Received: 4 September 2020 | Revised: 29 October 2021 | Accepted: 21 November 2021

DOI: 10.1002/env.2712

RESEARCH ARTICLE

WILEY

A nonstationary and non-Gaussian moving average model for solar irradiance

Wenqi Zhang[1] | **William Kleiber**[1] | **Bri-Mathias Hodge**[2,3] | **Barry Mather**[3]

[1]Applied Mathematics Department, University of Colorado at Boulder, Boulder, Colorado, USA

[2]Electrical, Computer and Energy Engineering Department, University of Colorado at Boulder, Boulder, Colorado, USA

[3]Power Systems Engineering Center, National Renewable Energy Laboratory, Golden, Colorado, USA

Correspondence
Wenqi Zhang, Applied Mathematics Department, University of Colorado at Boulder, Boulder, CO 80309, USA.
Email: wenqi.zhang@colorado.edu

Funding information
National Science Foundation, Grant/Award Numbers: DMS-1811294, DMS-1923062; U.S. Department of Energy, Grant/Award Number: DE-AC36-08GO28308

Abstract

Historically, power has flowed from large power plants to customers. Increasing penetration of distributed energy resources such as solar power from rooftop photovoltaic has made the distribution network a two-way-street with power being generated at the customer level. The incorporation of renewables introduces additional uncertainty and variability into the power grid. Distribution network operation studies are being adapted to include renewables; however, such studies require high quality solar irradiance data that adequately reflect realistic meteorological variability. Data from satellite-based products are spatially complete, but temporally coarse, whereas solar irradiances exhibit high frequency variation at very fine timescales. We propose a new stochastic method for temporally downscaling global horizontal irradiance (GHI) to 1 min resolution, but we do not consider the spatial aspect due to limited availability of the in situ irradiance measurements. Solar irradiance's first and second-order structures vary diurnally and seasonally, and our model adapts to such nonstationarity. Empirical irradiance data exhibits highly non-Gaussian behavior; we develop a nonstationary and non-Gaussian moving average model that is shown to capture realistic solar variability at multiple timescales. We also propose a new estimation scheme based on Cholesky factors of empirical autocovariance matrices, bypassing difficult and inaccessible likelihood-based approaches. The model is demonstrated for a case study of three locations that are located in diverse climates through the United States. The model is compared against competitors from the literature and is shown to provide better uncertainty and variability quantification on testing data.

KEYWORDS

Cholesky factor, high-resolution, statistical temporal downscaling

1 | INTRODUCTION

Renewable energy integration studies are increasingly used as renewable energy technologies, in particular solar power, are adopted in greater quantities. Such studies provide an analytical framework for evaluating a power system with high levels of variable renewable resources. One of the primary goals of a renewable energy integration study is to address stakeholder concerns that a power system can operate both reliably and efficiently with high penetration of renewable resources. As a major renewable resource, solar irradiance data are required at high temporal resolutions, but direct

observational irradiance data are scarce and are geographically limited, so simulated irradiances are often necessary. Moreover, since solar power integration studies examine future scenarios with larger amounts of installed solar power, simulated irradiances become an integral part of these studies. It crucial that simulated irradiances accurately represent variability at multiple timescales so that the impact on the power system is appropriately represented in grid planning studies. For power system integration studies, solar power output is often needed at the minute or second resolution. Because the in situ irradiance measurements are limited, there is a need to combine available historical in situ measurements with satellite-based irradiance estimates as the basis for generating solar irradiance time series used in such studies.

In the literature of solar irradiance modeling, two major aspects for electric grid applications are forecasting and resources assessment. Solar forecasting approaches enable grid operators to better predict how much solar energy will be added to the grid while solar resource assessment provides the means to accurately determine the availability of power generation on different time scales. There is a rich literature on solar forecasting techniques, including some statistical approaches (Das et al., 2021; Iversen et al., 2014). Our temporal downscaling approach pertains to solar resource assessment. Thus, we focus on a the modeling problem, and not forecasting. Producing simulated solar irradiance data is a problem that has been of interest for many years. Balouktsis and Tsalides (1986) simulate hourly solar insolation data in Athens, Greece using a method in which annual and daily periodicities are first removed to create a stationary random process that is simulated and then retransformed into a time series. Aguiar et al. (1988) use a Markov approach to produce daily solar radiation values based on radiation data from over 300 months of data at nine different locations. Graham and Hollands (1990) treat atmospheric transmittance as a substitute variable for irradiance and simulate its value by decomposing it into a trend and random component to produce hourly irradiance values based on data from three Canadian cities with different climate characteristics. Glasbey and Allcroft (2008) propose a spatiotemporal auto-regressive moving average model with a Matérn noise process for simulating realizations of energy output. Hocaoglu (2011) utilize hidden Markov models based on the relationship between ambient temperature and solar radiation to generate time series data for two cities in Turkey. Ngoko et al. (2014) also use Markov models to generate synthetic minutely solar irradiance data from hourly weather data but they started from daily clearness index values. A similar approach was taken in Bright et al. (2015), though they included cloud cover scenarios in the model. Lave et al. (2013) utilized a wavelet-based model to simulate the power output of a single solar PV plant from a single irradiance sensor, similar to Marcos et al. (2011). Resampling and clustering techniques are also used in the solar irradiance downscaling literature (Lave et al., 2012; Perez & Fthenakis, 2015). In Zhang et al. (2018a), synthetic datasets are produced using log-additive non-Gaussian mixture models and resampling techniques. Zhang et al. (2018b) extends the work of Zhang et al. (2018a) and proposes a stochastic method for generating synthetic data using properly transformed mixtures of random variables and a clear-day rule to determine when stochastic downscaling is applied.

Our goal is to conditionally model solar irradiance given satellite-based estimates at a single spatial location. In particular, the National Solar Radiation Database (NSRDB; Sengupta et al., 2018) contains estimates of multiple types of irradiance at a 4 km spatial resolution and 30 min time resolution over the continental United States for the period 1998–2016. The NSRDB data product combines satellite data, numerical weather prediction data, as well as in situ measurements. However, the 30 min resolution is limited in its use for understanding the impact of irradiance variability on certain types of grid operations. Given the NSRDB estimate, we produce conditional realizations of 1-min irradiance that capture plausible high frequency trajectories consistent with the NSRDB. The statistical challenges are comprised of handling diurnal and seasonal variations in both mean and covariance structure, as well as the non-Gaussianity of irradiance.

The statistically novel aspects of this work include a new nonstationary and non-Gaussian moving average model for stochastic temporal downscaling. Moving average models constitute a classic approach to modeling stationary processes, while nonstationarity is traditionally accommodated through differencing, forming the celebrated autoregressive integrated moving average (ARIMA) class. However, differencing is a fairly restrictive type of nonstationarity which Box et al. (2015) terms homogeneous nonstationary processes. There are a large number of approaches to handling nonstationary and non-Gaussian processes in the time series literature. Some involve adapting Gaussian processes within the ARIMA framework Beran (1995), but at the cost of computationally expensive estimation algorithms. Spectral approaches for non-Gaussian processes are considered by Lii and Rosenblatt (1982), Cheng (1990), and Lii and Rosenblatt (1992), while spectral nonstationarity is taken up in Velasco (1999b), Velasco (1999a), and Robinson (1995). The Whittle pseudo-maximum likelihood estimator for Gaussian processes was extended to the nonstationary setting in Velasco and Robinson (2000). Chapman et al. (2020) uses a nonparametric approach to locally estimate the variance

based on a wavelet framework. In Porcu et al. (2020), a detailed discussion on a bilevel skew-t stochastic model for handling non-Gaussianity in wind field speed data is given. In this work we consider an alternative approach to handling nonstationarity, non-Gaussianity and estimation within the ARMA framework simultaneously.

Our estimation idea bypasses difficult likelihood-based approaches, and relies on modeling Cholesky factors of autocovariance matrices. The Cholesky factor of an autocovariance matrix has two attractive properties: first, its inverse is a decorrelating transformation on the time series, and second, its entries can be interpreted as parameters in an inhomogeneous moving average representation. Statistical modeling of the Cholesky matrix has primarily been explored in the longitudinal data analysis literature. Pinheiro and Bates (1996) propose unconstrained parameterizations of a covariance matrix using Cholesky decomposition. Pourahmadi (1999) start with the modified Cholesky decomposition of the precision matrix, and then parameterize the covariance matrix as a linear combination of covariates. Zhang and Leng (2012) decompose the covariance matrix into generalized moving average parameters in the moving average Cholesky decomposition, while Lee et al. (2017) combine the autoregressive and moving average modeling of the covariance matrix. Our approach to exploiting the Cholesky factor for estimation is novel, easily allows for incorporation into a non-Gaussian setup and also for nonstationarity.

The temporal downscaling model is broken into two components: the first is a stochastic decision rule that determines whether a given time period will be modeled as clear or non-clear (e.g., cloudy). In non-clear time periods, the model is a correlated mixture of Laplace random variables whose coefficients vary diurnally and seasonally. As direct likelihood calculations are inaccessible, we introduce a new estimation approach exploiting empirical Cholesky factors and decorrelated residuals. Our example focuses on location Eugene, Oregon from year 2010 to year 2013 and uses 30 min estimates of global horizontal irradiance (GHI) from the National Solar Radiation Database (NSRDB; Sengupta et al., 2018), combined with 1 min in situ pyranometer measurements of GHI from a subset of the University of Oregon's Solar Radiation Monitoring Laboratory. The model is trained based on data of years 2011–2013 and is validated on year 2010. The proposed methodology is further validated on another two locations from the Surface Radiation Budget Network (SURFRAD). The outputs of our model are conditional ensembles of downscaled irradiance at the 1 min time resolution. The model is compared against existing models from the solar literature on held-out testing data, and exhibits good coverage properties as well as statistical representations of time dependence.

The remainder of this article is organized as follows. In Section 2, the data sources are introduced. In Section 3, we formulate the general model for temporal downscaling and then a detailed estimation procedure is described. Section 4 includes model comparison and detailed validation of the proposed methodology. Section 5 concludes with discussion and future work.

2 | DATA

The proposed model was originally trained based on one in situ observational time series dataset at Eugene, OR. The data include minutely observations of irradiance on a flat surface together with half-hourly satellite-derived GHI and clear-sky GHI. The pyranometer measurements are from the University of Oregon's Solar Radiation Monitoring Laboratory (SRML; Vignola & Perez, 2004). The data covers a period of 4 years from January 1, 2010 to December 31, 2013. We divide the period into a training and test set, with the training set covering the last three years and the test set the first year. The method is additionally validated on two other locations from SURFRAD. SURFRAD is a network of high quality observation stations supported by NOAA's Office of Global Programs (Augustine et al., 2000). These locations are located in variable climates across the continental USA. Both the SRML and SURFRD data are available at 1-min resolution. Table 1 summarizes the data used in this article.

Let $Y(t, d)$ denote GHI at minute of the day $t = 0, \dots, 1439$ and day of the year $d = 1, \dots, 365$. The notation t denotes minute-level data throughout the article. The NSRDB contains two estimates of irradiance at 30 min snapshots for each day of the year: GHI and clear sky GHI. Clear sky GHI is the amount of irradiance that would theoretically reach the ground under clear conditions (that is, no clouds or aerosols such as dust), while GHI is the estimate based on an empirical-physical model that incorporates satellite and in situ measurements as well as numerical weather prediction model output. Denote by $X(t, d)$ and $X_C(t, d)$ the linearly interpolated NSRDB estimated GHI and clear sky GHI, respectively. The choice of linear interpolation is simply to put the coarse-time NSRDB estimates at the same time resolution as the in situ data $Y(t, d)$, and was seen to work as well as more complicated interpolation schemes in exploratory analyses.

TABLE 1 Data used to train and validate the model

Database	Source	Training period	Testing period	Observations
Half-hourly data sets				
Eugene, OR	NSRDB	January 1, 2011 to December 31, 2013	January 1, 2010 to December 31, 2010	17,520
Boulder, CO	NSRDB	January 1, 2017 to December 31, 2017	January 1, 2018 to December 31, 2018	17,520
PSU, PA	NSRDB	January 1, 2016 to December 31, 2018	January 1, 2019 to December 31, 2019	17,520
1-min data sets				
Eugene, OR	SRML	January 1, 2011 to December 31, 2013	January 1, 2010 to December 31, 2010	525,600
Boulder, CO	SURFRAD	January 1, 2016 to December 31, 2018	January 1, 2019 to December 31, 2019	525,600
PSU, PA	SURFRAD	January 1, 2016 to December 31, 2018	January 1, 2019 to December 31, 2019	525,600

Note: All radiation measurements are in watts per meter squared.

3 | METHODOLOGY

In this section, we present our temporal downscaling framework along with exact details and specifications for our data example. In particular, we detail the two components of the downscaling model.

The basic idea behind the model is straightforward—for any time point we classify irradiance into one of three categories: (1) a clear period where clear sky GHI is achieved, (2) a non-clear period where a stochastic model provides variability, and (3) a non-clear period where effects such as cloud enhancement generate GHI no more than 20% higher than the current clear sky GHI. The cutoff of 20% is a standard level for such clear sky GHI exceedances (Zhang et al., 2018a). Note we only consider irradiance modeling during the day when it is strictly positive.

We propose the following log-additive temporal downscaling model,

$$\log Y(t,d) = Z(t,d)\log(X_C(t,d)) + (1 - Z(t,d))\min\left\{\log(X(t,d)) + \varepsilon(t,d), \log(1.2X_C(t,d))\right\} \quad (1)$$

where $\varepsilon(t,d)$ is the stochastic term that models variability due to extant atmospheric conditions and $Z(t,d)$ is a time series of Bernoulli random variables that determines whether a given time point is clear or not. When the sky is clear, we have $Z(t,d) = 1$, in which case $\log Y(t,d) = \log(X_C(t,d))$. If the sky is not clear, that is, $Z(t,d) = 0$, GHI is generated based on NSRDB GHI estimate with variability provided by our stochastic model, with the restriction that GHI cannot go beyond 120% of clear sky GHI. This log-additive model always results in a nonnegative GHI, and flips back and forth between clear and non-clear time periods. Additionally, our log-additive model is equivalent to modeling the so-called clear sky index that is common in the solar energy literature (Bright et al., 2015). Sections 3.1 and 3.2 detail the clear sky classification model and the stochastic time series model for $\varepsilon(t,d)$, respectively.

3.1 | Clear sky classification

The first step is determining whether a given time point is clear or not, by specifying the model for $Z(t,d)$. There is auxiliary information available in the NSRDB, apart from estimates of GHI, namely current cloud conditions represented by 13 classes (at each 30 min snapshot). We aggregate these 13 classes into two groups of clear and not clear (that includes water and ice) according to Zhang et al. (2018a). We use a logistic regression framework for the binary random variable $Z(t,d)$, incorporating cloud type from NSRDB as a predictor.

The NSRDB cloud type predictors are only available every 30 min. Thus, to fully describe the stochastic model for the binary process $Z(t,d)$, we require some additional notation. Denote by n_d the number of half-hour intervals during daylight hours on day d and let $I_1, \ldots, I_{n_d}$ be sets of minutes in half-hour intervals for this day that center on the hourly and half-hourly satellite data (e.g., $I_k = \{705, \ldots, 734\}$ represents 11:45 a.m.–12:14 p.m. for the satellite data available at $t = 720$, 12 p.m.). Note that $Z(t,d)$ with $t \in I_k$ and $k = 1, \ldots, n_d$ is just the binary response variable for interval I_k on day d with two situations of clear, $Z(t,d) = 1$ and not clear, $Z(t,d) = 0$.

TABLE 2 Maximum likelihood estimates with standard errors in parentheses for parameters of the clear/non clear interval logistic regression model

Parameter	ζ_0	ζ_1	ζ_2	ζ_3	ζ_4	ζ_5
Estimate	−4.35 (0.32)	3.78 (0.38)	1.95 (0.36)	2.01 (0.38)	−0.66 (0.03)	−0.30 (0.03)

Consider a triplet of intervals (I_{k-1}, I_k, I_{k+1}). Heuristically, if the NSRDB's cloud type in all three intervals are clear, we would expect to be more confident that I_k is indeed clear, that is, $Z(t,d) = 1$ for $t \in I_k$. On the other hand, if I_k is preceded or followed by a non-clear interval, then we would expect less confidence in I_k being clear. Thus, for interval k let C_k denote the clear classification from the NRSDB, while N_k denotes a non-clear classification; note that $C_k = C_k(d)$ depends on the day, which we drop from notation for ease of exposition.

The classification model for $Z(t,d)$ is based on a logistic regression with mean function

$$\zeta_0 + \zeta_1 \mathbb{1}_{[(C_{k-1},C_k,C_{k+1})]} + \zeta_2 \mathbb{1}_{[(C_{k-1},C_k,N_{k+1})]} + \zeta_3 \mathbb{1}_{[(N_{k-1},C_k,C_{k+1})]} + \zeta_4 \cos\left(\frac{2\pi d}{365}\right) + \zeta_5 \sin\left(\frac{2\pi d}{365}\right), \tag{2}$$

where $\mathbb{1}_{[(C_{k-1},C_k,C_{k+1})]}$, for example, is an indicator function of three clear sky NSRDB classifications corresponding to the triplet (I_{k-1}, I_k, I_{k+1}). The inclusion of the harmonics allows for seasonally varying probabilities. We do not include triplets of the form (N_{k-1}, C_k, N_{k+1}) due to separability. To determine the classification threshold, we perform a receiver operating characteristic (ROC) analysis and compute the Youden index. The threshold associated with the highest Youden index is 0.33, which we adopt for the ensuing analyses.

The available GHI data are not already classified into clear/not clear for training purposes. Thus, we develop an ad hoc classification rule to generate binary training data to estimate the parameters in (2). For each interval, we generate the so-called clear sky index $K(t,d) = Y(t,d)/X_C(t,d)$ which is a standard quantity in the solar energy literature (Bright et al., 2015). Then, we set $Z(t,d) = 1$ for $t \in I_k$if $\min_{t\in I_k} K(t,d) > 0.8$ and $\max_{t\in I_k} K(t,d) - \min_{t\in I_k} K(t,d) < 0.05$, and $Z(t,d) = 0$ otherwise. These two criteria are suggested by requiring the observed GHI to be similar to the theoretical clear sky GHI, and are seen to work well empirically, although other criteria can be imagined.

Parameters in (2) are fit by maximum likelihood on data from Eugene, OR during the years 2011–2013. Table 2 contains these estimates and standard errors based on the training data. As intuition would suggest, the effect of a triple of clear NSRDB classifications yields stronger influence on the probability of a true clear interval versus cases with neighboring non-clear NSRDB predictors.

Since the temporal resolution of the satellited-derived data is half-hourly, we only utilize the model (2) for $Z(t,d)$ at time points t in the center of a given interval I_k. It is not realistic that the transitions to/from $Z(t,d) = 0$ or $Z(t,d) = 1$ happen at exactly certain minutes before or after the hour/half-hour; to this end we propose randomizing the transition point. Exploratory analyses suggests a uniform distribution for the transition time is appropriate, which we adopt. To be explicit, let c_{k-1} and c_k denote the midpoints of intervals I_{k-1} and I_k, respectively. If interval I_{k-1} is classified as $Z(c_{k-1}, d) = 0$, and I_k as $Z(c_k, d) = 1$, then the transition time between the two regimes is sampled from $U(c_{k-1}, c_k)$. The time for transition from $Z(c_{k-1}, d) = 0$ to $Z(c_k, d) = 1$ is analogous.

3.2 | A nonstationary and non-Gaussian time series model

In the situation that a given time point with $Z(t,d) = 0$ and GHI is less than 120% of the clear sky GHI, model (1) includes the process $\varepsilon(t,d)$, to which we now turn. The first step is to disentangle diurnal and seasonal heteroskedasticity from correlation and other distributional assumptions, in particular we suppose

$$\varepsilon(t,d) = \sigma(t,d)\xi(t,d) \tag{3}$$

where Var $\xi(t,d) = 1$ for all t and d, and $\sigma(t,d)$ is a time and day-varying standard deviation (SD) function. The model for $\xi(t,d)$ is a nonstationary (and possibly non-Gaussian) moving average process

$$\xi(t,d) = \sum_{\ell=0}^{\infty} \alpha_\ell(t,d)\omega(t-\ell,d) \tag{4}$$

for some white noise process $\omega(\cdot,\cdot)$ and real-valued functions $\{\alpha_\ell(\cdot,\cdot)\}_\ell$. It is well known that stationary moving average (MA) and autoregressive (AR) processes are identical under an infinite lag representation. We use an MA specification in (4), rather than a more traditional AR specification, due to a physical motivation. Specifically, solar irradiance is affected by local cloud and aerosols that are moving across the domain. The direct effect on irradiance at a given site is a weighted combination of these clouds/aerosols, which is consistent with a MA representation, if we view ω as representing atmospheric phenomena.

3.2.1 | Modeling heteroskedasticity

We turn to the model for $\sigma(t,d)$ in (3). To assess the possibility of heteroskedastic residuals, we take empirical residuals $\varepsilon(t,d) = \log(Y(t,d)/X(t,d))$ and create empirical SDs that vary over time of day, and day of year. Specifically, we estimate the variability at time point t on day d, $\hat{\sigma}(t,d)$ as the empirical SD of residuals for time points in $[t-14, t+15]$ on days in $[d-14, d+15]$ using data from Eugene, OR during 2011–2013. The windowing in minutes and over days represents an assumption that the distribution of residuals varies slowly over time, and within a ± 15 min, or day window the second-order structure of the residuals does not change substantially. We leave out time intervals with data at the boundary where $t-14$ or $t-15$ have zeros due to morning/night.

Figure 1 shows a heat-map of $\{\hat{\sigma}(t,d)\}$ where the x-axis represents time of day (minutes) and the y-axis indexes day of year. Note that, due to the moving window empirical estimation procedure, empirical estimates are not available for the first and final 15 days of the year (and similarly truncated in morning and evening); however, our fitted model extrapolates values to these cases so that simulations can always be created. There are clear diurnal and seasonal cycles which is suggestive of a (log) linear model with periodicity whose frequencies respect the diurnal and seasonal variation. Thus, the form of a log linear model is

$$\log\sigma(t,d) = \beta_0 + \beta_1\cos\left(\frac{2\pi t}{1440} + \delta_1\right) + \beta_2\cos\left(\frac{2\pi d}{365} + \delta_2\right) + \beta_3\cos\left(\frac{2\pi t}{1440} + \delta_1\right)\cos\left(\frac{2\pi d}{365} + \delta_2\right), \tag{5}$$

where $\beta_0, \beta_1, \beta_2, \beta_3$ are coefficients for each parameter and δ_1, δ_2 are phase shifts. We estimate the model parameters by minimizing a sum of squared difference over $t = 1, \ldots, 1440$ and $d = 1, \ldots, 365$, noting that times for which irradiance is exactly zero (which depends on the day of year) are discarded. The empirical and fitted heatmaps are shown in Figure 1. The oval shape of non-missing values is due to seasonality in the Pacific Northwest, with more sunlight hours in the summer and fewer in the winter. The heatmap suggests enhanced variability of residuals at the shoulders of the day, in the morning and evenings; the proposed harmonic model captures the interaction between daylight hours and seasonality.

3.2.2 | Moving average coefficients

We now turn to the details of the moving average model (4), beginning with specification and estimation of the nonstationary moving average parameters $\{\alpha_\ell(\cdot,\cdot)\}_\ell$. Note there are 30 data points $\hat{\boldsymbol{\xi}}(t,d)$ for a given d and all $t \in I_k$; concatenate these into a time-ordered vector $\hat{\boldsymbol{\xi}}(I_k, d)$. Define an averaged interval and day specific empirical covariance matrix as

$$\hat{\Sigma}(I_k,d) = \frac{1}{100}\sum_{i=d-49}^{d+50}(\hat{\boldsymbol{\xi}}(I_k,i) - \overline{\boldsymbol{\xi}}(I_k,d))^{\mathrm{T}}(\hat{\boldsymbol{\xi}}(I_k,i) - \overline{\boldsymbol{\xi}}(I_k,d)), \tag{6}$$

where $\overline{\boldsymbol{\xi}}(I_k,d)$ is the empirical mean of $\hat{\boldsymbol{\xi}}(I_k,d)$. As we have three years of training data we average the empirical covariance matrices from all three before moving forward.

One can interpret the last row of the Cholesky factor of $\hat{\Sigma}(I_k,d)$ as empirical estimates of $\alpha_0(t,d), \ldots, \alpha_{29}(t,d)$. Figure 2 shows a smoothed version of the first 15 empirical moving average coefficients averaged across different seasons at different half-hourly time intervals; the smoothing is simply to help the eye identify patterns in the coefficients. Each plot stands for a representative morning (10:00 a.m.–10:29 a.m.), noon (12:00 p.m.–12:29 p.m.), and afternoon (2:00 p.m.–2:29 p.m.). In each plot, different colors represents different seasons. Plotting these final rows for varying I_k and d suggests that

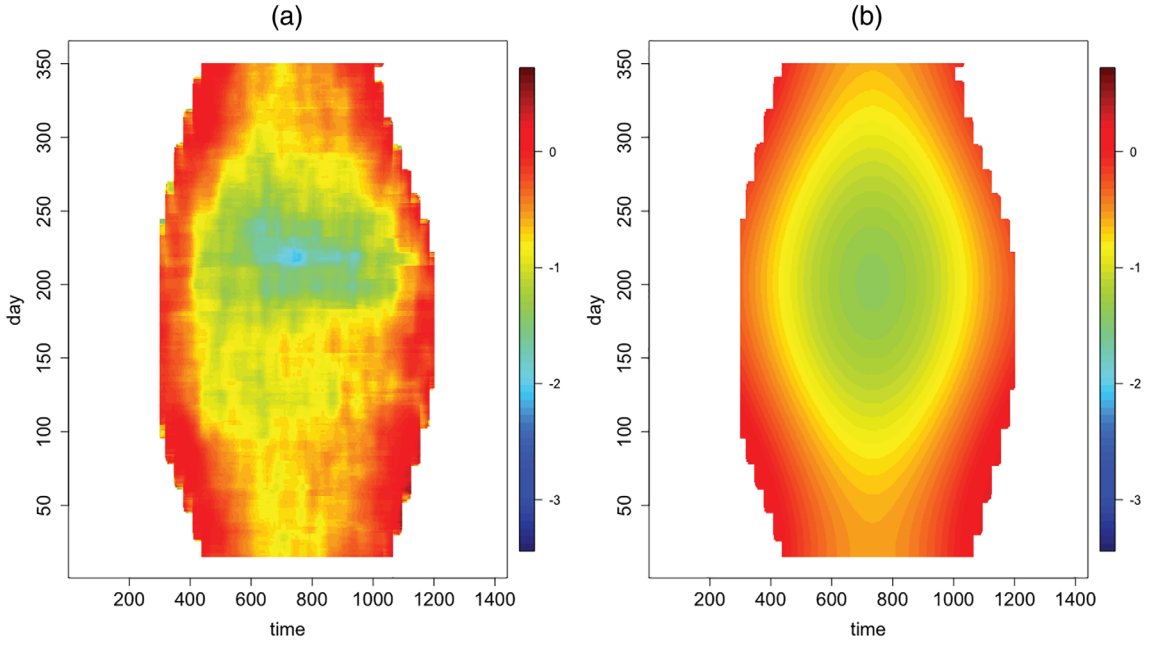

FIGURE 1 Panels (a) and (b) show the heatmaps of the empirical estimator $\log \hat{\sigma}(t, d)$ and modeled log SD $\log \sigma(t, d)$ (b), respectively. The unit of measure of the x-axis is minute

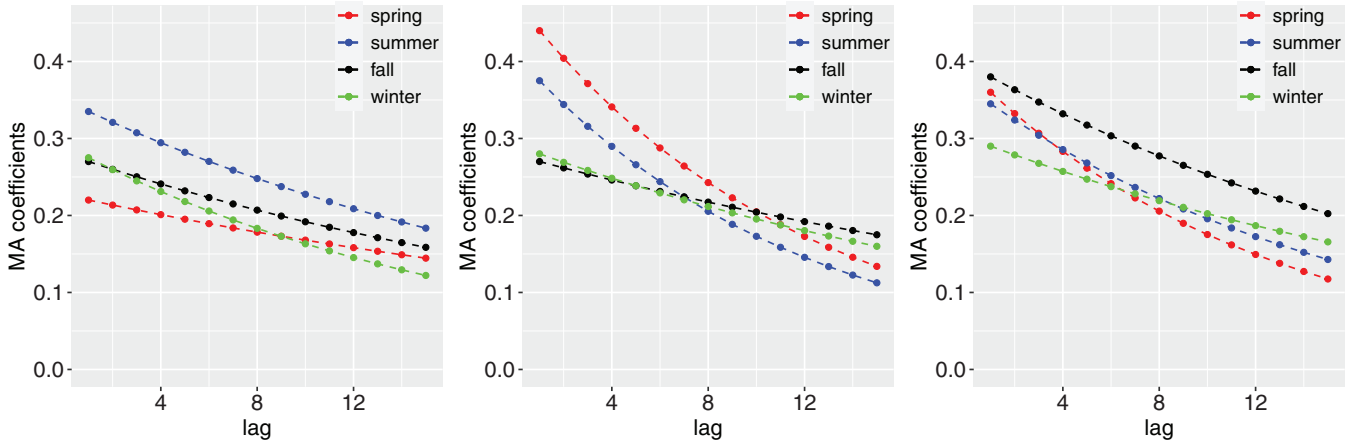

FIGURE 2 A smoothed version of the first 15 MA coefficients averaged across different seasons at different half-hourly time intervals. Columns indicate 30 min time intervals for a representative morning (10:00 a.m.–10:29 a.m.), noon (12:00 p.m.–12:29 p.m.), and afternoon (2:00 p.m.–2:29 p.m.)

(a) the moving average coefficients fall off approximately exponentially and (b) there is diurnal and seasonal variability in the coefficients.

Thus we propose the following model for the real-valued functions $\{\alpha_\ell(\cdot, \cdot)\}_\ell$,

$$\alpha_\ell(t, d) = \sum_{j=1}^{J} a_j(t, d) e^{-b_j(t,d)(t-\ell)^{c_j}}, \tag{7}$$

where J is the number of exponential basis functions and $a_j(t, d)$ and $b_j(t, d) > 0$ are the weights and exponential rates for the jth basis, respectively. The model (7) includes a standard AR(1) in the case that $J = c_1 = 1$ and a_1 and b_1 are constants. In the meantime, we make sure variances of each row are one consistently over time. However, based on exploratory analysis, we parameterize the weight and scale rate parameters as time-day surfaces where $\log(a_j(t, d))$ and $\log(b_j(t, d))$

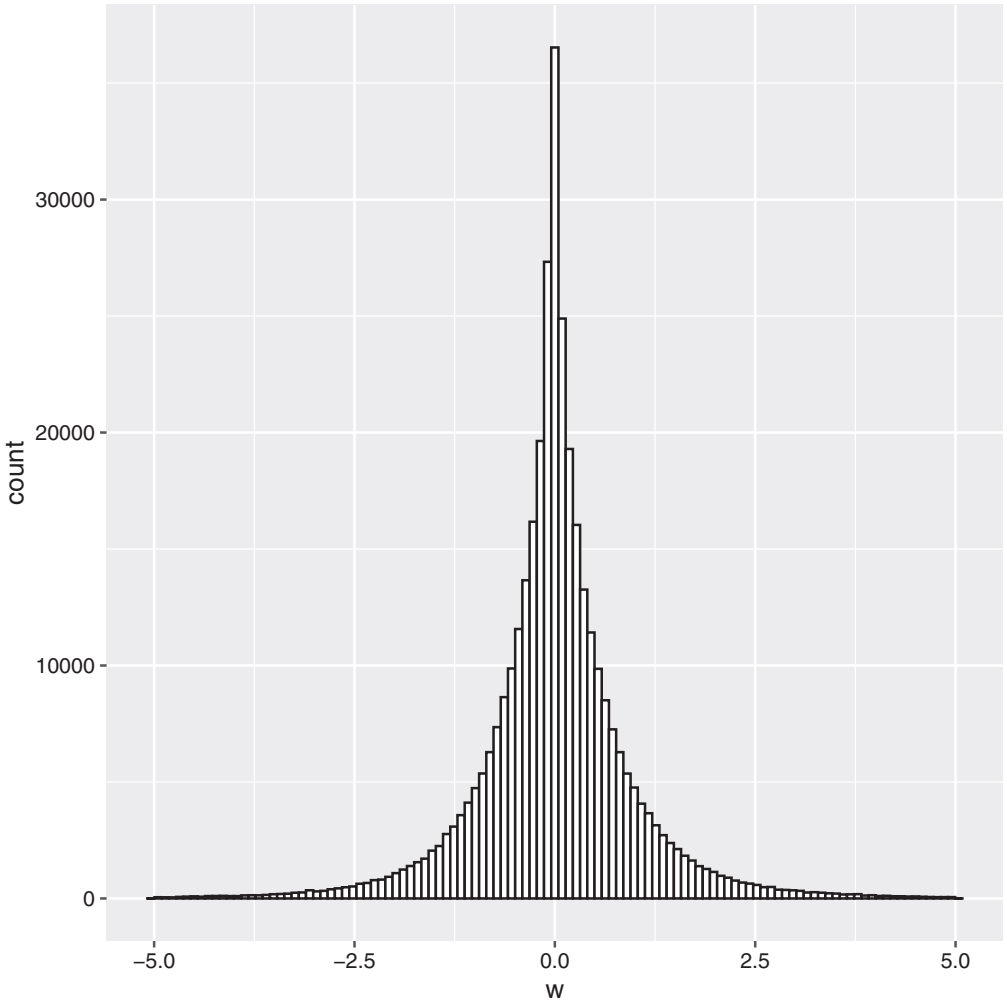

FIGURE 3 Histogram of estimates of $\omega(\cdot,\cdot)$

are linear functions of the form

$$\eta_{kj0} + \eta_{kj1}\sin\left(\frac{2\pi t}{24} + \delta_{kj1}\right) + \eta_{kj2}\sin\left(\frac{2\pi d}{365} + \delta_{kj2}\right) + \eta_{kj3}\sin\left(\frac{2\pi t}{24} + \delta_{kj1}\right)\sin\left(\frac{2\pi d}{365} + \delta_{kj2}\right) \tag{8}$$

for $k = a, b$.

Due to the choice of white noise model in the next section, direct likelihood-based estimation routines are difficult; we thus propose a weighted nonlinear least squares approach. As the final elements of the last row of the Cholesky factor of $\hat{\Sigma}(I_k, d)$ can be used to approximate the correlation structure of $\xi(t, d)$, we minimize the weighted squared difference between the sum of all such Cholesky-based estimates and the model with weights that are $\exp\{\frac{1}{5}\sum_{i=t}^{t+4} L_{ji}\}$ for $t = 1, \ldots, 15$ and otherwise 0 where $j = 30$.

We select J by comparing residual sums of squares between all available Cholesky estimates and the fitted model. The model with $J = 2$ is favored by this criterion with $c_1 = 1$ and $c_2 = 1.1$, whereas models with higher order sums do not substantially improve the fit.

Now the model for the real-valued function $\{\alpha_\ell(\cdot,\cdot)\}_\ell$ with two exponential functions is as follows:

$$\alpha_\ell(t, d) = a_1(t, d)e^{-b_1(t,d)(t-\ell)} + a_2(t, d)e^{-b_2(t,d)(t-\ell)^{1.1}} \tag{9}$$

where $a_1(t, d)$, $a_2(t, d)$, $b_1(t, d)$, and $b_2(t, d)$ are parameterized using model (8), respectively.

TABLE 3 The first four moments of decorrelated residuals

Moments	Mean	Variance	Skewness	Kurtosis
Value	0.0050	0.8988	0.0540	5.7734

TABLE 4 Akaike information criterion (AIC) for mixture models using Gaussian or Laplace components

Number of components	1	2	3	4	5
Gaussian	992,170	893,860	879,795	**877**, **550**	878,338
Laplace	385,260	**375**, **420**	376,520	376,422	376,360

Note: The lowest value is in bold.

3.2.3 | White noise model

The final step is to diagnose and model the driving white noise process of (4). Using the same setup as the prior section we calculate the Cholesky factor $\hat{L}(I_k, d)$ of $\hat{\Sigma}(I_k, d)$ for all $k = 1, \ldots, 30$ and $d = 1, \ldots, 365$. Define the decorrelated empirical residuals for interval I_k and day d as

$$\hat{\boldsymbol{\omega}}(I_k, d) = \hat{L}(I_k, d)^{-1} \hat{\boldsymbol{\xi}}(I_k, d). \tag{10}$$

Figure 3 contains a histogram of decorrelated residuals over all intervals and days. Generally the distribution is approximately symmetric and values are concentrated on the interval $[-5, 5]$ with heavier-than-Gaussian tails. The four moments of $\hat{\boldsymbol{\omega}}(I_k, d)$ are shown in Table 3. Based on the values of skewness and kurtosis, decorrelated residuals have a thin peaked distribution with heavy tails of large values.

We entertain two possible modeling approaches for the decorrelated residuals: mixtures of Gaussian or Laplace distributions. In both cases we assume the mixture component mean parameters are zero, and estimate weights and scale parameters by maximum likelihood using the decorrelated residuals. To assess quality of model fit we compare Akaike information criterion (AIC) values using differing numbers of mixture components. Table 4 compares AIC values for Gaussian or Laplace mixtures using anywhere between one to five components. The Laplace mixture fits are substantially better, and a mixture of two Laplace variables is favored over the four component mixture of Gaussians.

3.3 | Estimation discussion and overview

Here we outline the exploratory analysis approach we took to arrive at the nonstationary non-Gaussian moving average model. The exploratory manner in which we developed the model is intimately tied to estimation. The rough outline is

1. Empirically estimate $\sigma(t, d)$, develop and fit an appropriate parametric model, yielding fitted values $\hat{\sigma}(t, d)$
2. Calculate standardized residuals $\hat{\xi}(t, d) = \varepsilon(t, d)/\hat{\sigma}(t, d)$ and group into interval-localized vectors of residuals $\hat{\boldsymbol{\xi}}(I_k, d)$ for interval I_k
3. Empirically estimate a set of covariance matrices $\hat{\Sigma}(I_k, d)$ based on $\hat{\boldsymbol{\xi}}(I_k, d)$
4. Calculate the Cholesky factor of $\hat{\Sigma}(I_k, d)$, denoted $L(I_k, d)$
5. Identify the last row of $L(I_k, d)$ as estimates of moving average parameters, based on these develop and fit an appropriate parametric model for $\alpha_\ell(t, d)$
6. Calculate $\hat{\boldsymbol{\omega}}(I_k, d) = \hat{L}(I_k, d)^{-1} \boldsymbol{\xi}(I_k, d)$
7. Develop and fit an appropriate white noise model for $\omega(\cdot, \cdot)$ based on $\hat{\boldsymbol{\omega}}(I_k, d)$.

In the prior sections we detailed this exploratory procedure, which coincided with the estimation procedure. Here, Steps 5 and 6 warrant further discussion. The basic idea behind Step 5 is that if $\mathbf{Z} = (Z(1), \ldots, Z(n))^{\mathrm{T}}$ is a mean zero finite variance time series with covariance matrix Σ and L is the Cholesky factor of Σ, we can represent $\mathbf{Z} = L\mathbf{w}$ for an uncorrelated random vector $\mathbf{w}$. In particular, extracting the last row of L, we have $Z(n) = \sum_{i=1}^{n} L_{ni} w_i$, which is suggestive

of a moving average representation. Step 6 is simply the decorrelating nature of this transformation in that $L^{-1}\mathbf{Z}$ is a vector of uncorrelated random variables.

The exploratory approach is intimately linked to estimation of model parameters, which is done in a stepwise fashion, allowing for diagnosing different layers of the model. A more statistically principled approach, for instance, would be based on likelihoods. The moving average model is weighted combinations of Laplace mixtures (which is just a higher-order mixture of Laplace variables). The Laplace distribution is not an infinitely stable distribution, and the likelihood of a mixture of Laplace variables is complicated. Although we make no claims about the statistical efficiency of our approach, we find it attractive due to its link to model development. Moreover, as will be shown in the next section, the estimated model performs well in cross-validation experiments where it is not clear that a likelihood-based approach would provide any superior performance.

The first step is to empirically estimate $\sigma(t, d)$, develop and fit an appropriate parametric model. The fitted values are denoted as $\hat{\sigma}(t, d)$. For example, in the application of downscaling the satellite-derived solar irradiance in Section 3.2, a log linear model is proposed and parameters are estimated by minimizing a sum of difference over $t = 1, \ldots, 1440$ and $d = 1, \ldots, 365$. Then we calculate standardized residuals $\hat{\xi}(t, d) = \varepsilon(t, d)/\hat{\sigma}(t, d)$ and group into interval-localized vectors of residuals $\hat{\boldsymbol{\xi}}(I_k, d)$ for I_k where I_k is a set of minutes in intervals that are consistent with the time resolution of satellite data. A set of covariance matrices can be empirically estimated based on $\hat{\boldsymbol{\xi}}(I_k, d)$ as follows: Define an averaged interval and day specific empirical covariance matrix as

$$\hat{\Sigma}(I_k, d) = \frac{1}{2N} \sum_{i=d-N+1}^{d+N} (\hat{\boldsymbol{\xi}}(I_k, i) - \overline{\boldsymbol{\xi}}(I_k, d))^{\mathrm{T}} (\hat{\boldsymbol{\xi}}(I_k, i) - \overline{\boldsymbol{\xi}}(I_k, d)), \tag{11}$$

where $\overline{\boldsymbol{\xi}}(I_k, d)$ is the empirical mean of $\hat{\boldsymbol{\xi}}(I_k, d)$. The Cholesky factor of $\hat{\Sigma}(I_k, d)$ is calculated and denoted as $L(I_k, d)$. Each row of $L(I_k; d)$ are time-varying moving average coefficients and show similar exponential decay from the diagonal entry. We thus use the last row as empirical estimates of moving average parameters. Based on these, we develop and fit an appropriate parametric model for $\alpha_\ell(t, d)$. The statistical justification for this idea is that if $\mathbf{V} = (V(1), \ldots, V(n))^{\mathrm{T}}$ is a mean zero finite variance time series with covariance matrix Σ and L is the Cholesky factor of Σ, we can represent $\mathbf{V} = L\mathbf{w}$ for an uncorrelated random vector $\mathbf{w}$. In particular, extracting the last row of L, we have $V(n) = \sum_{i=1}^{n} L_{ni} w_i$, which is suggestive of a moving average representation. Pourahmadi (1999) uses modified Cholesky decomposition (MCD) to obtain both unconstrained and interpretable parameters and assumes low-order polynomial dependencies among the innovation variances and autoregressive parameters. The method is developed for the *inverse* Cholesky matrix, whereas our approach directly utilizes the structure of the raw Cholesky factor, which is subsequently linked to a parametric model. The parameters are then estimated using a weighted nonlinear least squares approach. The white noise process is empirically estimated as $\hat{\boldsymbol{\omega}}(I_k, d) = \hat{L}(I_k, d)^{-1} \boldsymbol{\xi}(I_k, d)$, which is simply the decorrelating nature of this transformation in that $L^{-1}\mathbf{V}$ is a vector of uncorrelated random variables. For the last step, we develop and fit an appropriate (and possibly non-Gaussian) white noise model for $\omega(\cdot, \cdot)$ based on $\hat{\boldsymbol{\omega}}(I_k, d)$. In our application, weights and scale parameters for mixture of Laplace distributions are estimated by maximum likelihood.

4 | VALIDATION

The model outlined in Section 3 was fit to pyranometer measurements and NSRDB data at Eugene, OR for the years 2011–2013 and tested for the year 2010. First, we check the efficiency of the proposed Cholesky-based estimation approach based on a small simulation study. Then, we compare the proposed logistic regression model with other clear sky detection methods. In order to validate the adequacy of the downscaled distributions, we hold out in situ data for the same location for the year 2010. We also compare the approach with a simpler time series model, a Gaussian moving average model. Next, we check autocorrelation and coverage properties of the proposed model where we include some results from Boulder and Penn State University (PSU) to show the generality of our approach. Finally we close this section with uncertainty quantification.

4.1 | Efficiency of the Cholesky-based estimation approach

We consider a small simulation study using our proposed moving average estimation procedure to examine if the approach can suitably estimate parameters of a time series model. In particular, essentially matching rows of the Cholesky

TABLE 5 Averaged Cholesky-based estimates with standard errors using k realizations for selected simulations using our proposed model

k	$a_1 = 0.830$	$a_2 = 1.448$	$b_1 = 0.253$	$b_2 = 0.052$	$w_1 = 0.801$	$w_2 = 0.199$	$s_1 = 0.739$	$s_2 = 0.168$
50	0.784 (0.112)	1.398 (0.161)	0.156 (0.112)	0.053 (0.043)	0.768 ($<10^{-5}$)	0.232 ($<10^{-5}$)	0.775 (0.013)	0.180 (0.013)
100	0.825 (0.091)	1.442 (0.121)	0.250 (0.100)	0.052 (0.030)	0.788 ($<10^{-5}$)	0.212 ($<10^{-5}$)	0.740 (0.005)	0.170 (0.006)
300	0.829 (0.062)	1.446 (0.079)	0.252 (0.073)	0.052 (0.019)	0.800 ($<10^{-5}$)	0.200 ($<10^{-5}$)	0.736 (0.002)	0.167 (0.005)

factor is not expected to be a statistically efficient approach, but it is straightforward to implement and bypasses difficult likelihood-based approaches for our non-Gaussian setup. As simulations based on our proposed model are indeed nonstationary and non-Gaussian, we perform a small test on estimating parameters of selected simulations of irradiance under our model for various sample sizes.

The simulations of standardized residuals, $\xi(t, d)$, are defined by parameters a_1, a_2, b_1, b_2 in the moving average model, location parameter μ and scale parameter s in the white noise model. In the first step of the simulation study, we simulate $k \in \{50,100, 300\}$ independent realizations of 30-min standardized residual processes for a representative morning (9:00 a.m.–9:29 a.m.), noon (12:00 p.m.–12:29 p.m.), and afternoon (3:00 p.m.–3:29 p.m.). Following our procedure, we then calculate empirical covariance matrices of each realization and average these to estimate a covariance matrix $\hat{\Sigma}$. The Cholesky factor L of $\hat{\Sigma}$ is then calculated and we identify the last row of L as estimates of moving average weights. The parameters a_1, a_2, b_1, b_2 in the moving average model are estimated by minimizing the weighted squared difference between the sum of all such Cholesky-based estimated and the model with the same weights that used in Section 3.2.2. Then location parameter μ and scale parameter s in the white noise model are estimated using maximum likelihood based on the decorrelated residuals. This process is repeated 500 times to estimate standard errors of the point estimates.

Table 5 presents the representative findings of this small study for noon (12:00 p.m.–12:29 p.m.) on June 20, including average point estimates and standard errors of all parameters. We see that for $k = 300$ (the same sample sizes used in the data example) there is reasonable efficiency in estimating the model parameters, with slight bias. The bias becomes apparently severe as the sample size is reduced, but seems to be practically negligible for the sizes used in our data study.

4.2 | Model comparison

4.2.1 | Comparison with other clear sky detection methods

In the solar energy literature there is sustained interest in developing and validating clear sky identification techniques (Gueymard et al., 2019). Time scale turns out to be an important factor in considering clear sky detection methods, with usual time scales being daily, hourly or minutely. In the following we consider two competing methods from the solar energy literature that focus on sub-daily data.

Two clear sky detection methods with similar input requirements are Reno and Hansen (2016) and Polo et al. (2009). Reno and Hansen (2016) develop a 5-step method using irradiance data and corresponding clear sky irradiance estimates using a moving window of irradiance observations during any given time period to their clear sky modeled counterpart. Threshold values are established empirically, based on one year of data at a single site. The period is identified as clear if the threshold value of each criterion is not reached. Polo et al. (2009) use the daily Linke turbidity factor in identifying clear days. For each day, empirical correlations between observed and clear sky GHI are calculated. The determinant of the correlation coefficient matrix can be used as a measure of the strength of relationship between the two time series: values close to zero indicate stronger correlation between measured and clear sky GHI. Thus, Polo et al. (2009) classify a period as clear if the determinant is lower than a certain threshold value (0.005 is suggested, which we adopt).

We compare classifications from our logistic regression model with Polo et al. (2009) and Reno and Hansen (2016) by training all three on data from Eugene, OR during the years 2011–2013, and validating on 2010. Table 6 contains comparisons of summary statistics from each classifier. Accuracy describes how often the classifier is correct. Specificity shows true negative rate while sensitivity represents true positive rate. The proposed logistic model substantially outperforms both ad hoc models in terms of accuracy and specificity, but tends to underpredict true clear days in favor of non-clear days.

TABLE 6 Validation statistics for clear interval classification rules of (Polo et al., 2009; Reno & Hansen, 2016) and the proposed logistic regression model

Method	Accuracy	Sensitivity	Specificity	Pos. pred value	Neg. pred value
Polo	0.72	0.98	0.70	0.27	0.99
Reno	0.68	***0.99***	0.64	0.24	***0.99***
Logistic	***0.87***	0.88	***0.86***	***0.42***	0.98

Note: Bold values indicate the best performance for a particular statistic.

TABLE 7 The averaged likelihood information using MA(q)

q	1	2	3	4	5	6
Likelihood	1243	1569	3422	4251	5695	5702

4.2.2 | Comparison with a nonstationary Gaussian moving average model

Here we compare our model against a more traditional nonstationary Gaussian moving average time series model to motivate the importance of modeling the heavy-tailedness of the innovations. A qth-order moving average process is defined by

$$\xi(t) = \omega(t) + \theta_1\omega(t-1) + \cdots + \theta_q\omega(t-q), \tag{12}$$

where $\xi(\cdot)$ and $\omega(\cdot)$ are standardized residuals and mean zero unit variance white noises, respectively, while $\{\theta_i\}_{i=1}^{q}$ are moving average coefficients. To fairly mimic our nonstationary approach, we generalize (12) to

$$\xi(t,d) = \omega(t) + \theta_1(I_k,d)\omega(t-1) + \cdots + \theta_q(I_k,d)\omega(t-q), \tag{13}$$

for some fixed q and where t is in 30 min interval I_k.

We estimate and parameterize (13) as follows. Recall that in our new model's estimation scheme, the standardized residuals are first calculated as $\hat{\xi}(t,d) = \epsilon(t,d)/\hat{\sigma}(t,d)$ and grouped into interval-localized vectors of residuals $\hat{\boldsymbol{\xi}}(I_k,d)$ for interval I_k. To estimate the parameters of model (13) we first estimate interval-dependent coefficients $\theta_1(I_k,d),\ldots,\theta_q(I_k,d)$ by maximum likelihood using $\hat{\boldsymbol{\xi}}(I_k,d)$, and assuming samples from different years are independent. We explore different possible model complexities by comparing likelihood values. Table 7 contains the likelihood value averaged over all 30-min time intervals using moving average models estimated by maximum likelihood with different orders $q = 1,\ldots,6$. Clearly, averaged likelihoods increase with q, but a point of diminishing returns is hit around $q = 5$; we thus adopt this value and compare the proposed model against (13) for the case $q = 5$.

With interval-dependent coefficients varying across time of the day and day of year, we fit a time-day surface model for each $\theta_i(I_k,d)$ as a linear function of the form

$$\psi_{i0} + \psi_{i1}\sin\left(\frac{2\pi k}{48} + \delta_{i1}\right) + \psi_{i2}\sin\left(\frac{2\pi d}{365} + \delta_{i2}\right) + \psi_{i3}\sin\left(\frac{2\pi k}{48} + \delta_{i1}\right)\sin\left(\frac{2\pi d}{365} + \delta_{i2}\right), \tag{14}$$

where $\psi_{i0},\psi_{i1},\psi_{i2},\psi_{i3}$ are coefficients for each parameter and δ_{i1},δ_{i2} are phase shifts. These surface parameters are estimated by minimizing a sum of squares difference between the prior interval-dependent moving average coefficients and the model over $k = 1,\ldots,48$ and $d = 1,\ldots,365$.

Using both the proposed model, and the nonstationary Gaussian MA(5) model, we simulate GHI $Y(t,d)$ for the year 2010, and compare the downscaled distribution coverage intervals, a critical quantity in probabilistic resource assessment (Zhang et al., 2015). To examine calibration of the proposed model, empirical coverage against nominal confidence levels based on 1000 downscaled ensembles are calculated. Table 8 contains the corresponding average coverage probabilities at different nominal levels. The proposed model has better coverage at nearly all nominal levels, with slight overdispersion above the 90%. The mean squared error of empirical-to-nominal coverages is 0.04

TABLE 8 Average coverage probabilities at different nominal levels validating at Eugene, OR in 2010 based on 1000 downscaled ensembles

	Nominal levels (%)																		
	5	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	**65**	**70**	**75**	**80**	**85**	**90**	**95**
MA(5)	3.3	6.7	10.0	15.0	20.0	25.0	30.0	33.3	38.3	43.3	50.0	55.0	61.7	66.7	71.7	76.7	83.3	***90.0***	***95.0***
Proposed	***4.8***	***9.6***	***11.7***	***16.7***	***21.7***	***26.7***	***32.6***	***38.7***	***45.3***	***49.3***	***54.3***	***58.5***	***65.0***	***70.0***	***75.0***	***80.0***	86.7	91.5	97.5

Note: Bold values indicate closest to nominal coverage.

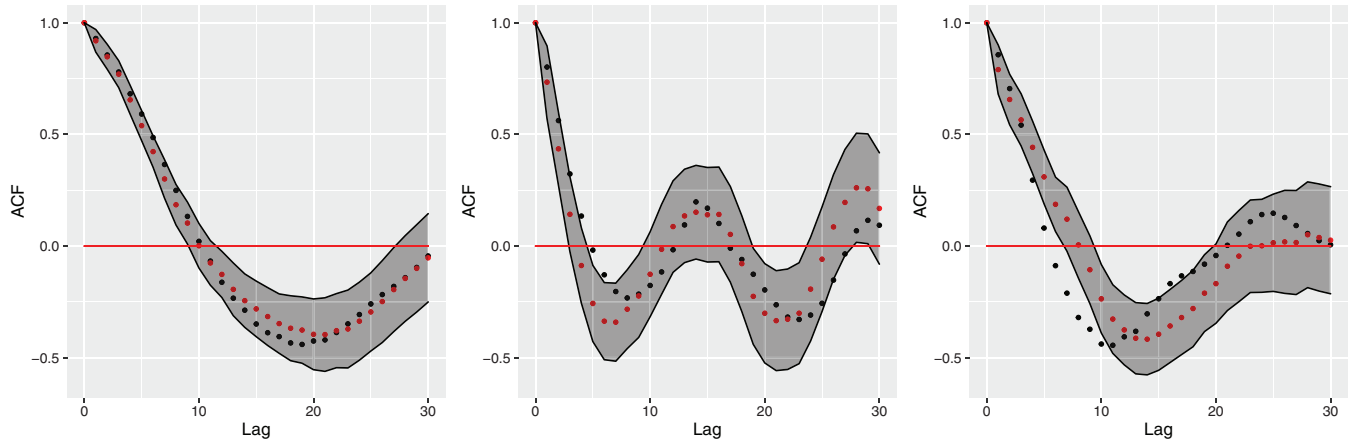

FIGURE 4 Empirical (black dot) and model (red dot) autocorrelations with 95% confidence interval (grey band) on GHI at 1 min time resolution split by time of day on a representative day for summer (June 20) at Eugene, OR. Columns indicate examples of 30 min time intervals for a representative morning (9:00 a.m.–9:30 a.m.), noon (12:00 p.m.–12:29 p.m.), and afternoon half-hour (3:00 p.m.–3:29 p.m.)

for the MA(5) model, and 0.01 for the proposed model, indicating substantially better coverage performance on average.

4.3 | Autocorrelation and coverage properties

We now break out validation of the proposed model by time of day and day of year to show that it is necessary to consider both diurnal and seasonal variations. Here we show representative results based on a summer day. Figure 4 shows representative results of empirical and model autocorrelations on irradiance split by time of day on a representative day in summer (June 20) at Eugene, OR. The representative half-hour time periods for time of day are 9:00 a.m.–9:30 a.m. (morning), 12:00 p.m.–12:29 p.m. (noon), and 3:30 p.m.–3:29 p.m. (afternoon). First, apparent diurnal variability in temporal structure is present. The model not only captures general trends, but also the important correlations at the first few lags, which implies that our proposed method adequately adapts to diurnal variation. The 95% confidence interval shows good coverage of the empirical autocorrelations with only a slight overestimation of high-lag autocorrelations in the afternoon.

Figure 5 contains reliability plots split out by time of day and season for three locations with different climates throughout the US. Using the downscaling framework in Section 3, we demonstrate the value of our method for two additional case studies at locations Boulder, CO and Penn State University, PA. The training period and testing period for both locations are summarized in Table 1. Each reliability plot contains boxplots of empirical coverage against nominal coverage; a perfectly calibrated simulation would follow the identity line. Each coverage percentage is based on 1000 downscaled ensembles, which is repeated 100 times as represented by the boxplot. As can be seen from Figure 5, the median of each boxplot is generally close to the identity line, with slight underdispersion in winter. Such plots indicate that our model is well calibrated, both at different times of day, but also in different seasons. As a final validation, we consider the marginal probability density function (pdf) of GHI in Eugene, OR for the year 2010, shown in Figure 6. This figure contains kernel smoothed estimates of empirical GHI, and GHI based on a single trajectory from the proposed model. Our model seems to

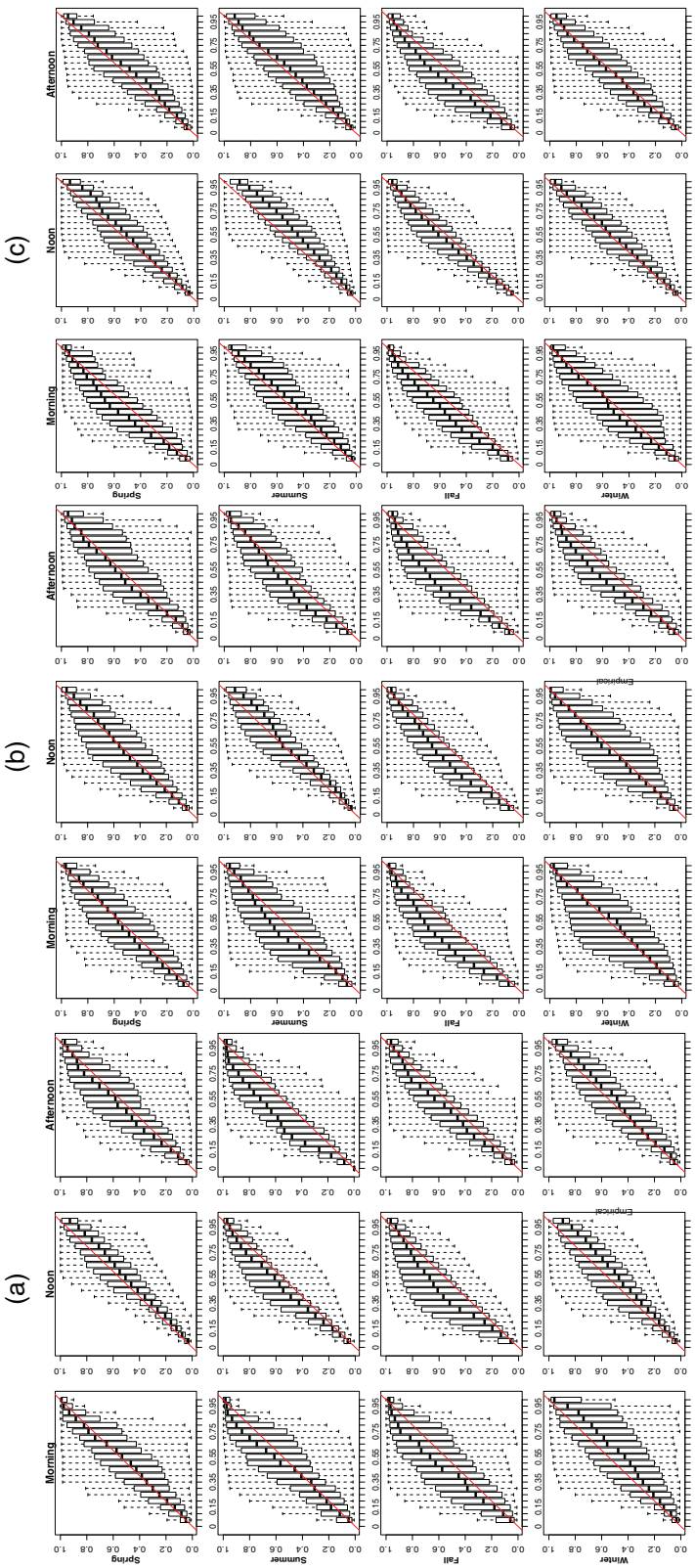

FIGURE 5 Reliability plots for (a) Eugene, (b) Boulder, and (c) PSU based on 1000 independent downscaled realizations at 1-min time resolution. Within each site, columns index morning, noon, and afternoon periods and rows index spring, summer, fall, and winter. Axis units are percentage. In each plot, the *x*-axis denotes empirical coverage while the *y*-axis shows nominal coverage

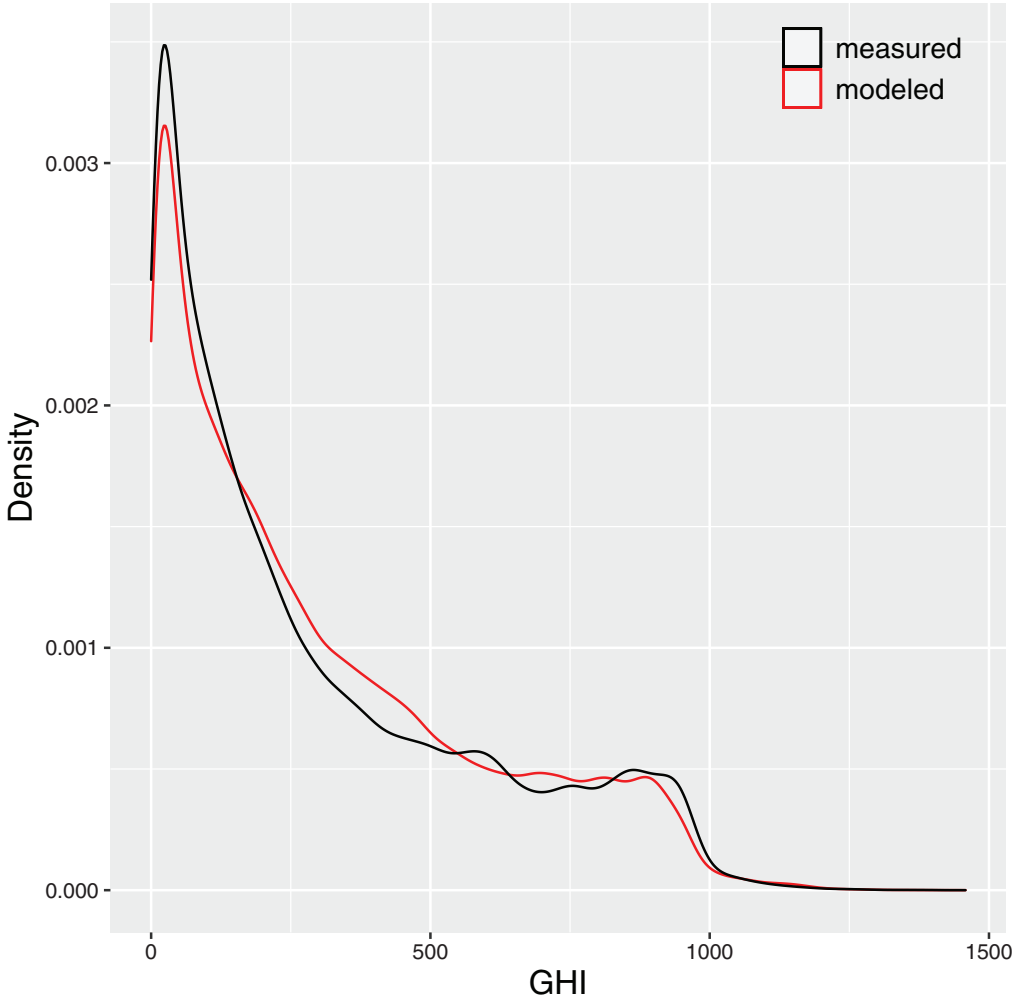

FIGURE 6 Probability density distribution of GHI over Eugene's simulated and in situ data in 2010

accurately reproduce the general behavior of the empirical pdf, but moves some probability from lower values to slightly higher values. The drop around GHI= 900 is due to the constraint of maximal possible GHI during the sunniest months in summer, which is also captured by the model.

4.4 | Uncertainty quantification

Our main interest is in developing a model that can appropriately identify and capture dependencies and nonstationarities of irradiance processes. One component to validation is a detailed quantification of parametric uncertainty; for instance, all parameters in Section 3 are uncertain. We adopt a bootstrap procedure to estimate parametric uncertainty; however, in order to maintain the temporal dependency structure, we cannot sample the data independently. Thus we adopt a moving block bootstrap (MBB) method. The MBB method was proposed by Künsch (1989). Instead of taking random observations from the initial sample like the bootstrap process for independent data, the MBB samples blocks and concatenates them. Therefore, the time series structure of original data is preserved within each particular block of data.

We apply the MBB for estimation of parameters in Section 3. Recall that we have three-year period of training data and here we form the blocks by choosing n consecutive days of data. We first randomly choose a start-point s from $U(1, 1065)$ and then n is then sampled from $U(30, 1095 - s)$. So that the length of the blocks at least covers one month of data. We generate 1000 bootstrap resamples and perform our estimation procedure. Due to the large number of parameters included in our model, we illustrate this assessment of parametric uncertainty with the white noise model parameters based on our empirical training data and 1000 sampled blocks in Table 9, respectively. Here w_i and s_i denotes weights

TABLE 9 Estimation of parameters in white noise model using empirical training data and 1000 sampled blocks, respectively

	w_1	w_2	s_1	s_2
Empirical	0.8013	0.1990	0.7450	0.1668
MBB	0.8138 (0.0602)	0.1880 (0.0165)	0.7580 (0.0530)	0.1528 (0.0225)

Note: Standard errors are in parentheses.

and scales for the ith component of Laplace mixtures where $i = 1, 2$. Standard errors with respect to sampled blocks are in parentheses. The results show that the estimation of the parameters based on 1000 sampled blocks is consistent with what we have based on our original training data.

5 | CONCLUSIONS AND FUTURE WORK

We have presented a postprocessing method that downscales satellite-based estimates of GHI from 30 min snapshots to 1 min time resolution. The model enables the accounting for both seasonal and temporal variability. Such high resolution data is important for renewable energy integration studies and grid operational management. Overall, the proposed downscaling model is shown to appropriately capture variability and distributional characteristics at higher time frequency than is available in typical satellite-based products. The method exhibits good statistical calibration at most nominal levels. The success of our model is attributable to both the nonstationary covariance structure and non-Gaussian noise driving the moving average model.

Future work might consider extending this method to spatially correlated fields of irradiance: spatial consistency is an important factor that affects how implied photovoltaic variability smooths and aggregates across geographical regions. Extension to the space-time setting is a natural next step that we are currently exploring, but unfortunately access to high quality high resolution solar irradiance data is highly limited, with most datasets being proprietary.

ACKNOWLEDGMENTS

This work was authored by Alliance for Sustainable Energy, LLC, the Manager and Operator of the National Renewable Energy Laboratory for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Solar Energy Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes. Kleiber, Hodge and Zhang's research was partially supported by NSF Grant DMS-1923062. Kleiber and Zhang's research was also supported by NSF Grant DMS-1811294.

REFERENCES

Aguiar, R. J., Collares-Pereira, M., & Conde, J. P. (1988). Simple procedure for generating sequences of daily radiation values using a library of Markov transition matrices. *Solar Energy*, *40*(3), 269–279.

Augustine, J. A., DeLuisi, J. J., & Long, C. N. (2000). SURFRAD–a national surface radiation budget network for atmospheric research. *Bulletin of the American Meteorological Society*, *81*(10), 2341–2358.

Balouktsis, A., & Tsalides, P. (1986). Stochastic simulation model of hourly total solar radiation. *Solar Energy*, *37*, 119–126.

Beran, J. (1995). Maximum likelihood estimation of the differencing parameter for invertible short and long memory autoregressive integrated moving average models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(4), 659–672.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Bright, J. M., Smith, C. J., Taylor, P. G., & Crook, R. (2015). Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data. *Solar Energy*, *115*, 229–242.

Chapman, J.-L., Eckley, I. A., & Killick, R. (2020). A nonparametric approach to detecting changes in variance in locally stationary time series. *Environmetrics*, *31*(1), e2576.

Cheng, Q. (1990). Maximum standardized cumulant deconvolution of non-Gaussian linear processes. *The Annals of Statistics*, *18*(4), 1745–1783.

Das, S., Genton, M. G., Alshehri, Y. M., & Stenchikov, G. L. (2021). A cyclostationary model for temporal forecasting and simulation of solar global horizontal irradiance. *Environmetrics*, e2700.

Glasbey, C. A., & Allcroft, D. J. (2008). A spatiotemporal auto-regressive moving average model for solar radiation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *57*(3), 343–355.

Graham, V., & Hollands, K. (1990). A method to generate synthetic hourly solar radiation globally. *Solar Energy*, *44*(6), 333–341.

Gueymard, C. A., Bright, J. M., Lingfors, D., Habte, A., & Sengupta, M. (2019). A posterior clear sky identification methods in solar irradiance time series: Review and preliminary validation using sky imagers. *Renewable and Sustainable Energy Reviews*, *109*, 412–427.

Hocaoglu, F. (2011). Stochastic approach for daily solar radiation modeling. *Solar Energy*, *85*(2), 278–287.

Iversen, E. B., Morales, J. M., Møller, J. K., & Madsen, H. (2014). Probabilistic forecasts of solar irradiance using stochastic differential equations. *Environmetrics*, *25*(3), 152–164.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, *17*(3), 1217–1241.

Lave, M., Kleissl, J., & Arias-Castro, E. (2012). High-frequency irradiance fluctuations and geographic smoothing. *Solar Energy*, *86*(8), 2190–2199.

Lave, M., Kleissl, J., & Stein, J. (2013). A wavelet-based variability model (WVM) for solar PV power plants. *IEEE Transactions on Sustainable Energy*, *4*, 501–509.

Lee, K., Baek, C., & Daniels, M. J. (2017). ARMA Cholesky factor models for the covariance matrix of linear models. *Computational Statistics & Data Analysis*, *115*, 267–280.

Lii, K. S., & Rosenblatt, M. (1982). Deconvolution and estimation of transfer function phase and coefficients for non-gaussian linear processes. *The Annals of Statistics*, *10*(4), 1195–1208.

Lii, K. S., & Rosenblatt, M. (1992). An approximate maximum likelihood estimation for non-Gaussian non-minimum phase moving average processes. *Journal of Multivariate Analysis*, *43*(2), 272–299.

Marcos, J., Marroyo, L., Lorenzo, E., Alvira, D., & Izco, E. (2011). From irradiance to output power fluctuations: The PV plant as a low pass filter. *Progress in Photovoltaics: Research and Applications*, *19*(5), 505–510.

Ngoko, B. O., Sugihara, H., & Funaki, T. (2014). Synthetic generation of high temporal resolution solar data using Markov models. *Solar Energy*, *103*, 160–170.

Perez, M. J., & Fthenakis, V. M. (2015). On the spatial decorrelation of stochastic solar resource variability at long timescales. *Solar Energy*, *117*, 46–58.

Pinheiro, J. C., & Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, *6*(3), 289–296.

Polo, J., Zarzalejo, L. F., Martín, L., Navarro, A. A., & Marchante, R. (2009). Estimation of daily Linke turbidity factor by using global irradiance measurements at solar noon. *Solar Energy*, *83*(8), 1177–1185.

Porcu, E., Rysgaard, J., & Eveloy, V. (2020). Discussion on a high-resolution bilevel skew-t generator for assessing Saudi Arabia's wind energy resources. *Environmetrics*, *31*(7), e2651.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, *86*(3), 677–690.

Reno, M. J., & Hansen, C. W. (2016). Identification of periods of clear sky irradiance in time series of GHI measurements. *Renewable Energy*, *90*, 520–531.

Robinson, P. M. (1995). Gaussian semiparametric estimation of long-range dependence. *The Annals of Statistics*, *23*(5), 1630–1661.

Sengupta, M., Xie, Y., Lopez, A., Habte, A., Maclaurin, G., & Shelby, J. (2018). The National Solar Radiation Data Base (NSRDB). *Renewable and Sustainable Energy Reviews*, *89*, 51–60.

Velasco, C. (1999a). Gaussian semiparametric estimation for non-stationary time series. *Journal of Time Series Analysis*, *20*(1), 87–127.

Velasco, C. (1999b). Non-stationary log-periodogram regression. *Journal of Econometrics*, *91*(2), 325–371.

Velasco, C., & Robinson, P. M. (2000). Whittle pseudo-maximum likelihood estimation for nonstationary time series. *Journal of the American Statistical Association*, *95*(452), 1229–1243.

Vignola, F., & Perez, R. (2004). *Solar resource GIS Data Base for the Pacific northwest using satellite data* (FC26-00NT41011). University of Oregon.

Zhang, J., Draxl, C., Hopson, T., Delle Monache, L., Vanvyve, E., & Hodge, B. M. (2015). Comparison of numerical weather prediction based deterministic and probabilistic wind resource assessment methods. *Applied Energy*, *156*, 528–541.

Zhang, W., Kleiber, W., Florita, A. R., Hodge, B. M., & Mather, B. (2018a). Modeling and simulation of high frequency solar irradiance. *IEEE Journal of Photovoltaics*, *9*(1), 124–131.

Zhang, W., Kleiber, W., Florita, A. R., Hodge, B. M., & Mather, B. (2018b). A stochastic downscaling approach for generating high-frequency solar irradiance scenarios. *Solar Energy*, *176*, 370–379.

Zhang, W., & Leng, C. (2012). A moving average Cholesky factor model in covariance modelling for longitudinal data. *Biometrika*, *99*(1), 141–150.

How to cite this article: Zhang, W., Kleiber, W., Hodge, B. -M., & Mather, B. (2021). A nonstationary and non-Gaussian moving average model for solar irradiance. *Environmetrics*, e2712. https://doi.org/10.1002/env.2712