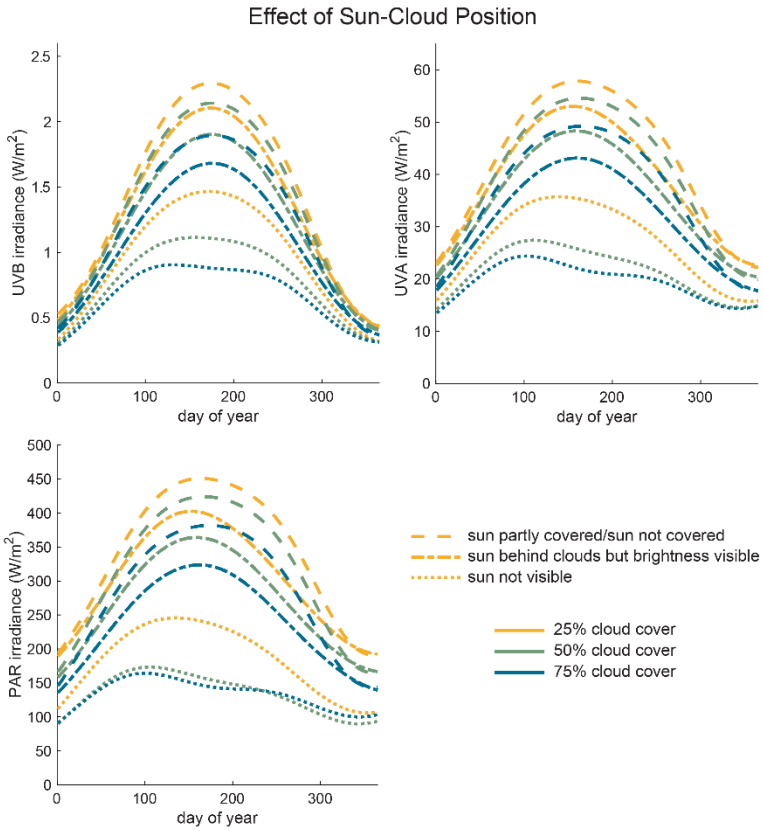

Figure S41. Neural network modeled results for irradiance regions (UVB, UVA, and PAR) as a function of cloud cover and sun position in the sky.

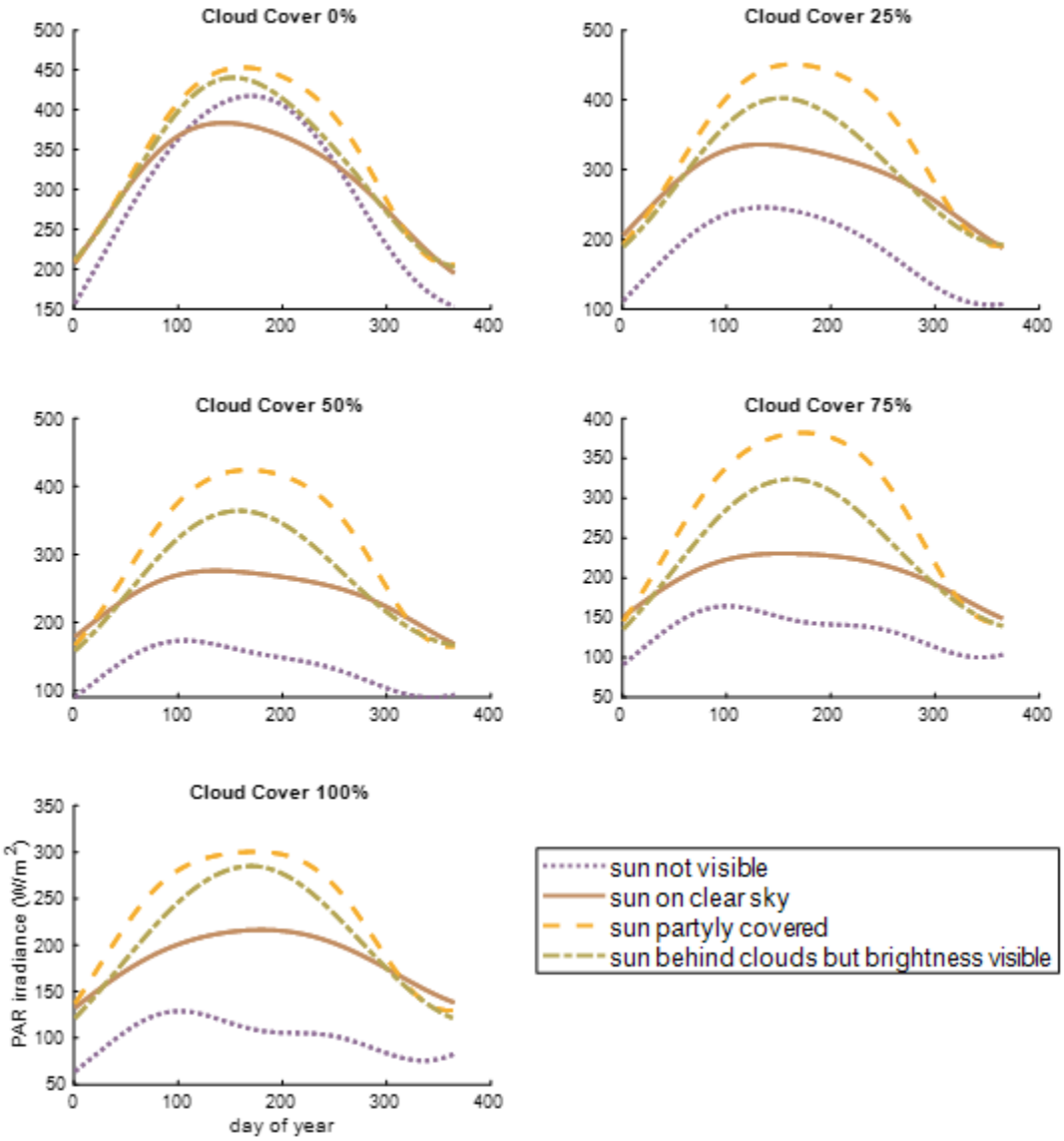

Figure S42. Neural network irradiance results over a year at increments of cloud cover (0, 25, 50, 75, 100%) as a function of the sun flag.

Section S11. Wavelength Dependence with Neural Network Model

S11-1. Initial Optimization of Shallow Neural Network for Wavelength-Dependent Irradiance.

The neural network was initially optimized by comparing the results for 2 and 3 hidden layers and between 8 and 26 nodes for wavelengths from 295 nm to 400 nm (in 5 nm increments). Not enough data was present for a viable signal at 290 nm to base the model on. Day of year, time of day, and cloud coverage were used as inputs. The cloud data was from the BRBG algorithm, which consists of only one value for total cloud cover. All input data was normalized beforehand and un-transformed before calculation of the RMSE and relative RMSE.

The Levenberg-Marquardt learning function and the hyperbolic tangent sigmoid transfer function were chosen from the beginning because of their better performance for UVB, UVA, and PAR data. The data splitting and validation was the same as previously described. Values for all wavelengths were predicted simultaneously. For the initial optimization, this means there were 22 outputs in the output layer that were predicted.

The results showed a wavelength-dependence for how well the neural network performed (Figure S43). Wavelengths of 295 and 300 nm had the highest relative RMSE, but there was a quick decline with a minimum at 325 nm. Surprisingly, the rRMSE began to rise again for higher wavelengths.

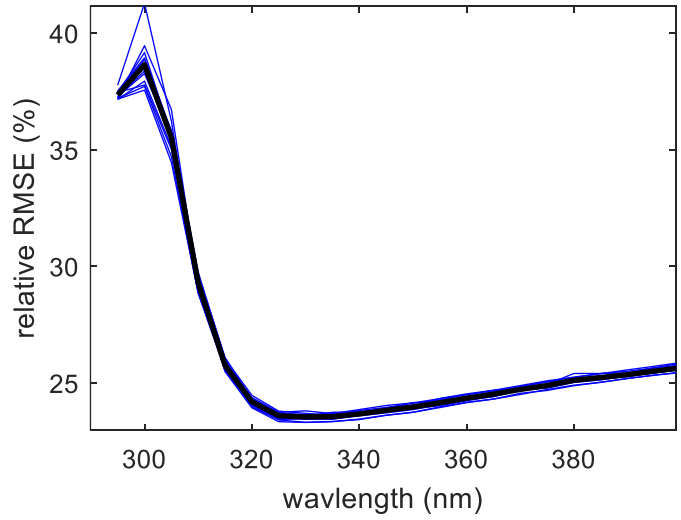

Figure S43. Calculated rRMSE of the validation set as a function of wavelength. The blue lines represent the 14 neural network architectures evaluated and the black line is the average of the 14 lines.

The overall performance of the neural network architecture was evaluated by taking the average of the rRMSE values for the 22 wavelengths evaluated. The results showed almost no variation when increasing the number of nodes from 8 to 26 or the number of hidden layers from 2 to 3.

Therefore, it appears a neural network with only 2 hidden layers and a smaller number of nodes is sufficient to model this relationship.

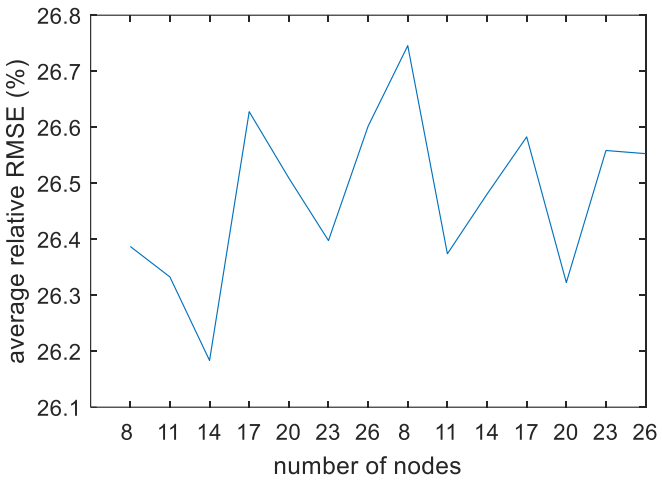

Figure S44. Calculated relative RMSE average over all wavelengths evaluated. The first 7 x-values are for 2 hidden layers and the rest are for 3 hidden layers.

S11-2. Neural Network for Wavelength-Dependent Irradiance

The final model was trained with 2 hidden layers with 16 nodes each to use the same architecture as the neural network models for UVB, UVA, and PAR data. The final output was for wavelengths from 290 nm to 550 nm at increments of 2 nm. The model was trained without using cloud data (2 inputs: day of year and time of day) and using BRBG total cloud cover (3 inputs).

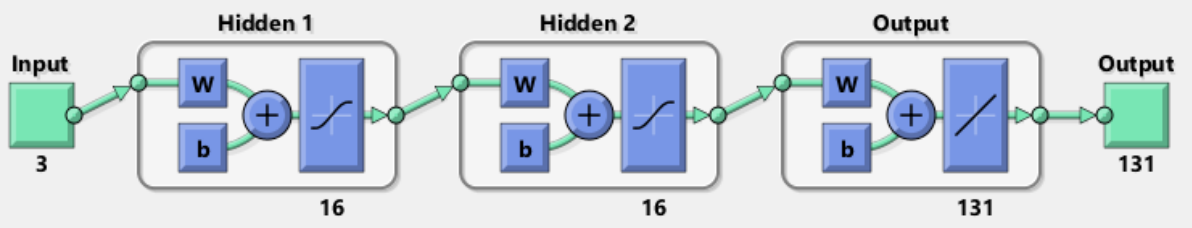

Figure S45. Neural network model set-up run on MATLAB; 2 hidden layers with 16 nodes (neurons).

The model error for the final model had similar trends to those observed in model optimization. The lowest wavelengths had a substantially higher relative RMSE and the error had a minimum around 325 nm. Between 322 nm and 550 nm, the rRMSE slowly increased from 24% to 29%. By using the cloud data in the model, the rRMSE decreased by approximately 14 percentage points.

cloud cover data = none

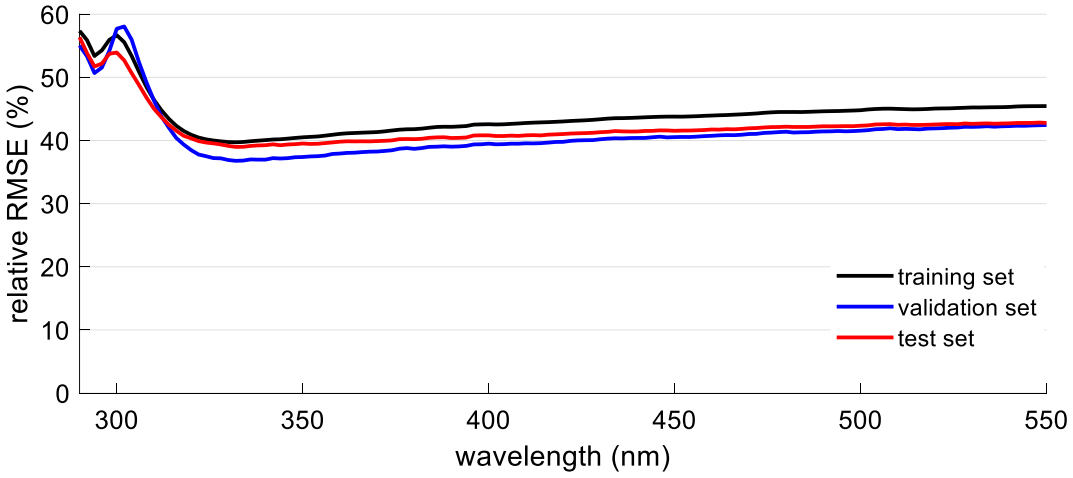

Figure S46. Relative root mean squared error for solar noon wavelengths between 290 and 550 nm (every 2 nm) for the training set and validation set when using only time of day and day of year (no cloud data) to train a neural network with 2 inputs, 2 hidden layers with 16 nodes each, and 131 outputs.

cloud cover data = BRBG total cloud cover

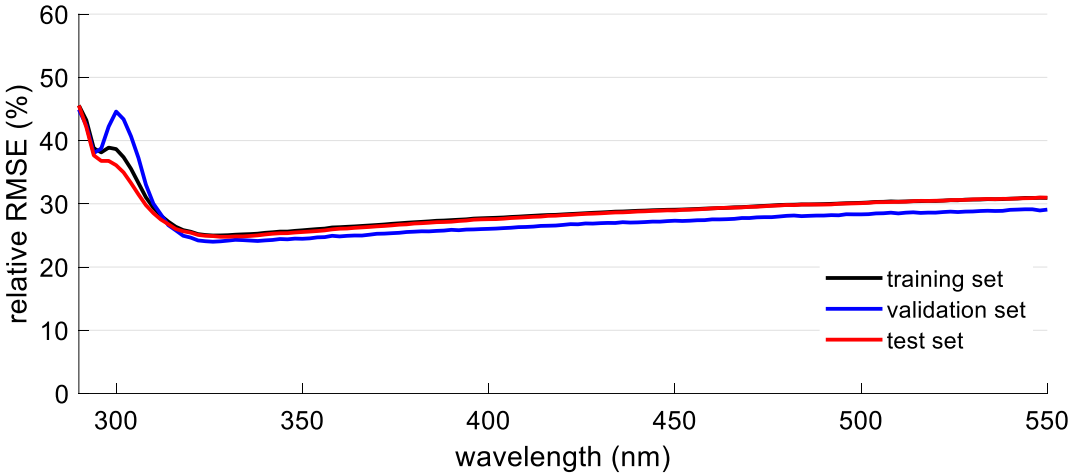

Figure S47. Relative root mean squared error for solar noon wavelengths between 290 and 550 nm (every 2 nm) when using BRBG total cloud cover to train a neural network with 3 inputs, 2 hidden layers with 16 nodes each, and 131 outputs.

S11-3. Results for Wavelength-Dependent Solar Noon Irradiance Neural Network

The modeled results for solar noon are shown in Figure S48-S49 in terms of absolute irradiance and as a fraction of the value modeled for clear skies (0% cloud cover). In general, the effect of cloud cover on irradiance appears to be independent of wavelength. There are variations in this trend below 320 nm. However, because these variations (1) go in both directions, (2) are more severe in winter when absolute irradiances are lower, (3) are in the region where the calculated rRMSE are higher, and (4) the magnitude of these variations are less than the calculated errors, it is likely that this observation is an artefact from imprecise measurements or instrument noise.

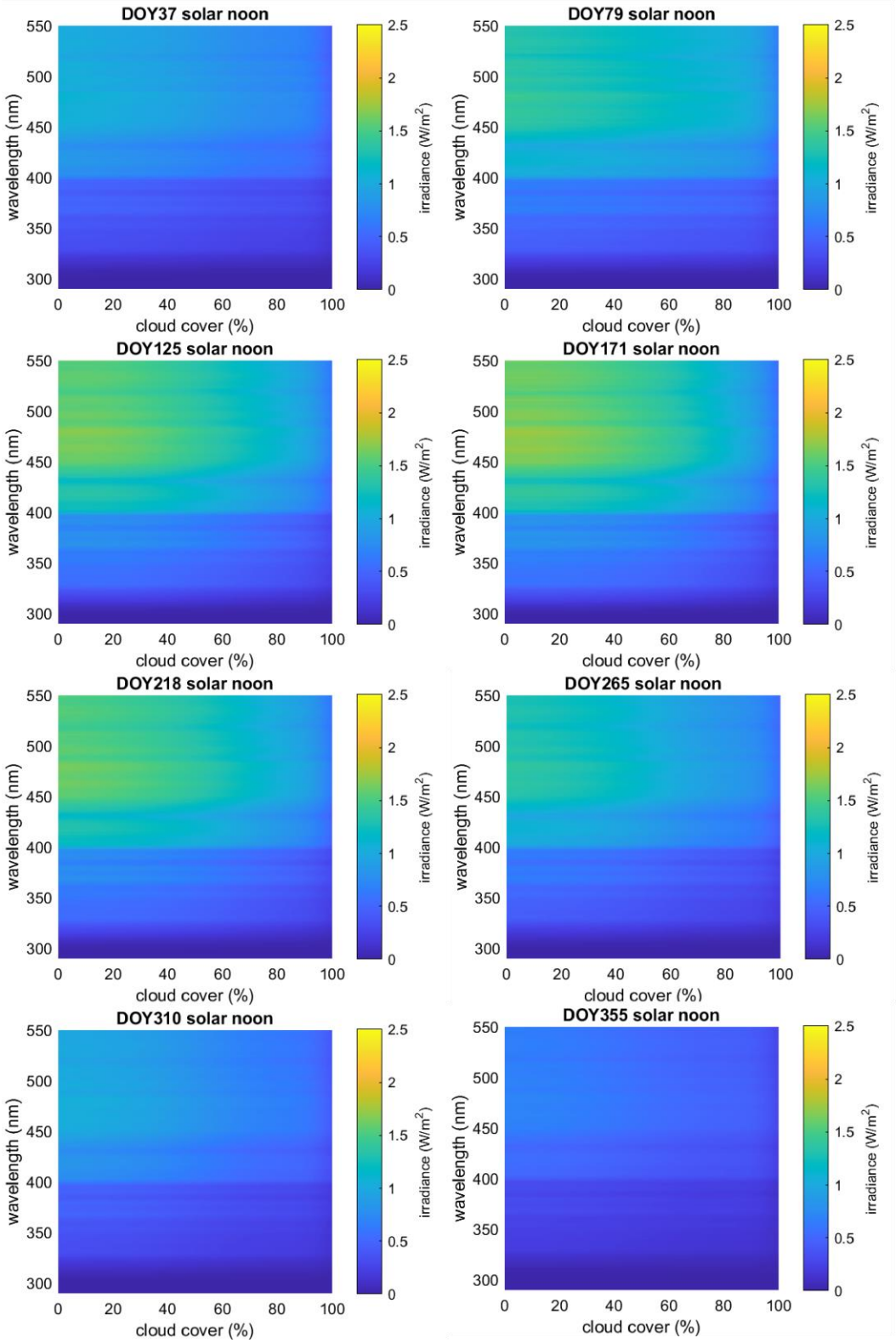

Figure S48. Modeled irradiance values as a function of wavelength and cloud cover for solar noon and eight days of the year (solstices and equinoxes in the right column, the midpoint between those days in the left column).

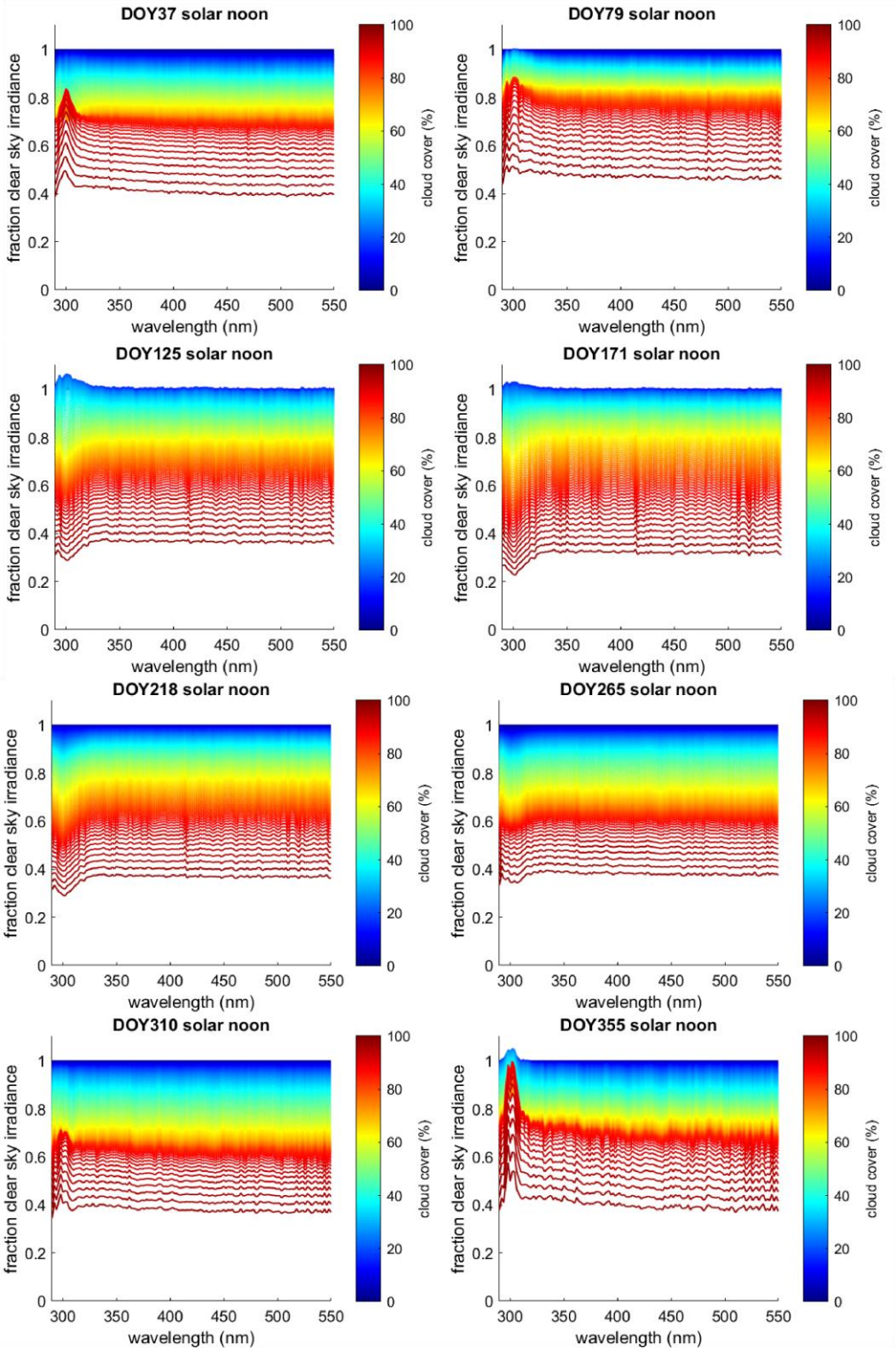

Figure S49. Modeled irradiance as a fraction of clear sky values as a function of wavelength and cloud cover for solar noon and eight days of the year (solstices and equinoxes in the right column, the midpoint between those days in the left column).

S11-4. rRMSE Neural Network for Wavelength-Dependent Daily Irradiance

Consistent with the results for UVB, UVA, and PAR, using values of daily cloud cover and daily irradiance resulted in smaller values of rRMSE. The model errors were highest around 300 nm. These errors likely stem from the higher instrument noise for lower wavelengths, which means wavelengths above 320 nm can be modeled with a higher degree of confidence. The lower values of rRMSE for daily irradiance is hypothesized to be caused by the averaging of data which could "average out" some of the noise of the measurements.

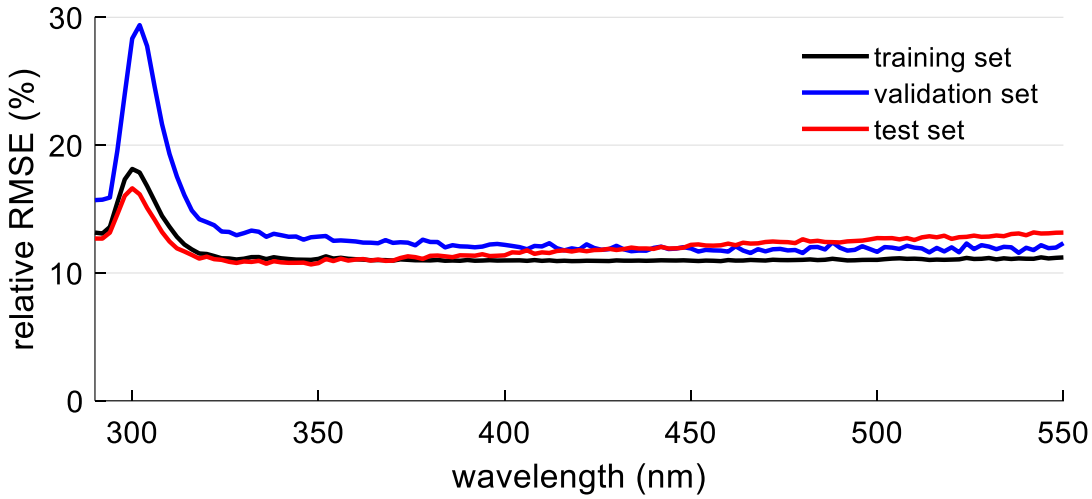

Figure S50. Relative root mean squared error for daily irradiance prediction for wavelengths between 290 and 550 nm (every 2 nm) when using BRBG cloud cover data to train a neural network with 2 inputs, 2 hidden layers with 16 nodes each, and 131 outputs. The inputs were day of year and the daily average cloud cover and the outputs were daily irradiance in W/m^2.

S11-5. Model results for Wavelength-Dependent Daily Irradiance Network

Similar trends were observed for daily irradiances as for solar noon irradiances. There was more wavelength-to-wavelength variation for daily irradiances, which may be from the smaller number of total observations used to train the model (n=735). Still, the overall trends were very similar for solar noon and daily irradiances, which gives us a higher degree of confidence in the magnitude of the effect cloud cover can have and that this effect is not wavelength dependent.

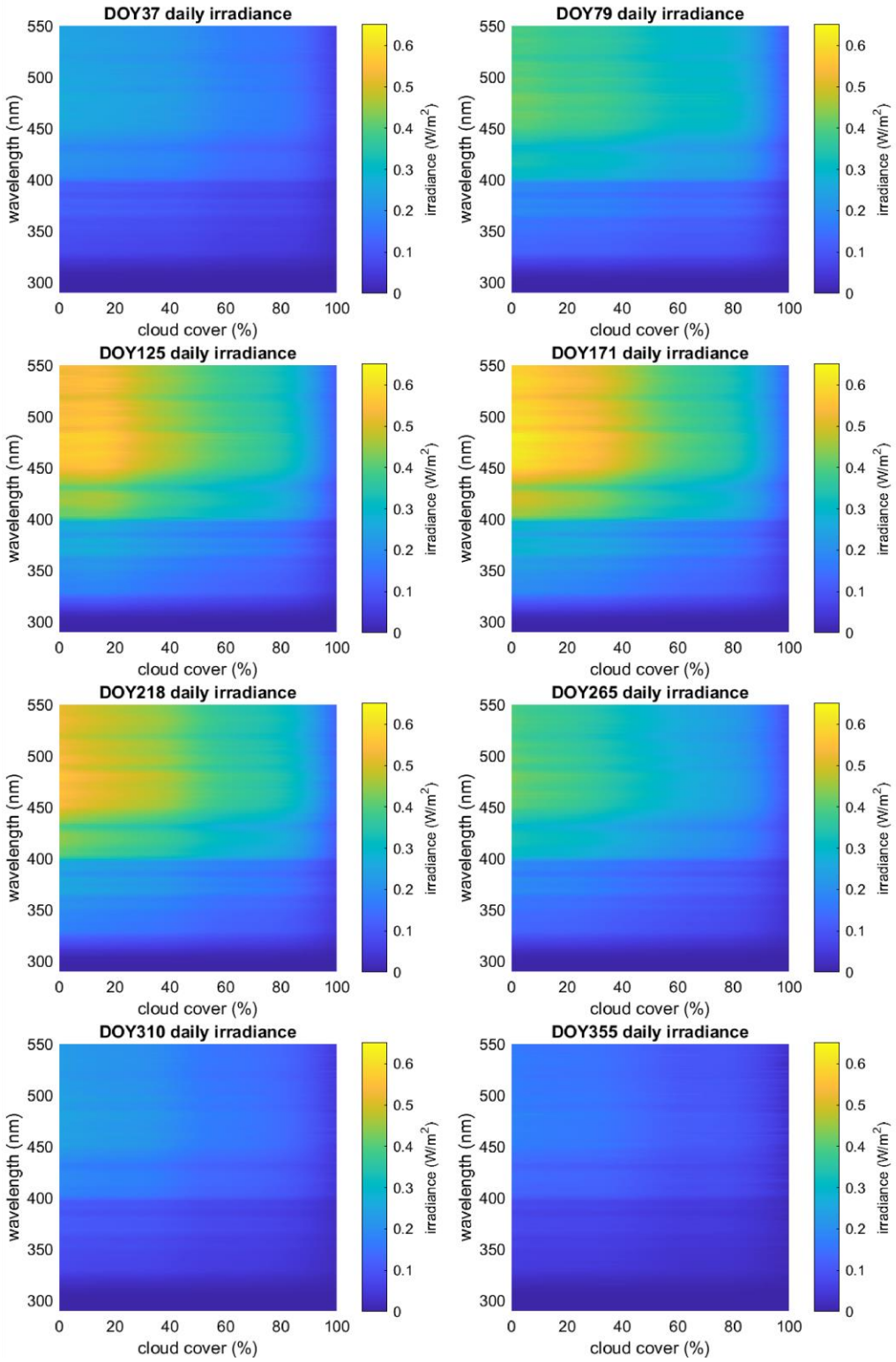

Figure S51. Modeled daily irradiance values as a function of wavelength and cloud cover and eight days of the year (solstices and equinoxes in the right column, the midpoint between those days in the left column).

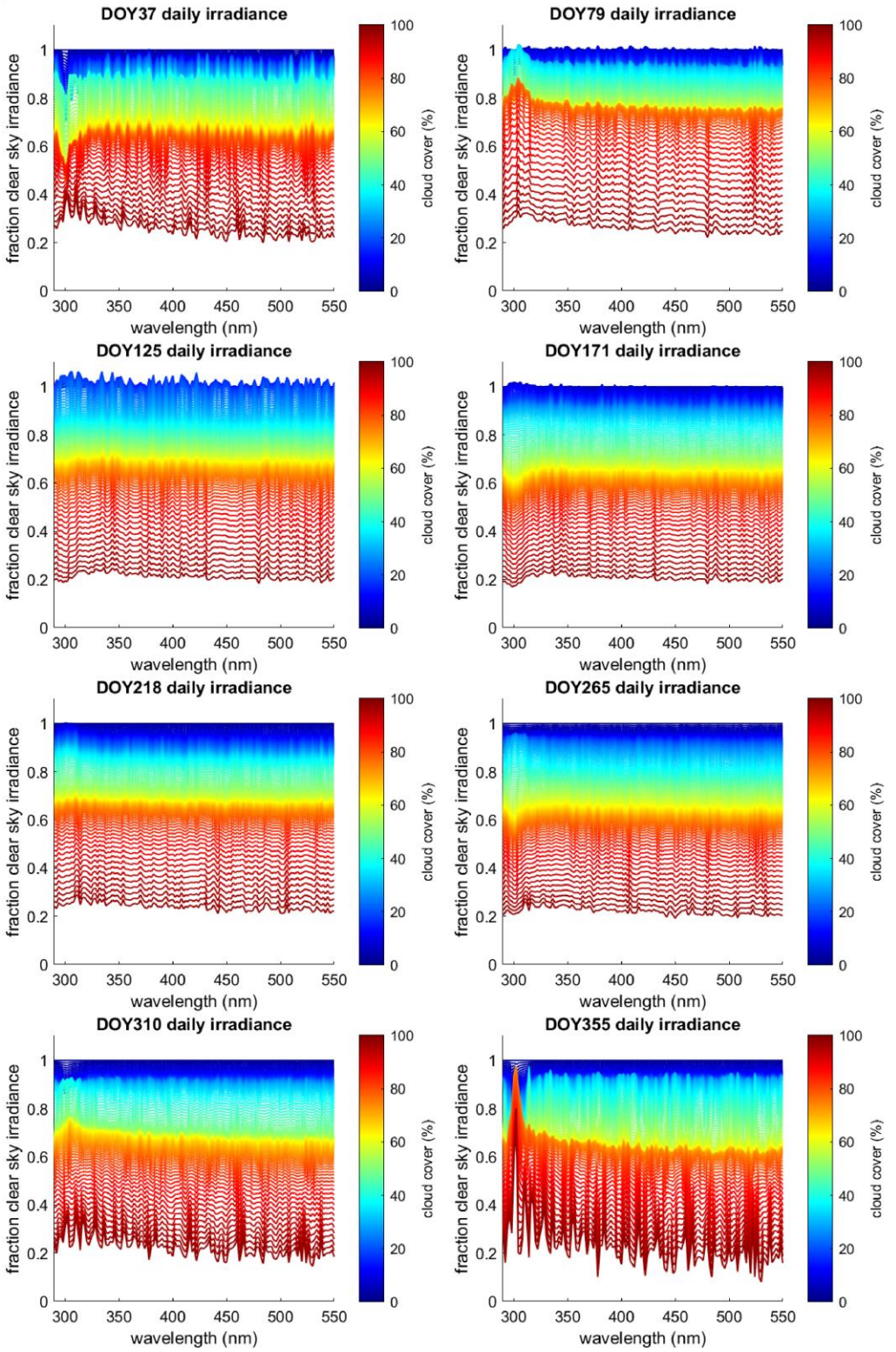

Figure S52. Modeled daily irradiance as a fraction of clear sky values as a function of wavelength and cloud cover and eight days of the year (solstices and equinoxes in the right column, the midpoint between those days in the left column).

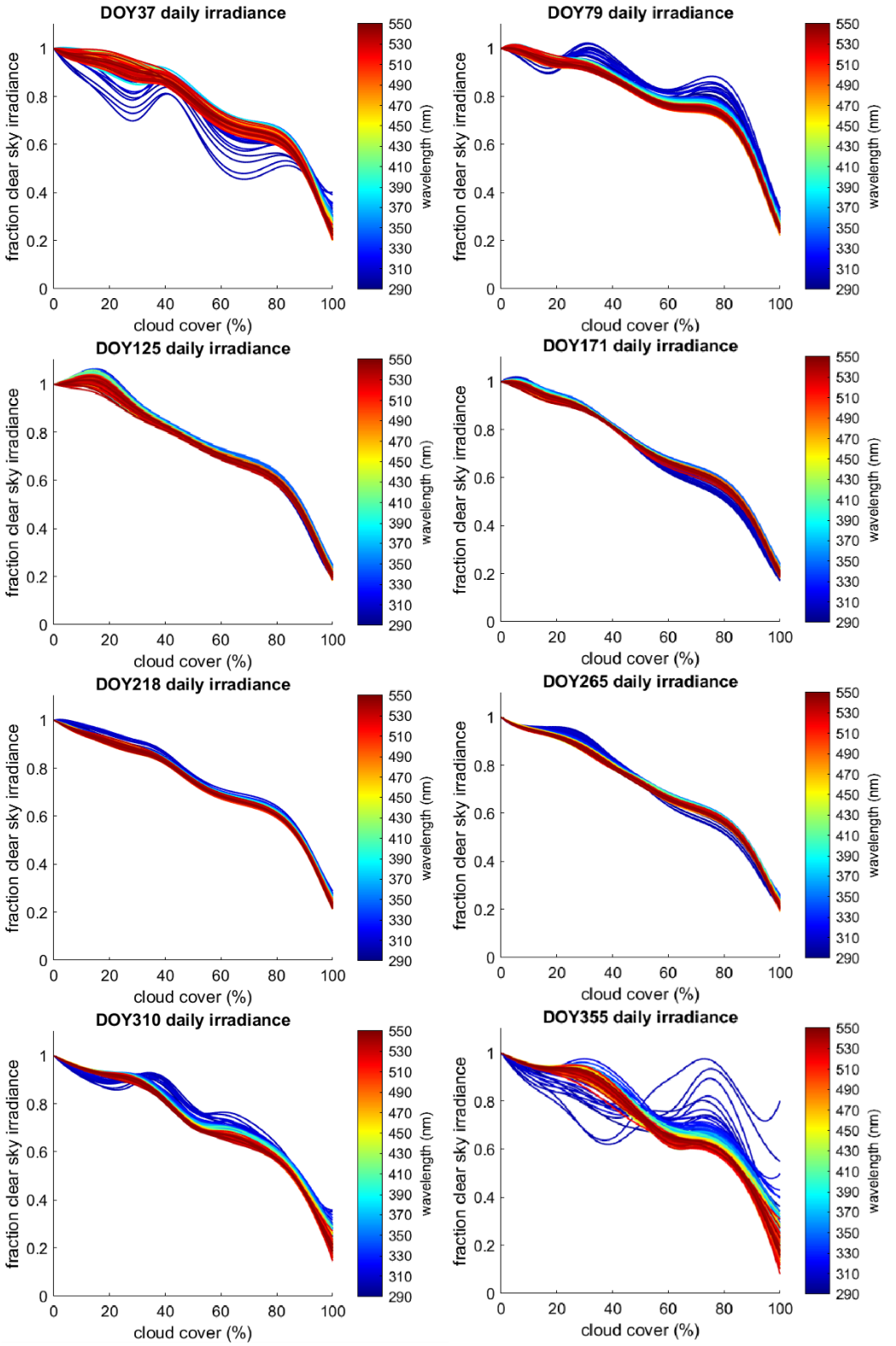

Figure S53. Modeled daily irradiance as a fraction of clear sky values as a function of cloud cover and wavelength for eight days of the year (solstices and equinoxes in the right column, the midpoint between those days in the left column).

Section S12. Neural Network equations for Solar and Daily Irradiance

Table S11. Equations for daily and solar noon irradiance as a function of cloud cover for individual UVB, UVA, and PAR regions and collectively. Variable y is fraction of clear sky irradiance and x is percentage of cloud cover in the sky.

	UVB	UVA	PAR	ALL
Daily	$y=-4.315E-05x^2$ $-0.002340x+1$	$y=-4.039E-05x^2$ $-0.002328x+1$	$y=-4.264E-05x^2$ $-0.002641x+1$	$y=-4.237E-05x^2$ $-0.002425x+1$
Solar Noon	$y=-4.452E-05x^2$ $-0.0005196x+1$	$y=-5.414E-05x^2$ $-0.0005414x+1$	$y=-5.263E-05x^2$ $-0.0003139x+1$	$y=-4.986E-05x^2$ $-0.0003537x+1$

Table S12. Details about the measurements taken from other cloud irradiance studies compared to this work.

Study	Wavelengths (nm)	Spectroradiometer	Cloud Coverage	Location	Sampling Times
Grant et al. (2009) [a]	UVB at 290 nm (258-320 nm)	International Light SED240 with UVB filter	Hemispherical photographs (180^o) field of view	40.43°N West Lafayette, Indiana U.S.A	23 days summer, ~6 runs/day
Lubin and Frederick (1991) [b]	342.5-347.5	Palmer station spectroradiometer	Naked eye	64.46°S Antarctic Peninsula	Spring, 700 sample points
Frederick and Steele (1995) [c]	300-380	Eppley black and white pyranometer	NOAA's data summaries for O'Hare Intl' airport	41.88°N Chicago Illinois U.S.A	0900, 1200, and 1500 April-October
Ilyas (1987) [d]	295-390	Eppley UV radiometer	Unknown	5.2°N Penang, Malaysia	4 years, monthly averages

Schafer et al. (1996) [e]	290-320	Brewer MKIV spectroradiometer, horizontal UV irradiance 286.5-363.0 nm	Wide angle camera	35.66°N Black Mountain NC U.S.A	February-July
Kasten (1980) [f]	global	Moll-Gorczynski pyranometer & Schulze net pyrradiometer	Air Weather Station at Hamburg Fuhlsbüttel Airport (observed hourly)	53.6^{o} N Hamburg, Germany	10 years (1964-1973), 25,521 observations
Josefsson (2000) [g]	global	Kipp & Zonen CM1O/CM1 1 pyranometers[10]	Human observation, nearby airport	58.6°N Norrköping, Sweden	March 1983-Dec 1992, hourly observations (≈35,000)
Bais et al. (1993) [h]	290-325	Brewer spectroradiometer	Unknown (reported hourly from cities airport)	40°N Thessaloniki, Greece	Years
this work	290-700	EKO-WISER spectroradiometer	Yankee Total Sky Imager Model 880 & EKO/CMS-Schreder All Sky Imager Model ASI-16	39.74°N Golden Colorado, U.S.A	October 2017 to December 2020, 64,479 observations

Supporting Information References

1. A. Andreas and T. Stoffel, *NREL Solar Radiation Research Laboratory (SRRL): Baseline Measurement System (BMS); Golden, Colorado (Data)*, 1981.
2. B. Wessels,Validation of a cloud detection algorithm with an All-Sky Imager Masters Energy Science, Utrecht University, 2019
3. SRRL BMS Outdoor Spectral Data, https://midcdmz.nrel.gov/apps/spectra.pl?BMS, (accessed October, 2020).
4. SRRL BMS Daily Plots and Raw Data Files, https://midcdmz.nrel.gov/apps/day.pl?BMS, (accessed October 2020).
5. SRRL ASI-16 Sky Imager Gallery, https://midcdmz.nrel.gov/apps/imagergallery.pl?SRRLASI, (accessed October, 2020).
6. General Solar Position Calculations, https://gml.noaa.gov/grad/solcalc/solareqns.PDF, (accessed October, 2020).
7. SRRL BMS Instrument Descriptions and Histories, https://midcdmz.nrel.gov/srrl_bms/instruments.html, (accessed October, 2020).
8. J. N. Apell and K. McNeill, Updated and validated solar irradiance reference spectra for estimating environmental photodegradation rates, *Environ Sci Process Impacts*, 2019, **21**, 427-437.https://doi.org/10.1039/c8em00478a.
9. W. Josefsson and T. Landelius, Effect of clouds on UV irradiance: As estimated from cloud amount, cloud type, precipitation, global radiation and sunshine duration, *Journal of Geophysical Research: Atmospheres*, 2000, **105**, 4927-4935.https://doi.org/10.1029/1999jd900255.
10. T. Persson, *Measurements of Solar Radiation in Sweden 1983-1998*, SMHI, 2000.https://www.smhi.se/polopoly_fs/1.159515!/RMK_89.pdf.