Site Adaptation for Global Horizontal Irradiance in Argentina Using a Multilayer Perceptron

Germán A. Salazar[1,2], Christian Gueymard[3], Rubén D. Ledesma[1,2] and Olga Castro Vilela[4]

[1] Universidad Nacional de Salta, Salta (Argentina)

[2] Instituto de Investigaciones en Energía No Convencional, Salta (Argentina)

[3] Solar Consulting Services, NH (USA)

[4] Universidade Federal de Pernambuco, Recife (Brazil)

Abstract

Accurately quantifying the spatial and temporal availability of the solar resource is a critical step before the sizing of large solar power plants can be optimized and their bankability established. To remove the inherent biases in time series of satellite-based irradiance estimates, short-term local measurements are used for calibration purposes along with a site adaptation (SA) algorithm. Here, a machine-learning algorithm based on Multilayer Perceptron (MLP) is used to improve estimates of global horizontal irradiance (GHI) at two sites in northwest Argentina. The SA algorithm is trained with one year of hourly data using six regressors, including GHI and cloud cover, to ultimately calibrate the CAMS-Rad satellite-derived estimates that are available in the public domain. The MLP-based adaptive function is then applied to several years of those GHI estimates and compared against ground measurements. Significant-to-large reductions in bias are found, most particularly at the high-elevation site of El Rosal, demonstrating the strength of the MLP approach in SA applications. The goal of this development is to help improve the viability and bankability of solar projects.

Keywords: GHI, site adaptation, MLP, machine learning, elevation, irradiance

1. Introduction

To optimally size future solar power plants, it is necessary to quantify the spatial-temporal magnitude distribution of the local solar resource. If there are no measurement stations properly distributed in the region of interest with at least 10–15 years of quality-controlled data—which is the case in the vast majority of projects—it is possible to use model estimates based on imagery from weather satellites (Sengupta et al., 2024; Gueymard, 2022). These products typically span multiple decades and attempt to address the inherent spatial and temporal variability of solar radiation over large regions with more or less accuracy; e.g., (Salazar et al., 2020).

Evaluating the uncertainty in such modeled irradiance estimates is an important task. It can only be conducted at radiometric stations with high-quality irradiance data after elaborate quality control (e.g., Forstinger et al., 2023). For solar projects of small to medium size, the design and financing phases typically rely on satellite-derived irradiance model estimates only, assuming they can determine the solar resource with reasonable uncertainty. For large projects, however, the investments and risks are considerable, hence financing is often conditional to a period of on-site measurements obtained with a temporary solar resource station. Such a measurement campaign typically lasts about 12 months to sample all seasons. The project's analysts then have two different databases: a time series of at least 15 years of satellite-based irradiance estimates and a short "ground truth" time series of only one year. In that context, the critical goal of the site adaptation (SA) technique is to use the short-term measurements to calibrate the long-term modeled data and decrease the bias in the solar resource estimate that is derived from it.

Various methods have been proposed to operate the SA process, as reviewed in (Polo et al., 2016; Yang and Gueymard, 2021). In essence, an appropriate function is needed to calibrate the model estimates and make them fit the measured values with a high degree of confidence, while enabling a reliable correction of the long-term estimate time series over its whole duration (Figure 1). That site-adapted multiyear time series is then used to evaluate the solar resource characteristics of the project's site with negligible bias.

10.18086/swc.2025.04.40 Available at http://proceedings.ises.org

Figure 1: Site adaptation schematic: A one-year measured database (MDB) is used to find a function to adapt the overlap irradiance values (rectangle a) over a more extensive satellite database (SDB) of, e.g., 15 years. Then that function is applied to the entire SDB, and an adapted database (ADB) is obtained.

The present contribution attempts to improve this calibration methodology by using an advanced approach based on an efficient machine-learning (ML) algorithm, thus constituting a proof of concept. It incorporates irradiance measurements from two radiometric stations in South America where the radiometers and equipment follow strict procedures for calibration, maintenance, and data quality control (Alonso-Suarez et al., 2024).

The focus here is on global horizontal irradiance (GHI), which is the critical resource quantity in solar applications involving nearly horizontal collectors or PV modules, e.g., for installations close to the equator. The two sites selected to test the new approach are in the province of Salta, northwest Argentina (Salta Capital, lat. 24.7°S, and El Rosal, lat. 24.4°S), close to the Tropic of Capricorn. The adaptation algorithm uses GHI and cloud fraction data from the open-source CAMS-Rad satellite-derived database, among other simple variables, and an ML adaptation function using a Multilayer Perceptron (MLP) to reduce the inherent bias in the satellite estimates.

2. Methodology

To sense GHI, an Eppley PSP radiometer is used at the Salta station, whereas a Kipp & Zonen CMP3 is used at El Rosal. Each is connected to a CR1000 datalogger, with a 1-minute record frequency. These sensors are periodically recalibrated against a local standard, which in turn is calibrated against the national Argentinian standard. The latter is traceable to WRR. Quality control filters are applied, following (Nollas et al., 2023), to eliminate questionable data.

The two sites are distant by 150 km and at significantly different elevations. Table 1 provides their geographical and climatic data.

Table 1: Information on the stations under scrutiny here. The climate class refers to the Köppen-Geiger typology.

Site	Latitude	Longitude	Elevation (amsl)	Climate Class	Period
Salta	–24.73°	–65.41°	1233 m	Cwb	2009–2021
El Rosal	–24.39°	–65.77°	3355 m	Cwc	2013–2018, 2023–2024

The adaptative function for each site is derived using an MLP to perform the regression procedure. MLP is a well-known feedforward artificial neural network (ANN) structure. It is a parametric model with a basic structure of an input layer, one or more hidden layers, and an output layer (Rumelhart et al., 1986). The nodes of the hidden and output layers are artificial neurons using a nonlinear activation function. The most common learning algorithm in MLP is backpropagation combined with an optimization algorithm. This algorithm can minimize the error in its output compared to the expected results. For that reason, the measured dataset is subdivided into training, validation, and testing subsets:

- The training set is used to train the MLP model. This work tests various configuration options, as detailed in Section 3. All these options generate many models, which are tested in the final step to determine those with the lowest errors.
- The data used for validation determines when to stop the learning process, using ordinary least squares (OLS) and gradient-descent (GD) algorithms. The validation set is a 20% subset of the training set, consisting of randomly selected elements.
- Finally, the actual accuracy of the method is evaluated with performance metrics based on the testing subset, i.e., using the site-adapted estimates compared to unseen measured data.

It is emphasized that the input data must be normalized first to ensure that all variables contribute proportionally to the learning process, preventing those with larger numerical ranges from dominating the optimization. This stabilization leads to faster convergence and more reliable model training. Although MLP is prone to overfitting issues, there are ways to prevent or mitigate this issue (Abu-Mostafa et al., 2012). Two such techniques are used here (dropout rates and final metrics in the test set), even though a standard methodology for solving this problem has yet to be established.

The theory of Machine or Automatic Learning argues that, in addition to the relationship between the target variable (measured GHI) and the modeled variable (satellite modeled GHI), adding more regressor variables is beneficial if they provide accurate information. Thus, six independent regressors are used here to estimate the local adaptive function from the available data (Table 2). Most importantly, 1-minute model estimates of all-sky GHI and cloud fraction are obtained from the public-domain CAMS-Rad service (https://www.soda-pro.com/web-services/radiation/cams-radiation-service) at 0.14° spatial resolution and 1-minute temporal resolution. Additionally, clear-sky GHI_{cs} estimates are obtained from Version 2 of the ARG-P empirical model based on regional irradiance measurements (Salazar et al, 2010; Ledesma et al, 2022).

At each station, the end product consists of the long-term all-sky GHI time series derived, after site adaptation, from a few variables included in the CAMS-Rad satellite database.

Table 2: Inputs used by the MLP model.

Input #	Regressor	Definition
1	SZA	Solar zenith angle
2	m	Optical airmass
3	I_o	Top-of-atmosphere irradiance
4	GHI_{cs}	ARG-P v2 clear-sky global irradiance
5	GHI	CAMS-Rad all-sky global irradiance
6	CF	Pixel's mean cloud fraction

For Salta, the whole year 2009 (3,796 hourly values) was used for training, of which 760 hourly values were used for the validation set, and 2010–2021 (35,893 hourly values) to test the MLP-adapted time series. For El Rosal, the years used were 2013 (3,632 hourly values) for training, 727 hourly values for the validation set, and both 2014–2018 and 2023–2024 (20,979 hourly values total) for testing.

3. Implementation and results

The ML algorithms are implemented in Python using Spyder 5.0. The Multilayer Perceptron (MLP) model used is trained using a grid search on a predefined set of hyperparameters using the Keras library with a TensorFlow backend. The primary objective is to identify the configuration that minimizes the relative root mean-square error (rRMSE) in the validation set. The parameter grid is defined using five configuration factors, as described below.

i) *Dense layer network*: This parameter specifies the architecture of the fully connected layers in the neural network. Each tuple indicates the number of neurons in successive hidden layers. For example, (5, 10, 15) represents a network with three dense layers containing 5, 10, and 15 neurons, respectively. Increasing the number of layers or neurons generally enhances the model's capacity to learn complex nonlinear relationships. However, this may also raise the risk of overfitting if adequate regularization techniques—such as dropout—are not applied. Here, the tested combinations are (5, 10), (10, 20), and (5, 10, 20).

ii) *Dropout rate*: This defines the fraction of artificial neurons that are randomly "dropped out" (i.e., temporarily deactivated) during each training iteration (Srivastava et al., 2014). A dropout rate of 0.1, for example, means that 10% of the neurons are ignored at each update step. This stochastic regularization technique prevents units from co-adapting too strongly, thereby reducing overfitting and improving the model's generalization ability. Here, the tested options are 0.1 and 0.2.

iii) *Epoch*: The number of epochs corresponds to the total times the learning algorithm passes through the entire training dataset. Fewer epochs may lead to underfitting, while too many can cause overfitting. The optimal number of epochs is usually determined empirically, often in combination with early stopping criteria to prevent unnecessary training once convergence is achieved. Here, the tested options are 30, 60, and 90.

iv) *Batch size*: This indicates the number of training samples processed before the model's internal parameters are updated. Smaller batch sizes typically introduce more stochasticity in gradient estimation, which can help escape local minima but may slow convergence. In contrast, larger batches produce smoother gradients but can reduce generalization capacity. A fixed batch size of 20 is considered here.

v) *Learning rate*: This controls the step size in the optimization process, determining how much the model weights are adjusted in response to the estimated error at each iteration. A high learning rate accelerates convergence but risks overshooting the optimal solution, whereas a low value ensures more stable learning but requires more epochs to reach convergence. Two learning rates are considered here: 0.01 and 0.1.

During the search, each combination of parameters is trained and evaluated independently, resulting in 36 models. Each MLP model obtained in the validation set is subsequently applied to the test set, where the best models are selected to be those with the lowest relative mean bias error (rMBE) and rRMSE. These metrics are used to evaluate the performance of both the original CAMS-Rad estimated GHI and the site-adapted GHI time series vs. the measured GHI. Tables 3 and 4 provide details on the performance results obtained with the best adaptive functions (indicated by the # symbol) among the 36 models.

Using the training dataset, the magnitude of both rMBE and rRMSE decreases compared to the raw satellite data, as expected, but not equally at both sites. This indicates spatial differences in satellite model performance, most likely caused by one or more factors: the site's elevation, the pixel's topographic heterogeneity, and/or the satellite's viewing angle. The inability of MLPs to significantly improve models that already work well might also have an influence (Ledesma and Salazar, 2025).

Table 3: Train and test dataset metrics for Salta. The best metric value in each column is shown in bold italics. DLN: Dense layer network; DR: Dropout rate; LR: Learning rate.

GHI Data/Model	DLN	DR	Epoch	LR	Train set (one year)		Test set (12 years)	
					MBE (%)	RMSE (%)	MBE (%)	RMSE (%)
CAMS-Rad (original)	—	—	—	—	1.0	24.2	4.1	29.2
#7 MLP (site adapted)	(5, 10)	0.2	30	0.1	***–0.2***	24.0	***–0.3***	24.6
#18 MLP (site adapted)	(10, 20)	0.1	90	0.01	–0.3	***23.8***	–0.5	***24.4***
#19 MLP (site adapted)	(10, 20)	0.2	30	0.1	0.9	24.4	0.6	24.8
#20 MLP (site adapted)	(10, 20)	0.2	30	0.01	–0.5	23.9	–0.7	24.5
#25 MLP (site adapted)	(5, 10, 20)	0.1	30	0.1	–0.4	25.2	–0.7	25.6

Table 4: Train and test dataset metrics for El Rosal. The best metric value in each column is shown in bold italics. DLN: Dense layer network; DR: Dropout rate; LR: Learning rate.

GHI Data/Model	DLN	DR	Epoch	LR	Train set (one year)		Test set (12 years)	
					MBE (%)	RMSE (%)	MBE (%)	RMSE (%)
CAMS-Rad (original)	—	—	—	—	–24.3	40.9	–21.6	38.7
#5 MLP (site adapted)	(5, 10)	0.1	90	0.1	–0.9	***19.9***	–0.5	20.8
#27 MLP (site adapted)	(5, 10, 20)	0.1	60	0.1	***0.0***	***19.9***	***0.1***	***20.6***
#19 MLP (site adapted)	(10, 20)	0.2	30	0.1	–0.2	20.9	0.5	22.2
#20 MLP (site adapted)	(10, 20)	0.2	30	0.01	–1.4	***19.9***	–1.1	20.8
#25 MLP (site adapted)	(5, 10, 20)	0.1	30	0.1	–1.0	20.5	–0.6	21.4

The real performance of the proposed approach is determined from the test dataset. Minor discrepancies typically exist between the metrics of the training and test sets for the modeled vs. measured GHI values, mainly because of the sampling effect related to the differences in dataset size (Tables 3 and 4). One desirable criterion to select the best MLP model is its stability, which means that the performance needs to be conserved between the training and testing datasets. At Salta, for instance, model #7 exhibits quasi-identical metrics for both datasets. Overfitting is unlikely because the test set is much larger than the training set. At El Rosal, a similar finding is obtained with model #27.

The bias in the MLP-derived test results is shown in Figure 2 for each MLP model and both experimental sites. Unexpectedly, most models do not behave similarly, site-wise. The two MLP models with the lowest biases are #7 and #18 at Salta, and #5 and #27 at El Rosal, as also indicated in Tables 3 and 4. This site dependence is a hindrance because, in practice, the SA method needs to be applied at unseen sites, i.e., where the best MLP model is not known a priori. To guarantee generalizability, it is thus important to propose an MLP model that generates good results evenly over a large region. In the present case, models #19, 20, and 25 result in only marginally higher rMBE, and similar rRMSE, than the best site-specific models. Those three architectures would thus be recommended for site-adapting CAMS-Rad datasets over the northwest region of Argentina.

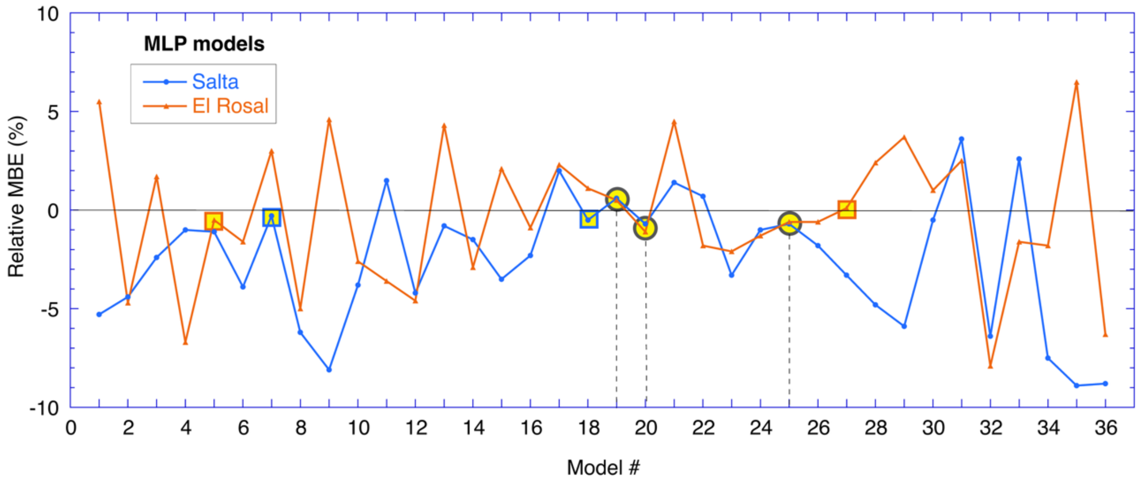

Figure 2: Performance of all 36 MLP models using the testing dataset at the two experimental sites. The square symbols indicate the two best models at each site in terms of rMBE. The circle symbols with vertical dashed lines indicate the three models that result in low bias at both sites.

Interestingly, all metrics show consistent improvements when using the MLP approach compared to the original CAMS-Rad model, which is the desired result. Most importantly, a large error reduction is found at El Rosal, where the long-term bias is reduced from –22% to ≈0%. This is extremely significant from a bankability standpoint, since the solar resource is effectively augmented by a factor of 22% by the SA method. The large rMBE in the original CAMS-Rad data at El Rosal, and the subsequent nullification of the bias after site adaptation, are analogous to what was obtained for another high-elevation site (Izaña, 2373 m amsl) by Yang and Gueymard (2021).

Figure 3 compares the measured GHI values with those modeled by CAMS-Rad and those adapted using MLP (option #7 for Salta and #5 for El Rosal) for the test sets. In the case of El Rosal, MLP clearly manages to “redistribute” the CAMS-Rad GHI data to correct its errors. Overall, this correction looks efficient but imperfect, in large part because of the considerable scatter in the original estimates. A significant correction can also be seen for low GHI values, where the SA method attempts to fix the systematic underestimation there. In the case of Salta, only a smaller bias correction is necessary, improving both the CAMS-Rad underestimation at low GHI and overestimation at high GHI.

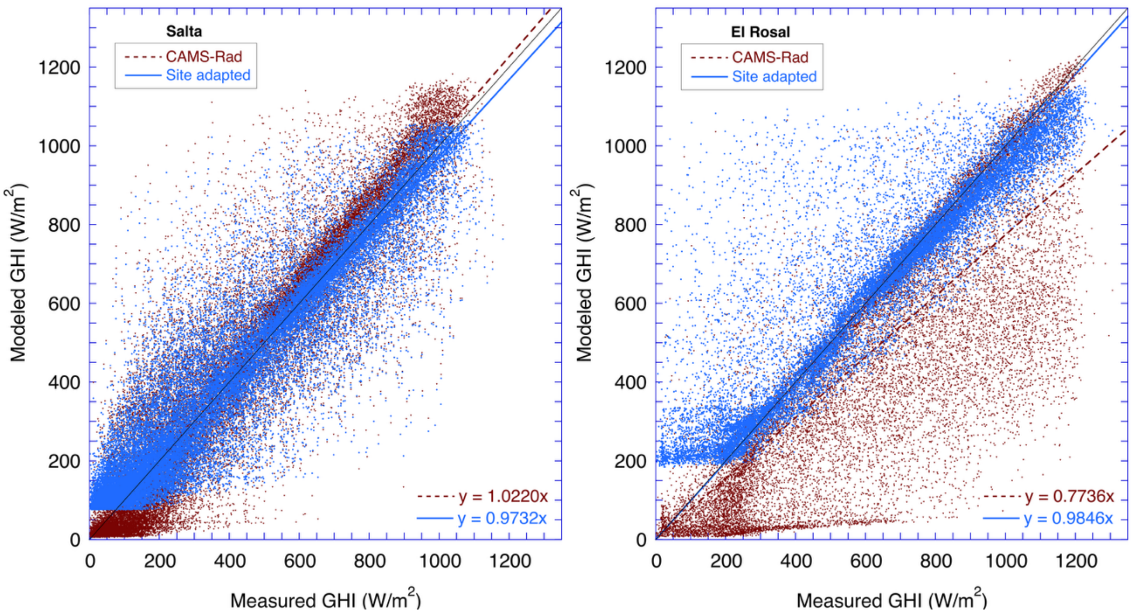

Figure 3: GHI estimates from CAMS-Rad before and after site adaptation compared to the measured at Salta (left) and El Rosal (right). The best-fit lines with no intercept are indicated.

4. Conclusion

This study has developed an innovative machine-learning method to improve satellite-derived estimates of GHI data in northwest Argentina. High-quality measurements from two sites were used to train and test the method: Salta (1233 m amsl) and El Rosal (3355 m amsl). To develop this novel Site Adaptation (SA) method, an optimal correction function was sought using a type of neural network within the ANN family known as Multilayer Perceptron (MLP).

The MLP hyperparameters used within the Python Keras library were similar to those found in the literature. Various options were tested to optimize the MLP, resulting in a total of 36 actual SA model architectures. The GHI datasets were partitioned into Train, Validation, and Testing sets, whose results were analyzed with conventional metrics, relative MBE and RMSE.

At El Rosal, the improvement in the modeled data set was found substantial, decreasing the bias by ≈20% in both the training and test sets. For Salta, only a slight bias improvement was observed. The origin of this improvement difference between the two sites might be caused by the high elevation of El Rosal and the fact that it is located on the edge of the Meteosat satellite imagery used by CAMS-Rad. This is a known factor that typically affects this kind of satellite-derived irradiance modeling, such as performed by the Heliosat model in this case. More research is necessary to investigate whether this issue also affects other satellite-derived databases around the edges of their domain, in comparison with, e.g., reanalysis-based irradiance estimates that do not directly depend on spaceborne observations.

Although MLP might improve data more efficiently than simpler methods, its implementation requires specialized hardware to handle large amounts of data. The computational costs of MLP must be taken into account, as they are not negligible. Moreover, the present results suggest that the optimal MLP architecture might be site-specific to some extent, and that there is no guarantee that the best configuration options obtained here are really optimal, considering the large number of possible configuration combinations that have not been tested yet. Although three specific MLP architectures have been found to work equally well at both experimental sites, the present approach needs to be expanded to more regions and more sources of estimated irradiance databases to verify its generalizability.

In conclusion, the MLP-based SA approach appears to be a promising tool for improving satellite-derived solar irradiance databases, thereby facilitating the bankability of solar power plant projects.

5. References

Abu-Mostafa, Y.S., Magdon-Ismail, M., Lin, H., 2012. Learning from data. A Short Course. AMLbook.com. ISBN 10:1-60049-006-9.

Alonso-Suarez, R. et al., 2024. Recomendaciones y Buenas Prácticas para la Medición y Registro de la Radiación Solar en Territorio. http://les.edu.uy/report/guia-mercosur-medicion-radiacion-solar-vfinal.pdf

Forstinger, A. et al., 2023. Worldwide Benchmark of Modelled Solar Irradiance Data. Report IEA-PVPS T16-05.

Gueymard, C.A., 2022. Solar Radiation Resource: Measurement, Modeling, and Methods. In *Comprehensive Renewable Energy*, 2nd Edition, T.M. Letcher (Ed.), Elsevier.

Ledesma, R., Salazar, G. A., 2025. Machine learning for site adaptation of satellite-derived solar irradiance in Northwestern Argentina. *Journal of Computer Science*. Advance online publication. https://doi.org/10.13140/RG.2.2.29111.02721.

Ledesma, R.D., Salazar, G.A., Castro Vilela, O., 2022. ARGP-v2 Un modelo práctico para la estimación de irradiancies global horizontal en condiciones de cielo claro para sitios de altura. *Avances en Energías Renovables y Medio Ambiente* 26, 283–289, ISSN 2796-8111 (in Spanish).

Nollas, F.M., Salazar, G.A., Gueymard, C.A., 2023. Quality control procedure for 1-minute pyranometric measurements of global and shadowband-based diffuse solar irradiance. *Renewable Energy* 202, 40–55.

Polo, J., et al., 2016. Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets. *Solar Energy* 132, 25–37.

Rumelhart, D., Hinton, G., Williams, R., 1986. Learning representations by back-propagating errors. *Nature* 323, 533–536.

Salazar G., et al., 2020. Solar irradiance time series derived from high-quality measurements, satellite-based models, and reanalyses at a near-equatorial site in Brazil. *Renewable and Sustainable Energy Reviews* 117, 109478.

Salazar, G.A, Hernández, A.L., Saravia, L.R., 2010. Practical models to estimate horizontal irradiance in clear sky conditions: Preliminary results. *Renewable Energy* 35, 2452–2460.

Sengupta, M., et al., 2024. Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications: Fourth Edition. Tech. Report NREL/TP-5D00-88300, doi:10.2172/2448063.

Srivastava, N., et al., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1, 1929–1958.

Yang, D., Gueymard, C.A., 2021. Probabilistic post-processing of gridded atmospheric variables and its application to site adaptation of shortwave solar radiation. *Solar Energy* 225, 427–443.