Received: 8 November 2022 | Revised: 10 February 2023 | Accepted: 7 March 2023

DOI: 10.1002/env.2800

RESEARCH ARTICLE

WILEY

Subordinated Gaussian processes for solar irradiance

Caitlin M. Berry[1] | William Kleiber[1] | Bri-Mathias Hodge[1,2,3]

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

[2]Department of Electrical, Computer and Energy Engineering and Renewable and Sustainable Energy Institute, University of Colorado, Boulder, Colorado, USA

[3]Grid Planning and Analysis Center, National Renewable Energy Laboratory, Golden, Colorado, USA

Correspondence
Caitlin M. Berry, Department of Applied Mathematics, University of Colorado, Boulder, CO, USA.
Email: caitlin.berry@colorado.edu

Funding information
National Science Foundation, Grant/Award Number: DMS-1923062

Abstract

Traditionally the power grid has been a one-way street with power flowing from large transmission-connected generators through the distribution network to consumers. This paradigm is changing with the introduction of distributed renewable energy resources (DERs), and with it, the way the grid is managed. There is currently a dearth of high fidelity solar irradiance datasets available to help grid researchers understand how expansion of DERs could affect future power system operations. Realistic simulations of by-the-second solar irradiances are needed to study how DER variability affects the grid. Irradiance data are highly non-stationary and non-Gaussian, and even modern time series models are challenged by their distributional properties. We develop a subordinated non-Gaussian stochastic model whose simulations realistically capture the distribution and dependence structure in measured irradiance. We illustrate our approach on a fine resolution dataset from Hawaii, where our approach outperforms standard nonlinear time series models.

KEYWORDS

non-Gaussian, nonstationary, renewable energy, simulation, spline

1 | INTRODUCTION

Decarbonizing the grid is a necessary step in aggressive climate action. In 2020 in the United States, electric power generation from all non-renewable sources produced 1.55 billion metric tons of CO_2 emissions, which is equivalent to about 0.85 pounds of CO_2 per kWh (EIA, 2021). In the effort to reduce these emissions as much as possible, renewable energy generation sources, such as solar photovoltaics (PV), will continue to play an increasing role in grid planning and operations moving forward. Traditionally, the power grid has been a one-way system with power flowing from large generators connected to the transmission grid down through the distribution system to consumers. With higher penetrations of distributed energy resources (DERs), like solar panels on individual homes, there is now power being generated at the distribution level with the potential to have reverse flow back into the transmission grid. A key to the success of the integration of solar power generation is to understand solar irradiance as a resource, which correlates very strongly with PV power output. Irradiance is a variable and uncertain resource, which can cause challenges in maintaining efficient and reliable grid operations (Bright et al., 2017; Lew et al., 2013; Nguyen et al., 2016). Having access to high-resolution (in both time and space) solar irradiance data would be beneficial for site planning, grid integration planning, and maintenance. While some sources of solar irradiance data do exist, like satellite data (the National Solar Radiation Database, NSRDB) (Sengupta et al., 2018), weather model data (WRF-Solar) (Haupt et al., 2016; Jimenez et al., 2016) and observational pyranometer data, each of these sources has drawbacks. Satellite and weather model data are typically coarse in time, whereas there is substantial variability of solar irradiance at very high time frequencies, which is only observed in direct pyranometer measurement data. Direct pyranometer data can be rich temporally, but tend to have limited spatial coverage and publicly available datasets are also very limited. Thus, access to synthetically generated solar irradiance time series that mimic observed statistical characteristics of real global horizontal irradiance (GHI) measurement data is critical for PV planning studies. The main goal of this work is to develop a stochastic model for high frequency solar irradiance processes that may then be used to address the dearth of available solar irradiance datasets. It is important to distinguish this goal from solar forecasting – while forecasting has great value for grid operators, and there are many studies that address this need, planning studies require unconditional synthetic datasets (Huang et al., 2012, 2013; Reikard, 2009; Sun et al., 2019; Weiss & Hays, 2004). We address solar resource variability quantification in that the goal of our model is to generate synthetic irradiance simulations which mimic the statistical properties of high-frequency in situ measurements. Said another way, this effort is meant to develop an irradiance stochastic weather generator (Wilks & Wilby, 1999).

In the solar modeling literature there have been myriad approaches proposed toward the goals of variability quantification and generation of simulated irradiance time series. Studies which focus on the former, developing and testing metrics of solar irradiance variability (Blaga & Paulescu, 2018; Castillejo-Cuberos & Escobar, 2020; el Alani et al., 2020), typically do not cover their use in generating or comparing simulated data to real measurements. Our work falls at the intersection of these two ideas where a main goal is synthetic simulation of irradiance data, which has been a subject of interest for decades. (Amato et al., 1986) develop a model for simulating daily solar irradiance using a Fourier analysis combined with a Markov process model, which is appropriate for a much coarser time resolution than our data example.

In more recent work, (Larrañeta et al., 2019) use a bootstrap approach to generate plausible meteorological years (PMYs) of GHI and direct normal irradiance (DNI) 1 min data, which requires a large amount of training samples (about 10–15 years). There have also been attempts at augmentation of a Weather Research and Forecasting (WRF) model designed for solar applications (WRF-Solar) with a physics-based ensemble to simulate solar irradiance at the 1 min and 1 h time scale (Prasad & Kay, 2020). These models at large tend not to perform well at finer temporal resolutions and are more focused on near-term forecasting of GHI. (Ramírez et al., 2021) synthetically generates hourly solar irradiance time series using a multivariate copula autoregressive algorithm, which performs well at capturing temporal dependencies at the hourly level, but has not been tested on high-frequency GHI data. (Zhang et al., 2022) models log clear sky index using a nonstationary moving average process, but also include a Bernoulli variable which serves to identify times of the day which are clear versus times in which clouds, aerosols, or dusts are present. Some distinguishing factors between our work here and (Zhang et al., 2022) is that ours is a continuous time model while theirs is discrete time. The model in (Zhang et al., 2022) is also conditioned on lower resolution data from the NSRDB, which is a downscaling exercise and different from the unconditional modeling we seek to accomplish. One final difference is that, similar to other papers in the literature, this model is built for the 1 min time scale, which has different variability than the 1 s level. Recent work from (Euán et al., 2022) does address high-frequency GHI simulation at the 1 s level using a self-exciting threshold autoregressive (SETAR) model to quantify and replicate the variability of daily irradiance time series for use in short term (1- and 5-step ahead) forecasting. While this has been shown to be quite effective for these short term time horizon forecasts, it has not been tested on the task of full-day one second resolution GHI simulations. In fact, markedly missing from the literature in this subject are models that can unconditionally generate multiple days of realistic GHI data simulations at the 1 s time scale, a method for which we present here. While studies like (Zhang et al., 2022) or (Euán et al., 2022) are successful, we feel that, because the problems they tackle (conditional downscaling and forecasting, respectively) are different enough from ours, direct comparison is not justifiable. Additionally, particularly at fine temporal resolution, GHI data tend to have intermittent high frequency variability and are highly non-Gaussian, which challenges the assumptions of most traditional time series models. As such, even advanced nonlinear time series models fail remarkably at modeling these data, as seen in Section 6.1. We develop a novel approach to handling such variability, relying on a randomly time-changed Gaussian process model at the core. The random time-change process uses a spline representation with random coefficients, and serves to capture clear and cloudy atmospheric conditions. An anamorphosis transformation allows for non-Gaussian behavior in the underlying irradiance process. The proposed estimation approach is stepwise, allowing for the practitioner to check each step of the model fit. We apply the approach to a difficult 1 s timescale pyranometer irradiance dataset from Hawaii, and find that our approach better captures distributional properties, intermittency, and variability than a competing state-of-the-art nonlinear time series model.

2 | IRRADIANCE DATA

We use a dataset from June 1, 2010 through August 31, 2010 of 1 s GHI pyranometer measurements. These measurements were collected by a pyranometer placed at the Daniel Inouye Airport in Oahu, Hawaii (Sengupta & Andreas, 2010).

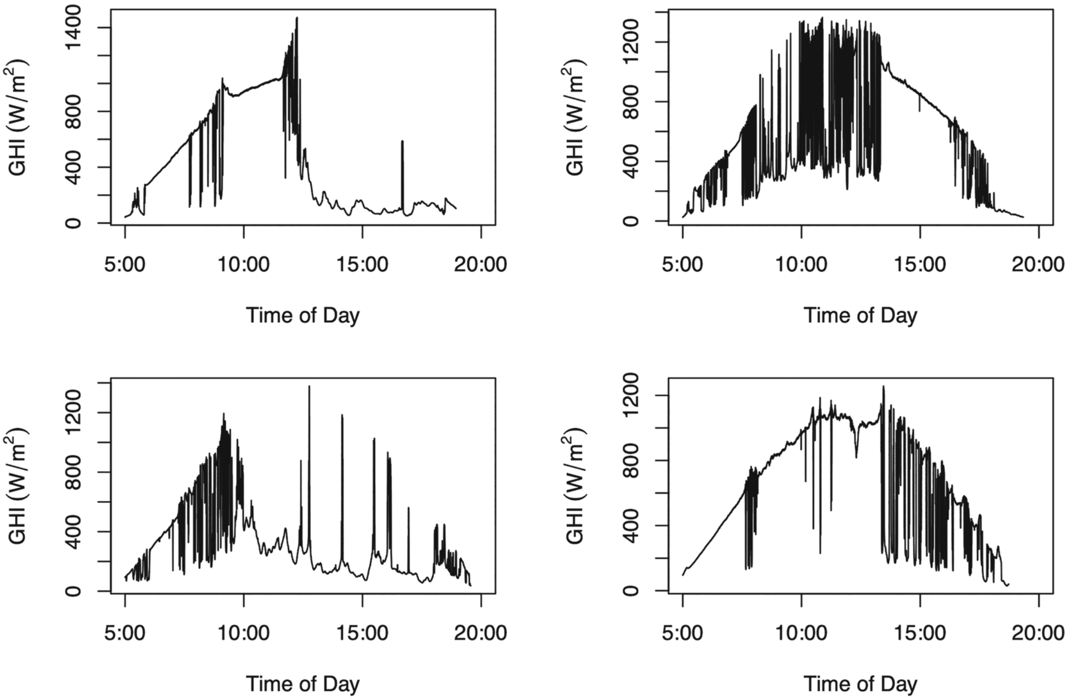

FIGURE 1 GHI measurements from pyranometer every 1 s for a selection of 4 different days during June–August 2010. Note the intermittent high-frequency variability in these data.

As we are interested only in daytime hours, solely the portions of the Oahu GHI data that occurred during daytime were used. Here we define "daytime" as measurements taken between the first and last time points of the day where GHI > 80 W/m^2, which is standard procedure in the literature (Yang et al., 2020). Figure 1 shows a selection of days of 1 s GHI measurement data from this dataset.

We use the National Solar Radiation Database (NSRDB) to obtain clear sky irradiance (CS) estimates for this location (Sengupta et al., 2018). This data is only available at the 30 min timescale and therefore, to get by-the-second CS, we interpolate the NSRDB measurements down to the 1 s level using a cubic spline. This gives us access to the CS time series for the location of the pyranometer, $\{X_C(t)\}$.

3 | MODEL

The goal of this work is to create a time series model, realizations from which behave statistically similarly to by-the-second observed GHI on any given day. Denote by $Y(t)$ the GHI process at time point $t \in \mathbb{R}$, and by $X_C(t)$ the corresponding clear sky irradiance. Clear sky irradiance is the amount of GHI expected under clear conditions with minimal cloud cover, aerosols, and dusts. Note that actual measured GHI can exceed the clear sky value under cloud enhancement events in which sunlight is reflected off of nearby clouds (Smith et al., 2017). We describe the model here in full generality, and in Section 5 detail specific choices for our Oahu dataset.

The clear sky index (CSI) is the ratio of measured GHI to clear sky GHI, which is the typical variable of interest in irradiance modeling (Lave et al., 2013; Zhang et al., 2019; Zhang et al., 2022). Let $Z(t)$ denote the log-CSI process,

$$Z(t) = \log\left(\frac{Y(t)}{X_C(t)}\right). \tag{1}$$

Our proposed model for GHI is

$$Y(t) = X_C(t)\exp\left(F_Z^{-1}\left(\Phi\left(G(S(t))\right)\right)\right) \tag{2}$$

where $F_Z(\cdot)$ is the cumulative distribution function (cdf) of $Z(\cdot)$, parameterized below, $\Phi(\cdot)$ is the cdf of a standard normal random variable, $G(\cdot)$ is a stationary, mean zero Gaussian process with a covariance function parameterized by $\boldsymbol{\theta}_G \in$

$\mathbb{R}^{d_G}$ (which often includes a variance, range and smoothness parameter, in which case $d_G = 3$) and $S(t)$ is a positive, nondecreasing stochastic process. The basic assumption is that actual GHI is a modulated version of the clear sky GHI by a stochastic multiplicative factor. Equivalently, we can write the model in terms of the log-CSI as

$$Z(t) = F_Z^{-1}\left(\Phi\left(G(S(t))\right)\right). \tag{3}$$

The composition of quantile and cdf is sometimes called an anamorphosis function, or a copula, and is often used to model non-Gaussian processes in the statistical climatology literature (Berrocal et al., 2008; Kleiber et al., 2012; Kleiber et al., 2023).

The cdf F_Z can be any parametric distribution, but the CSI process is highly non-Gaussian (see Figure 5 as an example). Thus, we model the corresponding probability density function (pdf), f_Z, as a mixture model,

$$f_Z(z;\boldsymbol{\theta}) = \sum_{k=1}^{K} \lambda_k f_k(z;\boldsymbol{\theta}_k) \tag{4}$$

where $\sum_{k=1}^{K} \lambda_k = 1$, each component pdf f_k has parameters $\boldsymbol{\theta}_k \in \mathbb{R}^{d_k}$ and $\boldsymbol{\theta} = (\lambda_1, \dots, \lambda_K, \boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}}, \dots, \boldsymbol{\theta}_K^{\mathrm{T}})^{\mathrm{T}}$. The specific component distributions $f_1, \dots, f_K$ can be chosen explicitly in a data-driven way based on the application.

We now turn to the nondecreasing stochastic process $S(t)$. To help motivate the use of $S(t)$, it helps to first understand the idea of subordination; such an idea is common in the Lévy process literature (Tankov, 2003). A subordinator is a particular type of Lévy process that can be used as a random time change to build new Lévy processes with certain desirable statistical properties; subordinators are by definition positive and nondecreasing. Heuristically, a subordinator can be thought of as a random process which, when composed with another random process as is done here with this model, has the ability to "speed up" or "slow down" time. Rather than using a formal Lévy process subordinator, we propose an I-spline adaptation, shown in Equation (5). In our proposed model, the "subordinator" $S(t)$ serves as a time change to a Gaussian process, $G(\cdot)$. From exploratory data analysis for our application below, comparing traditional Lévy subordinators to our proposal, we found improvement in capturing clear sky conditions in using the proposed I-spline adaptation, see Appendix A.2 for discussion. Because an I-spline is not strictly a subordinator in the technical Lévy sense, to distinguish our proposal, we refer to it as a *spline-subordinator*.

An I-spline is created using a basis of integrated M-spline basis functions (Ramsay, 1988) and thus can be written as:

$$S(t) = \sum_{\ell=1}^{N} \alpha_\ell \phi_\ell(t) \tag{5}$$

where $\phi_\ell(\cdot)$ are I-spline basis functions, $\alpha_\ell > 0$ are coefficients, and $N = \kappa + d + 1$, where κ is the number of knots in the I-spline, and d is the degree of the I-spline. The α_ℓ coefficients of the I-spline can be thought of as controlling how "quickly" or "slowly" the subordinator moves time with respect to the subordinated Gaussian process (SGP), where each ℓ corresponds to a window of real time in seconds. For example, let $\ell = 1, \dots, N$; in the context of GHI data, we can think of breaking the day up into N windows of time, where each window of time corresponds to an α_ℓ. If a particular α_ℓ is close to zero, then the corresponding I-spline basis function $\phi_\ell(t)$ effectively does not enter $S(t)$, and thus the domain over which its slope is non-negligible does not affect the slope of $S(t)$, whereas if $\alpha_\ell \gg 0$ then $S(t)$ increases dramatically over the effective domain of non-negligible slope of $\phi_\ell(t)$. Figure 2 uses a simple example to illustrate how the α_ℓ coefficients applied to the ϕ_ℓ affect the final $S(t)$. For a selection of $N = 14$ basis functions, $\{\alpha_\ell\}_{\ell=1}^{14} = \{14, 1, 5, 5, 0.1, 15, 20, 22, 0.3, 0.1, 0.01, 15, 0.1, 0.5\}$, the scaled versions of the basis functions are shown in the middle plot and the final form of $S(t)$ for $t = 1, \dots, 1000$ is shown in the bottom panel of Figure 2. One can see from this plot how the $\phi_\ell(\cdot)$ with larger α_ℓ values associated affect the slope of $S(t)$ more dramatically. Thus, larger α_ℓ values tend to correspond to sharper increases in $S(t)$ (a speeding of time), and smaller α_ℓ values correspond to times of flat, almost negligible increases, in $S(t)$ (a slowing of time).

Once the spline-subordinator is generated, as shown in Figure 2, Figure 3 illustrates the process of subordinating a stationary GP, which, when passed through the model (2), can replicate the kinds of behavior seen in 1 s GHI measurements.

For the proposed model, we assume α_ℓ are strictly positive, uncorrelated left censored normal random variables, $\alpha_\ell \sim N_{\text{censored}}(\mu_\ell, \sigma^2; \epsilon)$ with mean μ_ℓ, σ^2 is a constant variance for all ℓ, and ϵ refers to the censoring value, below which

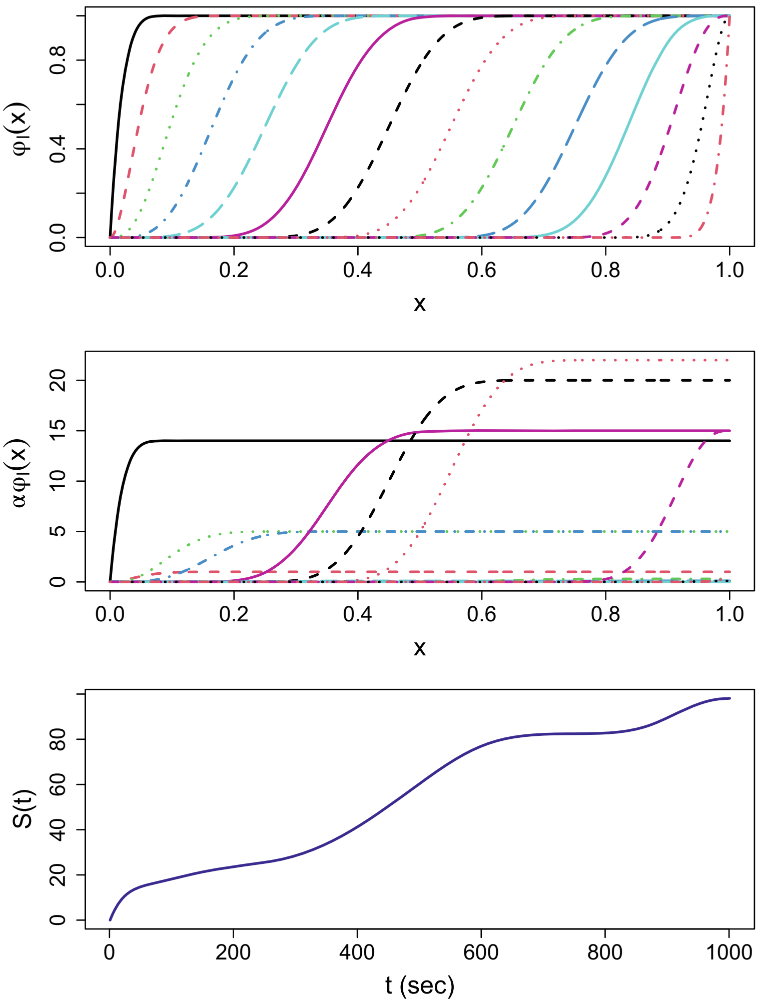

FIGURE 2 I-spline basis functions (top), scaled I-spline basis functions with various α (middle) and final spline form $S(t)$ (bottom).

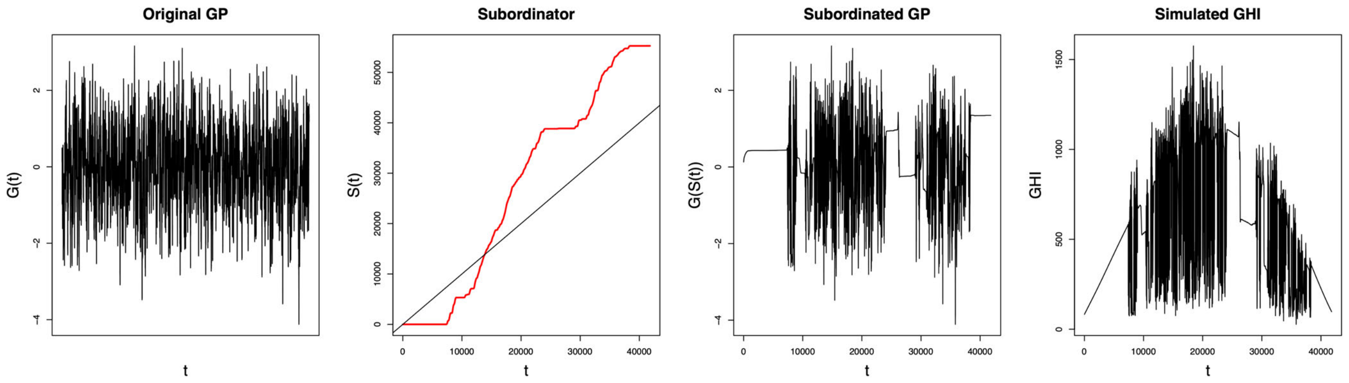

FIGURE 3 Example realization of stationary mean-zero GP, $G(t)$, (left) $S(t)$ (red) with linear time (black) (left middle), subordinated GP (right middle), and simulation of GHI using the SGP (right).

(uncensored values of) α_ℓ are assigned the value ϵ. We choose to model μ_ℓ with a B-spline

$$\mu_\ell = \sum_{k=1}^{M} \beta_k B_k(\ell) \tag{6}$$

where $\{B_k(\cdot)\}$ is the set of B-spline basis functions with coefficients $\{\beta_k\}$, and $\sigma^2 > 0$ a constant variance parameter for all ℓ (de Boor, 1978). The choice of M and knots are detailed in Section 5.

The flexible family of mean models for μ_ℓ allow periods during which $N(\mu_\ell, \sigma^2)$ is below the cutoff ϵ, but also vary above the cutoff in a continuous way. This is useful because there may be stretches of time, which correspond to clear points of the day, where we need time to "slow" with respect to the SGP. These correspond to times when $\alpha_\ell \approx \epsilon$. By having μ_ℓ modeled as a B-spline, we can achieve this behavior by μ_ℓ dipping below ϵ (see the red curve in Figure 6), thereby creating stretches of "time" where the $\alpha_\ell = \epsilon$. Then to generate a set of $\{\alpha_\ell\}_{\ell=1}^{N}$, unconditionally simulate N random variables distributed as $N(\mu_\ell, \sigma^2)$ and assign values below ϵ to ϵ. See Figure 6 for an example of the creation of this process.

4 | ESTIMATION AND SIMULATION

We begin this section with a description of our proposed estimation approach using a single day's worth of GHI measurement data. Expansion to multiple days of data is discussed in Section 4.4. Suppose we have a set of GHI observations, $\{Y(t)\}$, at $t = 1, \ldots, L$ where L represents the number of time points of non-negligible irradiance (i.e., irradiances during daylight hours and above the threshold discussed in Section 2). Time is indexed in seconds, and L varies depending on the day of year, but is on the order of tens of thousands.

Under the proposed model, techniques wherein we maximize the likelihood to find suitable parameters are not easily accessible because the joint distribution of $(S(1) \ldots S(L))^{\mathrm{T}}$, the values of the subordinator at each time point, does not have a closed form in general and therefore likelihood-based estimation approaches are not directly available. Even if closed-form likelihoods were available, we would be working in a high-dimensional parameter space, which would be a computational burden that the approach we propose avoids. Due to these limitations, we propose a multi-step procedure, which is not uncommon in the solar irradiance forecasting and modeling literature (Euán et al., 2022; Zhang et al., 2022). See Figure 4 for a summary overview of the steps of the estimation procedure.

4.1 | Marginal distribution

Given all available $Z(\cdot)$ data, we estimate the parameters of the mixture model $f_Z(z; \boldsymbol{\theta})$ in (4) by maximum likelihood using the expectation-maximization (EM) algorithm, assuming mutual independence between $Z(t_1)$ and $Z(t_2)$ for distinct t_1 and t_2 (Dempster et al., 1977). Although pairs of log-CSI are dependent, the joint distribution of $(Z(t_1), \ldots Z(t_N))^{\mathrm{T}}$ is complicated and this simplifying assumption works well in exploratory experiments, partly due to the volume of data we have at the 1 s level.

The marginal density f_Z is the density of log-CSI data for a stretch of days on the order of a month or a season. The interpretation of this with regard to our multi-day simulations is that we are, to a degree, assuming that each individual day has the same density as the aggregate of data. In reality, there will be some variability in the distribution f_Z on a daily basis, and, in theory, one could instead fit a new mixture model for each day of interest to account for this. However, for more than a few days of irradiance data, this would be computationally challenging. Since the goal of this work is to capture and quantify distributional qualities of GHI over, say, a season, this choice of common density f_Z performs well.

Estimation of $f_Z(z; \boldsymbol{\theta})$ here is done as an iterative process. First, select $D \in \mathbb{N}$ options for $\hat{K}$ and the components f_k for $k = 1, \ldots, \hat{K}$. This generates D number of possible models for f_Z. Second, estimate the MLEs $(\hat{\lambda}_k, \hat{\theta}_k)$ based on those choices using EM for each of the possible D choices of model. Given the various potentially reasonable choices for $\hat{K}$ and f_k, we fit a variety of models and select the model that balances goodness-of-fit of the quantile-quantile (Q-Q) plot and model simplicity. We use the Q-Q plot as a visual assessment of model fit, and choose a model such that the Q-Q plot is "close" to the identity line (this check can be done visually or using the nonparametric two-sample Kolmogorov-Smirnov (K-S) test (Pratt & Gibbons, 1981)) while using the smallest $\hat{K}$ possible. One could envision a quantitative information criterion that balances goodness-of-fit with model complexity, but for this application we find the visual assessment works well.

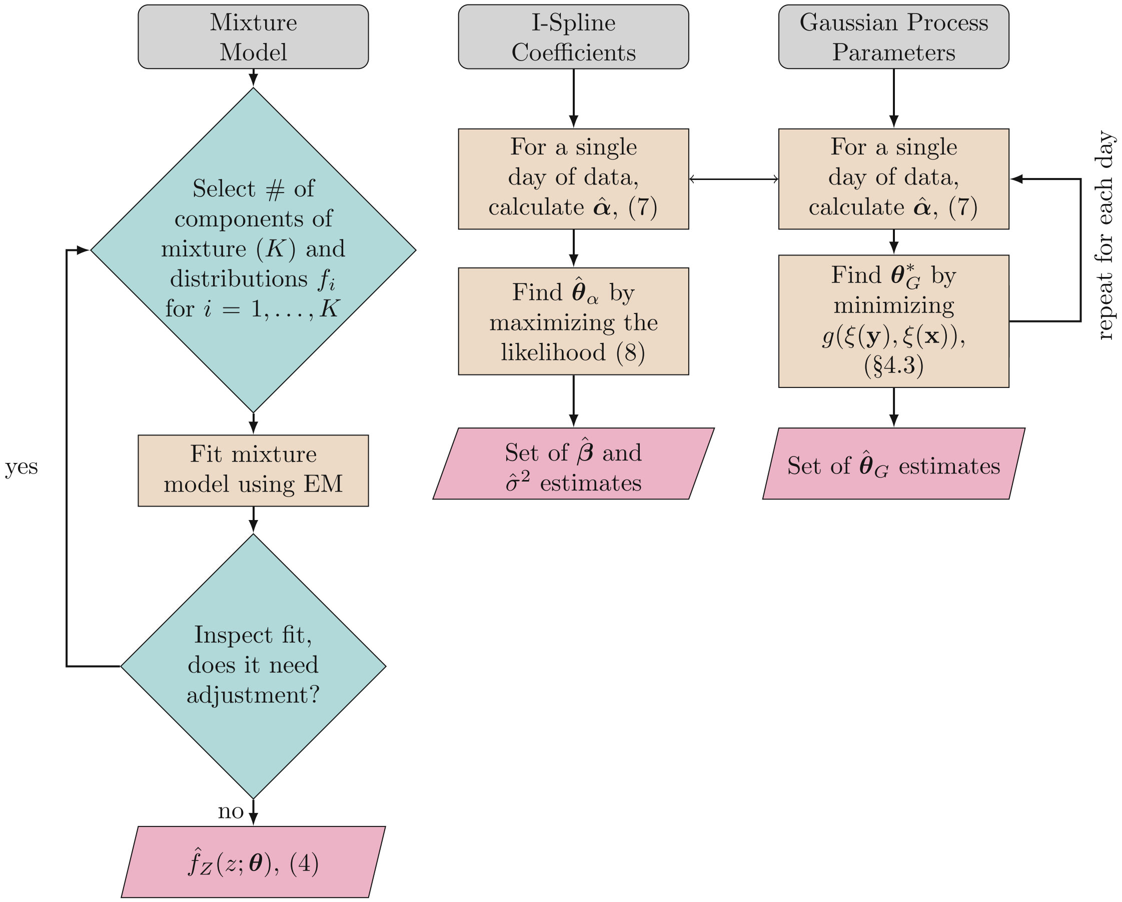
Mixture Model
Select # of components of mixture (K) and distributions f_i for $i = 1, \ldots, K$
Fit mixture model using EM
Inspect fit, does it need adjustment?
yes
no
$\hat{f}_Z(z; \boldsymbol{\theta})$, (4)
I-Spline Coefficients
For a single day of data, calculate $\hat{\boldsymbol{\alpha}}$, (7)
Find $\hat{\boldsymbol{\theta}}_\alpha$ by maximizing the likelihood (8)
Set of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ estimates
Gaussian Process Parameters
For a single day of data, calculate $\hat{\boldsymbol{\alpha}}$, (7)
Find $\boldsymbol{\theta}_G^*$ by minimizing $g(\xi(\mathbf{y}), \xi(\mathbf{x}))$, (§4.3)
Set of $\hat{\boldsymbol{\theta}}_G$ estimates
repeat for each day

FIGURE 4 Flowchart of estimation algorithm; diamonds represent decision points, rectangles process points, and parallelograms are outputs.

4.2 | Spline-subordinator

We now turn to the parameters governing the spline-subordinator, $S(t)$, defined in (5). Recall the connection of the α_ℓ coefficients to the behavior of the spline-subordinator: higher α_ℓ values tend to correspond to windows of time in the day where we see higher frequency variability (i.e., a "speeding" of time according to the spline-subordinator) and vice versa for $\alpha_\ell \approx 0$. With this motivation, we propose an approach to estimate the α_ℓ which utilizes a moving-window variance measure of the data. The scheme is implemented in two parts: first, empirically calculate a local variance measure of the data, second, fit a parametric model to that measure.

Let one day of log-CSI data be denoted $\mathbf{Z} = (Z(1), \ldots, Z(L))^{\mathrm{T}}$. What we define as the empirical local variance measure (ELVM) is calculated by dividing the L time points available in the day into N overlapping windows of time and taking empirical variances of the data within those windows. This ELVM is used as an intermediate estimate for the $\hat{\alpha}_\ell$. The goal is to match a potential $\hat{\alpha}_\ell$ to an estimated variance of the data within each window for $\ell = 1, \ldots, N$. In theory, the higher the empirical variance, the higher frequency variability we might expect to see in the measurements and therefore the higher we would want the active $\hat{\alpha}_\ell$ in that window of time to be – signifying a speeding of time. Alternatively, a low empirical variance in the window would produce a small $\hat{\alpha}_\ell$, which would have the effect of slowing time.

Denote by $\mathbf{Z}_{\ell,\delta} = (Z(\lceil k_\ell \cdot L \rceil - \delta), \ldots, Z(\lceil k_\ell \cdot L \rceil + \delta))^{\mathrm{T}}$, where $k_\ell \in (0, 1)$ is the ℓ th knot in the I-spline, $\delta = \lceil k_1 * L \rceil - 10$ is the half-width of the window over which we are taking the variance, $\lceil \cdot \rceil$ is the ceiling function, and let $\overline{\mathbf{Z}_{\ell,\delta}}$ be the mean of this vector. Then define the ELVM for $\hat{\alpha}_\ell$, as

$$\hat{\alpha}_\ell = \frac{(\mathbf{Z}_{\ell,\delta} - \overline{\mathbf{Z}_{\ell,\delta}})^{\mathrm{T}}(\mathbf{Z}_{\ell,\delta} - \overline{\mathbf{Z}_{\ell,\delta}})}{2\delta + 1}, \tag{7}$$

which is simply the empirical variance of $\mathbf{Z}_{\ell,\delta}$.

Once $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \dots, \hat{\alpha}_N)^{\mathrm{T}}$ is calculated, we can use this as data to estimate parameters for our proposed model for the α_ℓ using a likelihood-based approach. As mentioned in the model section, assume $\hat{\boldsymbol{\alpha}}$ represent realizations of independent random variables with a censored normal distribution (see Appendix for functional form). Assuming independence, the likelihood function for $\boldsymbol{\theta}_\alpha$ is

$$L(\boldsymbol{\theta}_\alpha | \hat{\boldsymbol{\alpha}}) = \prod_{\ell=1}^{N} f_\alpha(\hat{\alpha}_\ell; \theta_\ell) \tag{8}$$

where $\boldsymbol{\theta}_\alpha = (\boldsymbol{\beta}, \sigma^2)^{\mathrm{T}}$ is a size $(M+1) \times 1$ vector of the parameters for f_α, the pdf for a censored normal. We calculate the maximum likelihood estimator for $\boldsymbol{\theta}_\alpha$ by maximizing (8).

4.3 | Gaussian process parameters

Next we discuss estimation of $\boldsymbol{\theta}_G$, the parameters for the Gaussian process $G(\cdot)$. The parameters $\boldsymbol{\theta}_G$ define the covariance structure for the Gaussian process; exact specification of which covariance function we use is described in the model specifics section next, but typically $\boldsymbol{\theta}_G$ includes a variance, range and smoothness parameter. Let $\mathbf{y} = (y(1), y(2), \dots, y(L))^{\mathrm{T}}$ be a vector of historical by-the-second GHI data. A basic problem is that the joint distribution of $\mathbf{y}$ is intractable due to the complicated nature of the joint distribution of the underlying subordinator $S(t)$. Thus, we propose an approach which minimizes the L^2 distance of spectral estimators based on an adapted approximate Bayesian computation (ABC) algorithm (Beaumont et al., 2009).

In an ABC algorithm, first, a parameter value $\boldsymbol{\theta}_G^*$ is chosen. Then a realization of the data is generated according to the statistical model under $\boldsymbol{\theta}_G^*$, $\mathbf{Y} = (Y(1), \dots, Y(L))^{\mathrm{T}}$, at which point statistics based on the data and simulation are calculated, $\xi(\mathbf{y})$ and $\xi(\mathbf{Y})$, and a distance between those values is considered, $g(\xi(\mathbf{y}), \xi(\mathbf{Y}))$. If the distance is below a given threshold, the value of $\boldsymbol{\theta}_G^*$ is recorded, otherwise it is discarded; the algorithm then begins anew with a newly chosen value of $\boldsymbol{\theta}_G^*$. We propose an approach similar to an ABC algorithm, but where uncertainty in $\boldsymbol{\theta}_G$ is quantified using daily replicates of data, rather than posterior estimate values of $\boldsymbol{\theta}_G$.

We now define $\xi(\mathbf{Y}), \xi(\mathbf{y}), g(\cdot,\cdot)$ for this particular model. Define first $P_y(\cdot)$ to be the periodogram of the data $\mathbf{y}$:

$$P_y(\omega) = \frac{1}{L}\left|\sum_{k=0}^{L-1} y_k e^{-i\omega k}\right|^2 \tag{9}$$

for $\omega \in [0, \pi]$. We then smooth $P_y(\cdot)$ using a Daniell kernel with $C \in \mathbb{N}$ spans set to $\{sp_1, sp_2, \dots, sp_C\}$ to get the statistic $\overline{P}_y(\cdot)$. The choice to smooth comes from the desire to limit the high variance the periodogram tends to exhibit over ω varying from $[0, \pi]$. Smoothing eases this variability to better highlight the overall trend, and is common procedure in the time series literature (Brockwell & Davis, 1991). The number C of spans in the Daniell kernel can be chosen to yield the best results for specific data. Then we let $\xi(\mathbf{y}) = \log(\overline{P}_y)$. $\xi(\mathbf{Y})$ is created in an analagous manner using the realization from the model, $\mathbf{Y}$, as a replacement for the real measurement data $\mathbf{y}$. The distance metric $g(\cdot,\cdot)$ is the L^2 distance between $\xi(\mathbf{y})$ and $\xi(\mathbf{Y})$, given by

$$\sum_{j=0}^{J}\left(\log\left(\overline{P}_y(\omega_j)\right) - \log\left(\overline{P}_Y(\omega_j)\right)\right)^2 \tag{10}$$

where $\omega_j = \frac{2\pi j}{L}$ for $j = 0, \dots, \lfloor L/2 \rfloor$.

For each day of data, $\mathbf{y}$, and a given $\boldsymbol{\theta}_G$, we can generate $\mathbf{Y}$ using our proposed model. Then, instead of selecting various $\boldsymbol{\theta}_G^*$ and recording them if they meet a certain criteria, we first search for a $\boldsymbol{\theta}_G^*$ for this specific $\mathbf{y}$ by way of

$$\boldsymbol{\theta}_G^* = \underset{\boldsymbol{\theta}_G}{\operatorname{argmin}} \; \{g(\xi(\mathbf{y}), \xi(\mathbf{Y}); \boldsymbol{\theta}_G)\} \tag{11}$$

and save the $\hat{\boldsymbol{\theta}}_G = \boldsymbol{\theta}_G^*$ which minimizes this distance. This process is repeated for each $\mathbf{y}$, thereby allowing for estimation of $\boldsymbol{\theta}_G$ through repeated estimation using daily $\mathbf{y}$, which are available through the given data set.

4.4 | Multi-day parameter estimates

The estimation steps outlined in Sections 4.2 and 4.3 are designed for use with a single day of historical GHI data at a time. However, in practice, and in our example below, we have multiple days of data. We propose repeating the aforementioned estimation processes using multiple days of historical by-the-second GHI data, resulting in sets of estimated parameters $\{\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\boldsymbol{\theta}}_G\}$. As described in Section 4.5 below, these daily parameters can then be resampled to capture daily variability in the estimated models, as well as uncertainty in the individual parameter estimates.

4.5 | Simulation

Once the proposed model has been trained on a set of data, one can simulate realizations of one second GHI data one day at a time. Detailed here is the methodology used to generate these simulated data. Note that a specified $\boldsymbol{\Theta} := (\hat{\boldsymbol{\beta}},\ \hat{\sigma}^2,\ \hat{\boldsymbol{\theta}}_G)^T$, is all that is required to simulate from the model (2).

Let us first assume we have access to a set of distributions of estimated parameters for the component parameters of $\boldsymbol{\Theta}$. We may generate $\mathbf{Y}$, a realization of one day of GHI data, by first independently resampling M times from the distribution of $\hat{\beta}$ estimates and once from the distribution of $\hat{\sigma}^2$ estimates. Then $\hat{\mu}_\ell$ can be determined using Equation (6). Once $\hat{\mu}_\ell$ is calculated, we can generate $\hat{\alpha}_\ell$ by simulating N independent $N(\mu_\ell, \sigma^2)$ random variables and left-censoring them at ϵ. These $\hat{\alpha}_\ell$ can then be substituted into (5) to generate $S(t)$. Next, independently resample a $\hat{\nu}$ parameter and jointly resample a set of $\{\hat{\rho}, \hat{\nu}\}$ from the corresponding distribution of estimated parameters, which make up $\hat{\boldsymbol{\theta}}_G$. Once $\hat{\boldsymbol{\theta}}_G$ and $S(t)$ are known, using $X_C(t)$ for $t = 1, \ldots L$ and the estimated cdf model $\hat{F}_Z$, we can generate $\mathbf{Y}$ using the model (2).

5 | MODEL CHOICES AND ESTIMATES FOR HAWAII DATA

We now apply this model and estimation technique to irradiance data from Hawaii. We describe our specific model choices made with regard to the cdf F_Z, the spline-subordinator, and the Gaussian process. With the exception of the density function for $Z(\cdot)$, estimation schemes for all parameters in the model are performed on each single day of GHI measurement data separately.

Using these model choices for the Hawaii data, we train the model on 1 s GHI measurements from June and July. There are 61 days of data in these months so we have a set of 61 best estimates of the parameters $\boldsymbol{\beta}, \sigma^2, \rho, \nu, \nu$, which are also shown and remarked upon in the following sections.

5.1 | Model choices and estimates for F_Z

Figure 5 shows a histogram of $\{Z(t)\}$ for t in June, July, and August 2010. For this application, we assume $Z(t)$ has the same density, f_Z, for every day of June, July, and August. In general, f_Z varies day-to-day, but the months of June, July and August define a widely-accepted climatological period of summer, during which we assume a stationary climate. The histogram in Figure 5 is an empirical estimate of f_Z that we model using the mixture model of Equation (4). For the pdf $\hat{f}_Z$, we estimate $\hat{\boldsymbol{\theta}}$ which includes $\hat{K}$, the number of distributions in the mixture model, $\hat{\lambda}_k$, the proportions of the mixtures, and all associated $\hat{\theta}_k$, the sets of parameters for each of the member distributions of the mixture model, f_k.

Based on visual inspection of the somewhat bimodal histogram of log-CSI data in Figure 5, we determined that the general shapes of the component distributions look like they could be well fit by a combination of Gaussian and Gamma distributions. Therefore we originally chose $\hat{K} = 3$, where the components were two Gaussian and one Gamma. This fit the distribution of data in terms of having a Q-Q plot which followed the identity line relatively well; however, the tail (particularly at the top 0.5% of data) of the fit distribution $\hat{f}_Z$ for $\hat{K} = 3$ were slightly too light, thus concentrating more mass at high values of GHI than actual GHI measurement data can support. As a result, we decided to add a fourth component to this mixture model, which is a Pareto. As a result of this iterative process, we select the model for $\hat{f}_Z$ with $\hat{K} = 4, f_1(z; \theta_1)$ a Gaussian pdf with parameters $(\mu_1, \sigma_1^2), f_2(z; \theta_2)$ a Gaussian pdf with parameters $(\mu_2, \sigma_2^2), f_3(z; \theta_3)$ a Gamma pdf with parameters (α, β, c), and $f_4(z; \theta_4)$ a Pareto pdf with parameters (η, θ). The results of the maximum-likelihood estimates from the EM algorithm are shown in Table 1.

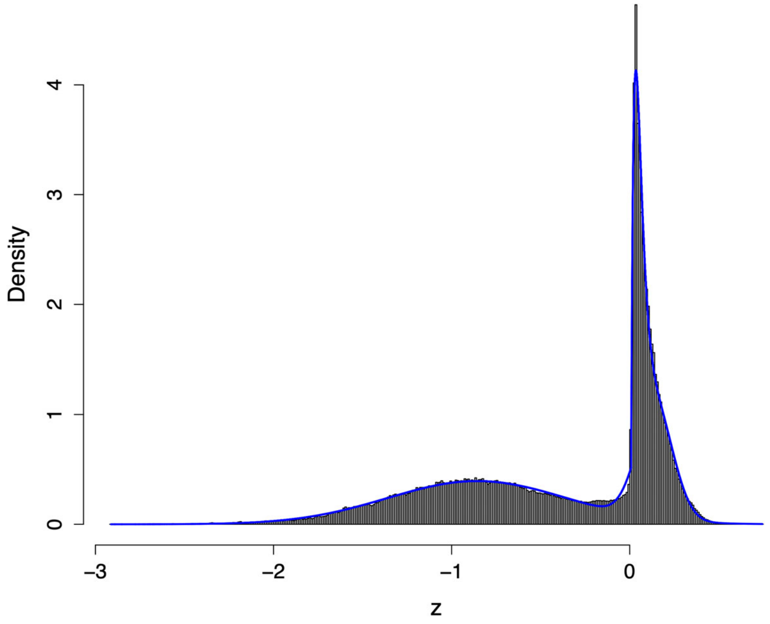

FIGURE 5 The estimated density $\hat{f}_Z$ (blue) along with the histogram of $Z(t)$ data for June, July, and August.

TABLE 1 The estimated parameters for the f_Z distribution function fit for Equation (4).

k	Parameters	Estimates
1	$(\lambda_1, \mu_1, \sigma_1)$	$(0.494, -0.857, 0.5)$
2	$(\lambda_2, \mu_2, \sigma_2)$	$(0.255, 0.142, 0.1)$
3	$(\lambda_3, \alpha, \beta, c)$	$(0.245, 1.978, 0.026, 0.007)$
4	$(\lambda_4, \eta, \theta)$	$(0.005, 9.357, 0.387)$

5.2 | Model choices for the spline-subordinator

With regard to the estimation of the spline-subordinator, recall that we need to choose the number of knots κ, the degree d, and the number of coefficients M for the B-spline, μ_ℓ.

We selected the number of equally-spaced knots to be $\kappa = 500$ and the degree to be $d = 4$. Since the amount of data in each day falls around 50,000 s, with this choice of κ, each α_ℓ in Equation (6) will take into account a little more than 1.5 min of 1 s GHI data. Aside from convenience, the justification for equally-spaced knots is that the data are evenly spaced in time so it isn't clear that more complicated knot placements (e.g., based on quantiles of available features of the data) would be an improvement. And even if they were an improvement, the level of complexity this adds to the problem at this stage is beyond the scope of this paper. We chose $d = 4$ to achieve a certain level of smoothness for the spline-subordinator. Finally, we choose $M = 25$ to allow for the most possibility for flexibility in the B-spline to fit the ELVM vector $\hat{\boldsymbol{\alpha}}$ while acknowledging we only have 505 data points with which to estimate $\boldsymbol{\beta}$, the coefficients of the B-spline; this choice balances between model flexibility and model parsimony.

An example realization of $\hat{\alpha}_\ell$ is shown in Figure 6. The mean trend μ_ℓ (B-spline) is shown in red and the final process in blue. The blue line in this figure illustrates how the anamorphosis censors the $N(\mu_\ell, \sigma^2)$ process (black) at $\epsilon = 0.001$, thereby creating some stretches where $\hat{\alpha}_\ell \approx 0$. This also further illustrates the motivation for the use of the B-spline mean trend in pulling the original uncorrelated $N(\mu_\ell, \sigma^2)$ process below ϵ, where it will be censored by the anamorphosis, resulting in stretches of time with near zero $\hat{\alpha}_\ell$ s, which correspond to clear times of day.

For the 61 days of June and July irradiance data, Figure 7a,b show the marginal densities of the set of β_k (treated as independent here, see Appendix A.3 for discussion) and σ^2. The β_k and σ^2 densities seem to be uni-modal in nature, centered around $\beta \approx 0$ and $\sigma^2 \approx 0.15$.

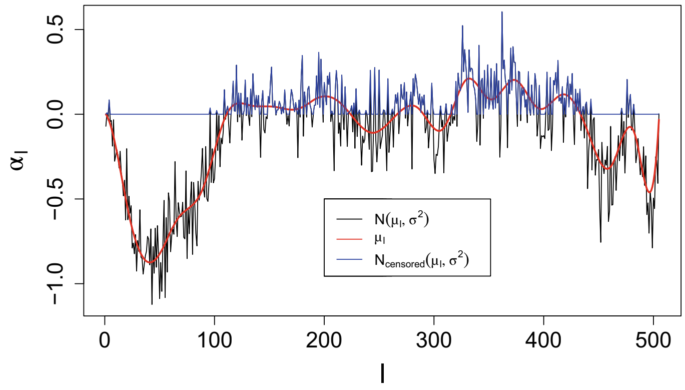

FIGURE 6 Simulation of $N(\mu_\ell, \sigma^2)$ for $\ell = 1, \ldots, 505$ with mean trend (red).

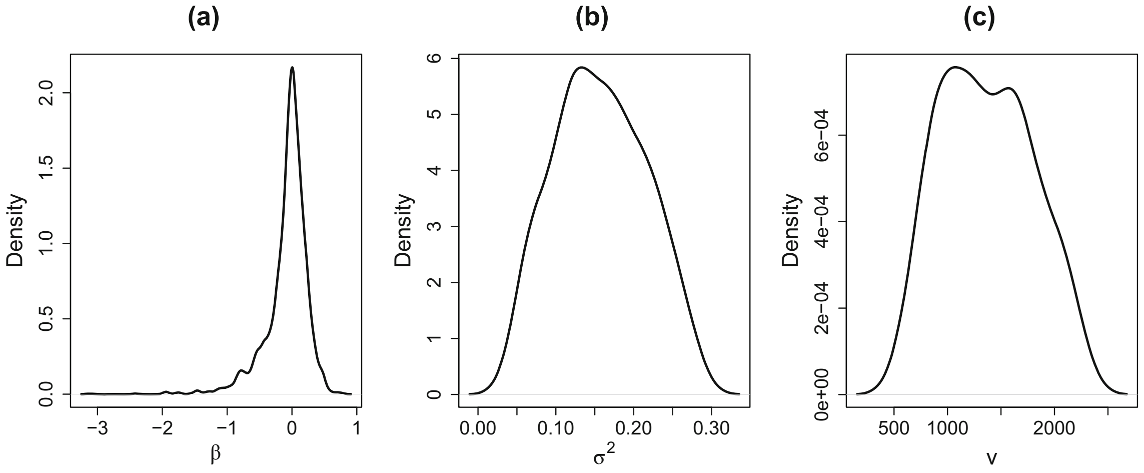

FIGURE 7 Distribution of β parameters used in α_ℓ model (a), σ^2 parameters for the α_ℓ model (b), and variance parameters of $G(\cdot)$ (c).

5.3 | Model choices and estimates for the Gaussian process

The final model choices are to specify the particular covariance function of the Gaussian process $G(\cdot)$, which populates $\boldsymbol{\theta}_G$, as well as C and $\{sp_1, \ldots, sp_C\}$ for the Daniell kernel smoother.

For this application we choose a Matérn covariance function for the Gaussian process which involves a variance parameter v, a range parameter ρ and a smoothness parameter ν (see Appendix for functional form of Matérn used). For these data, we select $C = 2$ with respect to the Daniell kernel smoother spans and $sp_1 = 55$ and $sp_2 = 201$. These specific choices were made based on a desire for a smoothed spectral estimate without an excess of bias that would occur from over-smoothing. When we set $C = 1$, there was not sufficient smoothing, resulting in a noisy spectral estimate, thus the choice to use $C = 2$. We chose sp_2 of about four times that of sp_1 so as to improve the smoothing without adding too much bias of the underlying spectral values at individual frequencies. Since these choices for C and sp_i, $i = 1, 2$ gave desirable results visually with regard to the periodogram, we chose not to pursue more complicated smoothing schemes for the sake of simplicity in the model.

Figure 8 shows the marginal densities of the 61 estimates from June and July of v and (ρ, ν) jointly. The parameter estimates for v, the variance parameter for the GP, yield a somewhat bimodal density with peaks around 1000 and 1600, which is interesting because the starting values for the optimization were chosen randomly, therefore this bi-modal distribution cannot be explained through a bias in starting values for the numerical optimization to find each $\boldsymbol{\theta}_G^*$. With regard to the joint density of $\{\rho, \nu\}$ in Figure 8, we also have a relatively bimodal surface where most of the density lies in the region where $1.2 \leq \nu \leq 2$ and $20 \leq \rho \leq 45$.

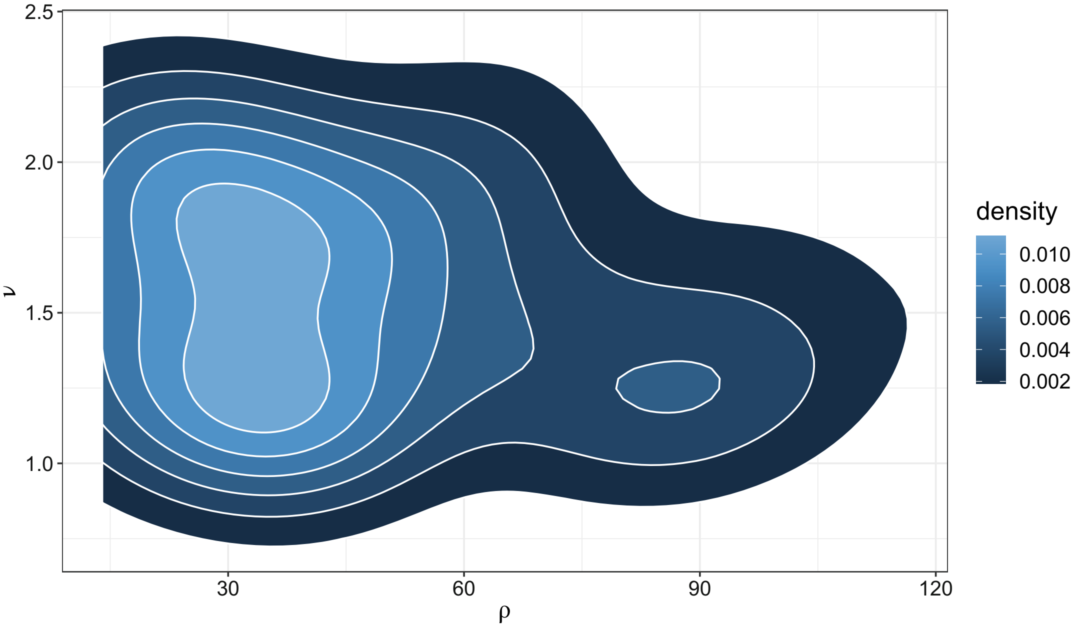

FIGURE 8 Joint density of smoothness (ν) and range (ρ) parameters of $G(\cdot)$.

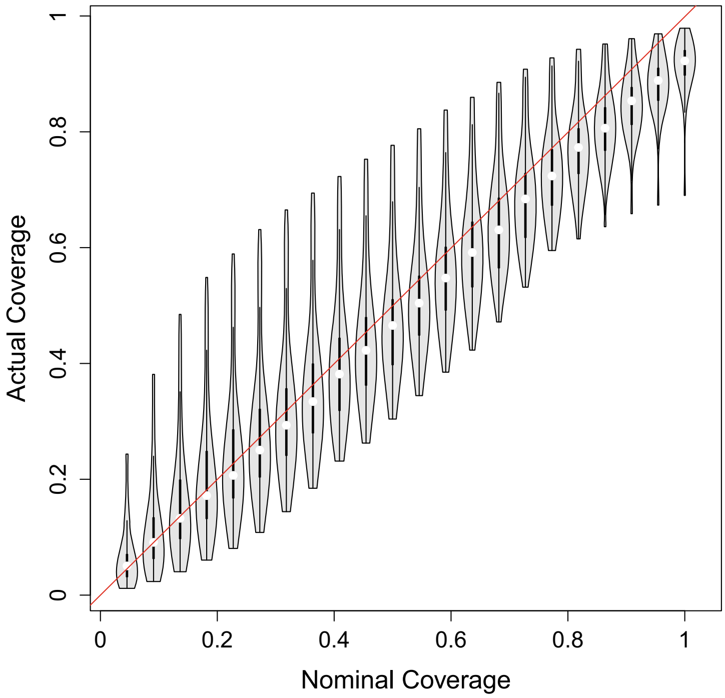

FIGURE 9 Violin plot of reliability results for August GHI simulations.

6 | VALIDATION FOR HAWAII DATA

In this section, we consider a suite of validation measures for the proposed model fitted to the Hawaii irradiance data. Recall that the goal is to capture the statistical properties of the irradiance process, and not to forecast any individual time or day of GHI. That is, this section assesses the unconditional distribution of modeled GHI to the actual distribution. To this end, we consider a cross-validation experiment in which the model is fit to data from June and July, whose implied statistical distributions are then compared to held-out August data.

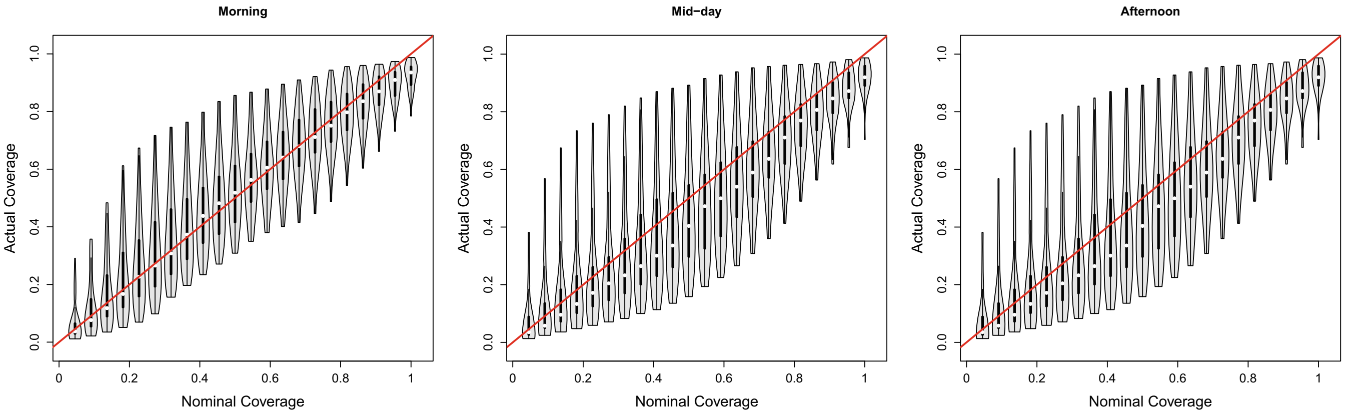

FIGURE 10 Reliability plots broken down by time of day: morning time (left), mid-day (middle), and afternoon/evening (right).

We train the model on the 61 days of June and July 2010, and use the simulation technique described above in Section 4.5 to generate 500 simulations of GHI at the 1 s resolution for each day of August 2010. To validate the model, we consider reliability plots, spectral density estimates, sample partial autocorrelation functions (PACFs), and sample autocorrelation functions of first differences (FDACFs), and marginal density estimates among other visual assessments.

Figure 9 contains a reliability plot assessing the statistical calibration of the model. For a given day in August, we generate 500 realizations of the stochastic model; at each second, we consider the actual coverage of confidence intervals represented by the ensemble of 500 simulations at nominal levels from 5% to 90% in 22 equal increments. For a well-calibrated model, we expect that reliability plot will follow the identity line where actual coverage approximately matches the nominal coverage. The violin plots capture variability over each day of August. Each violin in Figure 9 represents the shape of the distribution of actual percent coverage for a given nominal percent coverage over each of the 31 days of August. Generally the reliability plot indicates that our model is well-calibrated, with slight under dispersion at the largest intervals.

The behavior of GHI differs depending on the time of day, so it is natural to consider a more detailed assessment of reliability broken out by hours. Figure 10 shows three reliability plots grouped by time of day. In particular, we consider coverage in the morning (first third of the day), mid-day (middle third of the day), and afternoon (last third of the day). Note here the term "day" refers to total time of usable irradiance as defined in Section 2. We see that the mornings generally give the best reliability at most nominal coverages, while the mid-day and afternoon show slight under dispersion at mid-to-high coverage levels. Although there is slight under dispersion, the nominal levels are still captured within the interquartile range at each band.

Another set of statistics we use to compare the simulations to real GHI data is the estimated spectra (log periodograms), PACFs, and FDACFs. The functional boxplots (Sun & Genton, 2011) in Figure 11 show comparisons of these statistics for 500 GHI simulations versus those of the GHI measurement data for August 9. The simulations yield accurate statistics – notably neither the PACF nor the FDACF are trained statistics within the estimation procedure, so capturing these well indicates the model is appropriate for these held-out data.

We also show the density of GHI measurement data in August compared with the simulations overall, see Figure 12, as well as by times of day (morning, mid-day, afternoon), see Figure 13. The bimodal appearance of this density is interesting and is even, to some degree, mimicked in the morning/mid-day/afternoon breakdowns. This may be explained by the reaction of GHI to the existence of cloud events. If every day in August were clear, we would still expect to see a peak at higher GHI due to the nature of the clear sky curve. The existence of the second mode at lower GHI, however, is interesting and likely caused by cloudy days which cause higher incidence of lower GHI values. Our model matches the complicated density of high-frequency GHI throughout the day very well.

We also remark on the visual comparison of the simulated and measured GHI data. Figure 14 shows 8 days of GHI data. A selection of these 8 plots are real GHI measurement and a selection are days of simulated GHI data produced from our model. A question one might ask as a litmus test for the viability of this model is: can you tell which is which? While statistical or visual similarity each have varying levels of import, the two together make a successful model in the space of quantifying and assessing variability for solar irradiance data. The results shown here are evidence of the promise of this model.

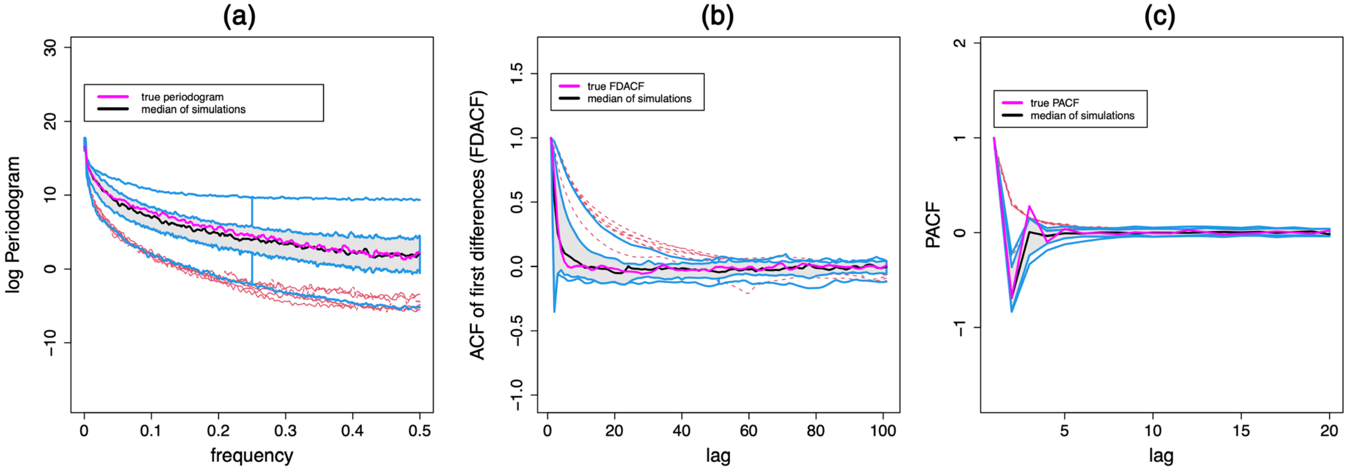

FIGURE 11 Functional boxplots of statistics of 500 simulations of one day of August compared with those of the real GHI measurement data. Log periodograms are shown in (a), sample autocorrelation function of first differences (FDACF) in (b), and sample partial autocorrelation function (PACF) in (c).

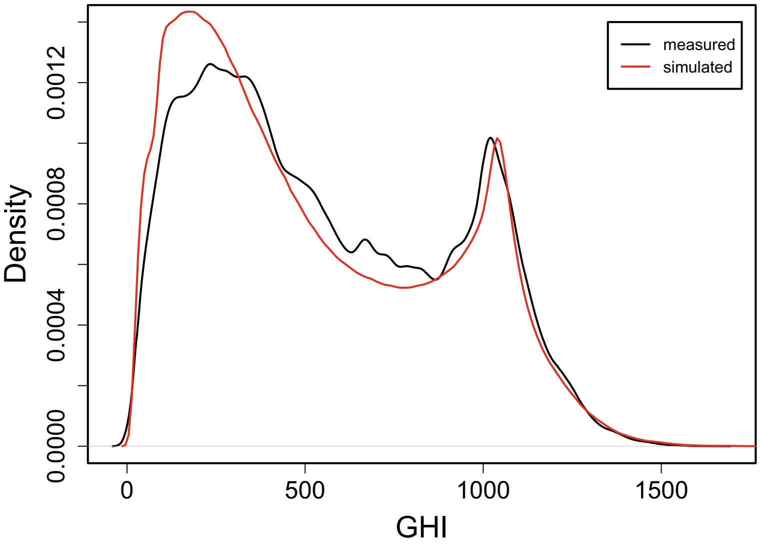

FIGURE 12 Density of August simulations compared with real GHI.

6.1 | Comparison to GARCH

We now turn to a comparison against a well-established nonlinear time series model. In particular, we consider a Generalized AutoRegressive Conditional Heteoskedasticity (GARCH) model (Berkes et al., 2003). The GARCH model is meant for time series in which there are periods of time with increased or decreased variability, which is certainly an appropriate concept for irradiance data modeling. As such, one might consider replacing $G(S(t))$ in Equation (2), with a GARCH(p,q) process. A GARCH(p,q) process, $\{w_k\}$ is defined by the equations

$$w_k = \sigma_k \epsilon_k \tag{12}$$

$$\sigma_k^2 = \omega + \sum_{1\le i\le p} \alpha_i w_{k-i}^2 + \sum_{1\le j\le q} \beta_j \sigma_{k-j}^2 \tag{13}$$

where $\omega \ge 0, \alpha_i \ge 0, \beta_j \ge 0$ are constants and ϵ_k are independent, identically distributed random variables (Berkes et al., 2003). Here ω is a mean term, the α_i and β_i are coefficients of the autoregressive terms in the GARCH process, and σ_k^2

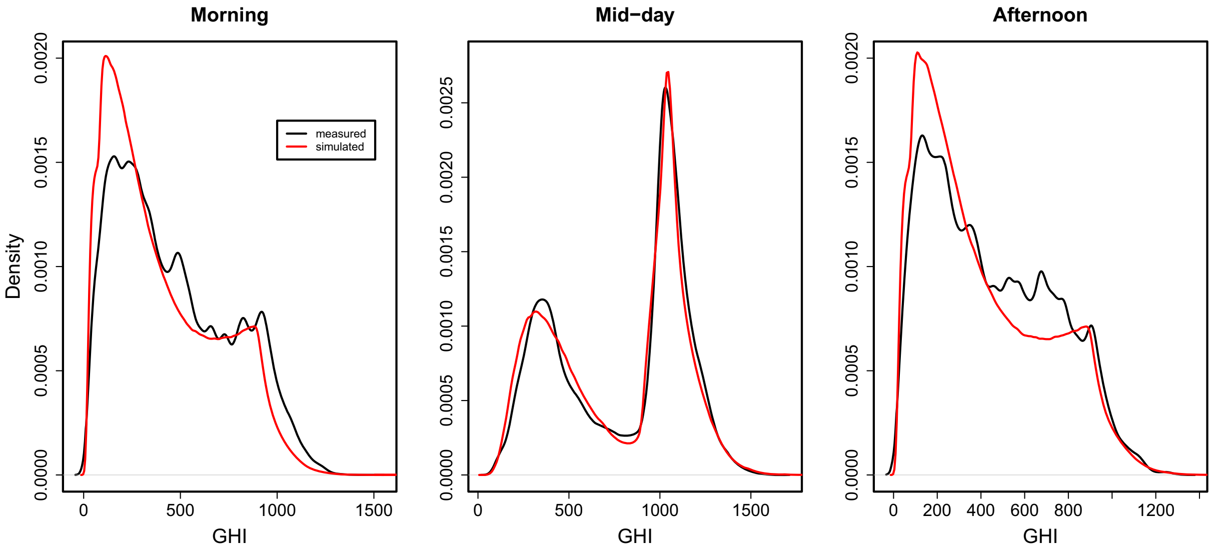

FIGURE 13 Density of August simulations compared with real GHI broken down by time of day.

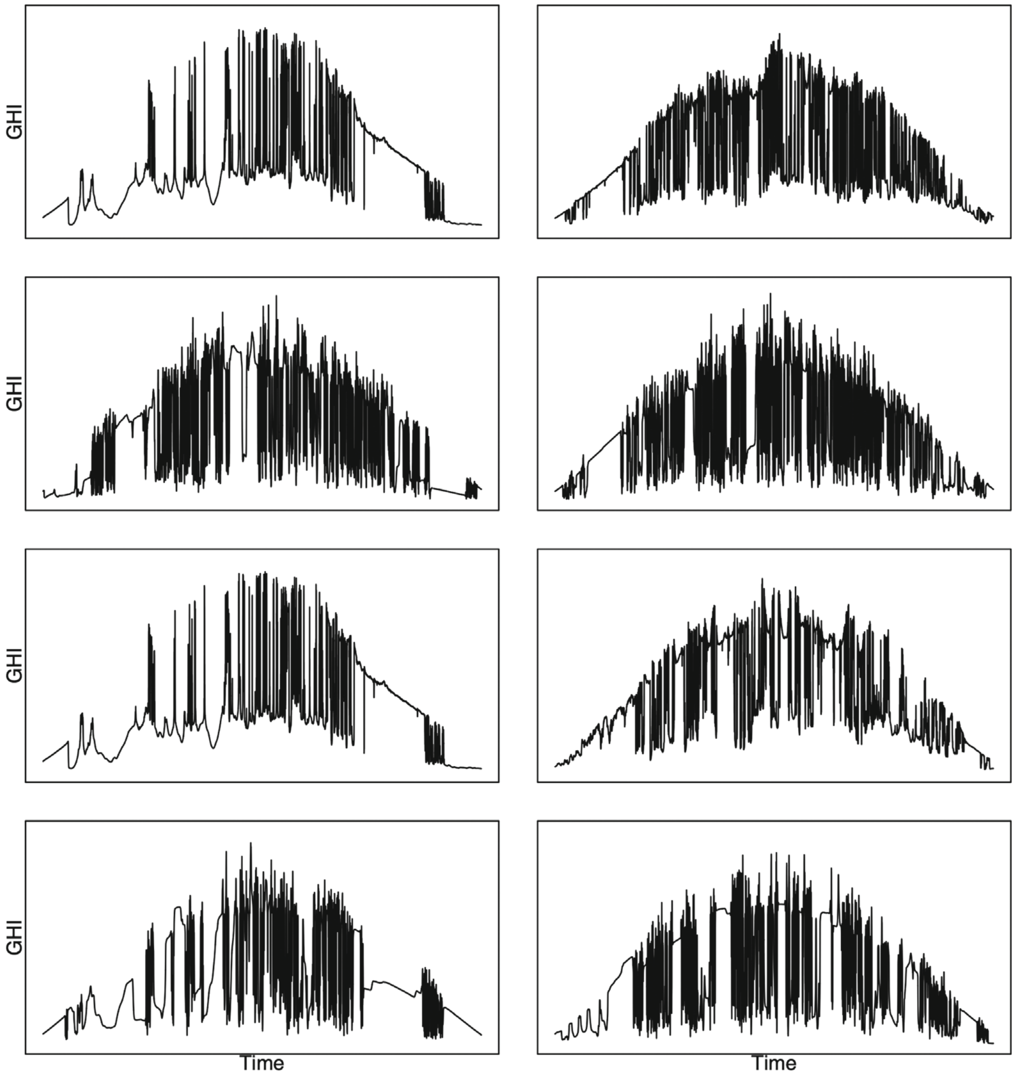

FIGURE 14 A selection of days of simulated GHI and real GHI measurement data.

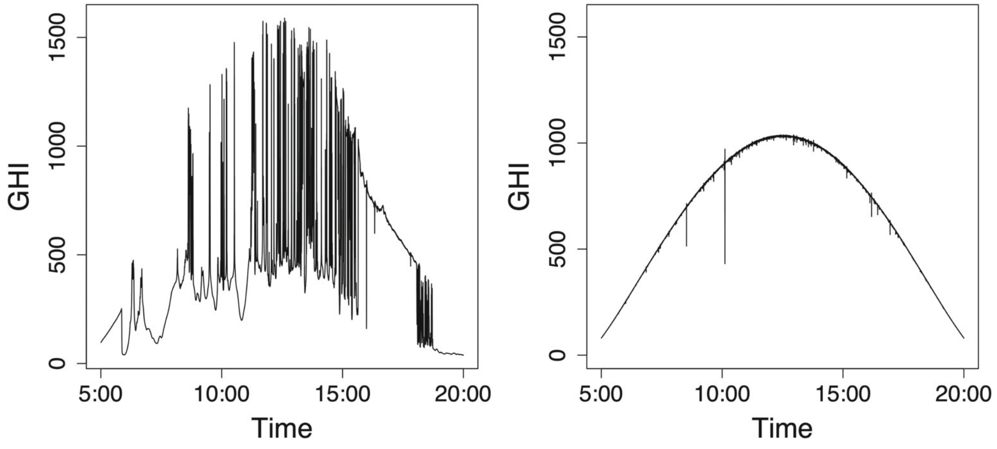

FIGURE 15 Comparison of real GHI measurement data (left) to a simulation using the GARCH model (right).

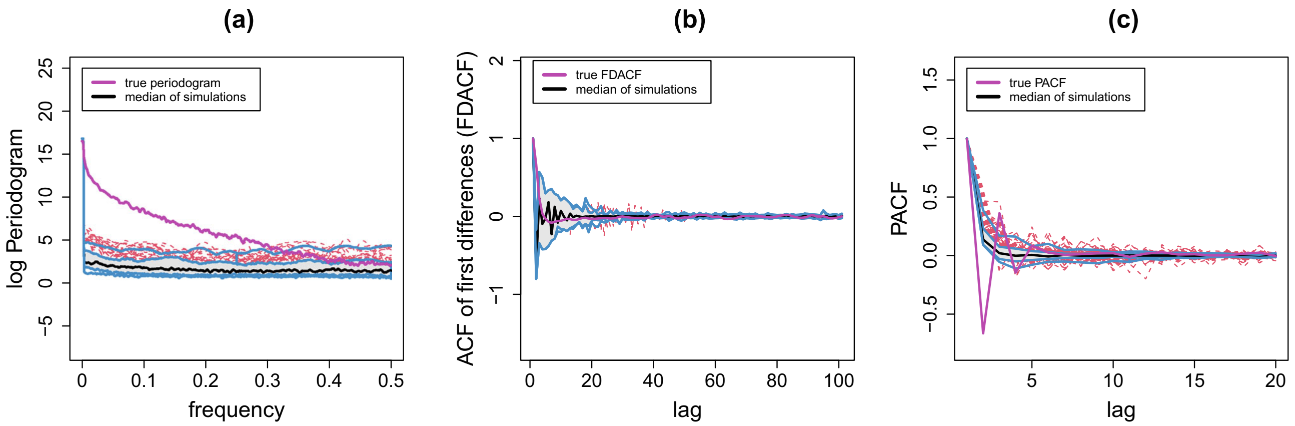

FIGURE 16 Functional boxplots of statistics of 500 GARCH-based simulations of 1 day of August compared with those of the real GHI measurement data.

is the variance of the process at time k. Thus, the variance of the GARCH at any given point is dependent on past squared observations and variances. Note that, while there may be overlap in notation used in the SGP model, these parameters are to be thought of as different quantities altogether and are recycled here because they are standard notation in the time series literature.

Let $\mathbf{Y}_g$ be a vector of simulated GHI data from a GARCH-based model, $\mathbf{X}_C$ be a vector of the clear sky GHI for that day, and F_Z be the cdf of the log-CSI data, which has been estimated based on the details outlined in Section 4.1. Then the GARCH-based model for a single day of 1 s GHI data can be written as

$$\mathbf{Y}_g = \mathbf{X}_C \exp\left(F_Z^{-1}(\Phi(\mathbf{W}))\right) \tag{14}$$

in much the same way as (2) where $\mathbf{W}$ is a vector of values generated from a GARCH(p, q) process. Then, given a set of actual GHI measurements, $\mathbf{y}$, choice of distribution of ϵ_k, and a selection of p, q, we may find the $\boldsymbol{\theta}_{\text{GARCH}} = \{\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q\}$ which will maximize the likelihood of the GARCH process.

We chose $p = 1, q = 1$ and $\epsilon_k \sim N(0, 1)$ for the sake of comparison to the most widely used GARCH process. Note that we also fit models with $p, q \in \{2, 3\}$, however they perform qualitatively similarly to the GARCH(1,1) model shown here. As the GARCH(1,1) is replacing the SGP in the model, we opt to estimate parameters for the process using a maximum likelihood approach, which is built into statistical packages in R. We fit parameters $\{\omega, \alpha, \beta\}$ using R's `garchFit` in the `fGarch` package (Wuertz et al., 2020).

Once the GARCH parameters are fit, we generate 500 GHI simulations using those parameters and test the GARCH-based model's ability to perform the task of mimicking the kinds of variability seen in real GHI measurement

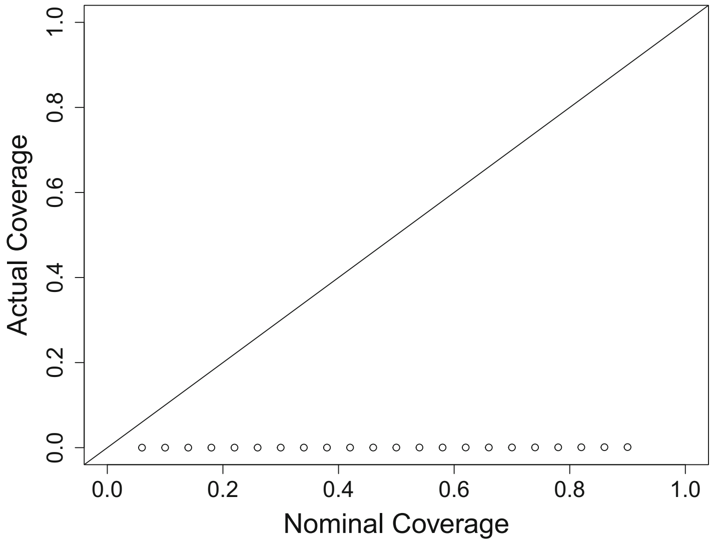

FIGURE 17 Reliability plot for 500 GARCH-based simulations of one day of GHI measurement data.

data using visual inspection and the same statistical metrics for comparison as shown in Section 6. Upon initial visual inspection of these simulations, it is clear that this model does not perform as well as our proposal at capturing the kinds of variability seen in GHI measurement data, see Figure 15. Figure 16 shows the analagous plots from Figure 11 for this GARCH-based model. In addition to its lack of visual similarity to real GHI data, this model does not perform as well as our proposal in terms of capturing the statistics–in particular with regard to the log-periodogram. The true log-periodogram of the day of data these simulations were based on generally falls well outside the range of the functional boxplot for the majority of frequencies. The PACF in Figure 16c shows that the simulations do not tell the same story as the true data in terms of how the process' partial autocorrelation looks–particularly at smaller lags. The FDACF does not show as drastic of differences from the truth as the log-periodogram and the PACF, but the general pattern of the FDACF is markedly different in the truth (magenta) as compared with the median of the simulations (black) in Figure 16b.

Figure 17 shows the reliability plot for these 500 simulations, which illustrates the inability of the GARCH-based model to capture realistic uncertainty in irradiance, further supporting the need for improved approaches for irradiance data simulation, such as our proposed model.

7 | DISCUSSION

In this work we propose a new non-Gaussian approach to modeling high frequency solar irradiance data. Understanding variability of irradiance resource data is critical for power planning studies with solar PV, and is a different task than forecasting. Our model uses a new spline-subordinator that randomly time changes a stationary Gaussian process, resulting in a stochastic process that exhibits periods of amplified variability between calmer periods, mimicking clear atmospheric conditions and weather fronts. Our estimation procedure is stepwise, which allows for the practitioner to check each stage of the model. We apply our model to a high frequency in situ irradiance data set in Hawaii, and find that our approach well-captures variability, intermittency, and non-Gaussian distributional properties better than a competing common nonlinear time series model. The parameters of our model could potentially be of use to grid researchers in understanding the variability of irradiance data at the 1 s level. While this latter point is not the focus of the current paper, there is possibility for interpretation of model parameters, which could be explored in future work, especially when comparing multiple data sets from different climatological regions. For example, recall that the α coefficients of the spline-subordinator could be interpreted to represent the level of variability seen in the irradiance data during a given time window (e.g., $\alpha \gg 0$ means more variability and $\alpha \approx 0$ means smoother, clearer times of day). Therefore, inspection of plots like those found in Figure A3 might be an interesting comparison point between estimates of different locations. Put another way, it might be of interest to grid researchers to know: does the distribution of these α change depending on time or location? This is outside the scope of the current proposal, but would certainly be a direction for the application of this model.

Future research may be devoted to generalizing this approach to the space-time setting, in which multiple correlated time series trajectories of irradiance are necessary. Temporal nonstationarity of irradiances due to, say, seasonality, may be captured using a seasonally-varying model, which may be an interesting route of future research. Other investigations might include alternative subordinators that arise in the Lévy process literature.

ACKNOWLEDGMENTS

This research was supported by NSF grant DMS-1923062. The authors thank a reviewer for helpful comments that improved the exposition of the manuscript.

ORCID

William Kleiber https://orcid.org/0000-0003-0411-9108

REFERENCES

Amato, U., Andretta, A., Bartoli, B., Coluzzi, B., Cuomo, V., Fontana, F., & Serio, C. (1986). Markov processes and fourier analysis as a tool to describe and simulate daily solar irradiance. *Solar Energy*, *37*(3), 179–194. https://doi.org/10.1016/0038-092X(86)90075-7

Beaumont, M. A., Cornuet, J. M., Marin, J. M., & Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, *96*, 983–990.

Berkes, I., Horváth, L., & Kokoszka, P. (2003). GARCH processes: Structure and estimation. *Bernoulli*, *9*(2), 201–227. https://doi.org/10.3150/bj/1068128975

Berrocal, V. J., Raftery, A. E., & Gneiting, T. (2008). Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Annals of Applied Statistics*, *2*, 1170–1193.

Blaga, R., & Paulescu, M. (2018). Quantifiers for the solar irradiance variability: A new perspective. *Solar Energy*, *174*, 606–616.

Bright, J. M., Babacan, O., Kleissl, J., Taylor, P. G., & Crook, R. (2017). A synthetic, spatially decorrelating solar irradiance generator and application to a LV grid model with high PV penetration. *Solar Energy*, *147*, 83–98. https://doi.org/10.1016/j.solener.2017.03.018

Brockwell, P. J., & Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer. https://doi.org/10.1007/978-1-4419-0320-4

Castillejo-Cuberos, A., & Escobar, R. (2020). Understanding solar resource variability: An in-depth analysis, using chile as a case of study. *Renewable and Sustainable Energy Reviews*, *120*, 109664. https://doi.org/10.1016/j.rser.2019.109664

de Boor, C. (1978). *A Practical Guide to Splines*. Springer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

EIA (2021). *How Much Carbon Dioxide is Produced per Kilo Watt Hour of US Electricity Generation?* [U.s. energy information administration FAQs]. Energy Information Administration (EIA). https://www.eia.gov/tools/faqs/faq.php?id=74&t=11

el Alani, O., Ghennioui, H., & Ghennioui, A. (2020). Intra-day variability quantification from ground-based measurements of global solar irradiance. *Int J Renew Energy Res*, *10*, 1576–1587.

Euán, C., Sun, Y., & Reich, B. J. (2022). Statistical analysis of multi-day solar irradiance using a threshold time series model. *Environmetrics*, *33*(3), e2716. https://doi.org/10.1002/env.2716

Haupt, S. E., Kosovic, B., Jensen, T., Lee, J., Jimenez Munoz, P., Lazo, J., Cowie, J., McCandless, T., Pearson, J., Wiener, G., Alessandrini, S., Delle Monache, L., Yu, D., Peng, Z., Huang, D., Heiser, J., Yoo, S., Kalb, P., Miller, S., Rogers, M., & Hinkleman, L. (2016). *The sun4cast® Solar Power Forecasting System: The Result of the Public-Private-Academic Partnership to Advance Solar Power Forecasting*. UCAR/NCAR. OpenSky. 10.5065/D6N58JR2

Huang, J., Korolkiewicz, M., Agrawal, M., & Boland, J. (2013). Forecasting solar radiation on an hourly time scale using a coupled auto regressive and dynamical system (CARDS) model. *Solar Energy*, *87*, 136–149. https://doi.org/10.1016/j.solener.2012.10.012

Huang, R., Huang, T., Gadh, R., & Li, N. (2012). Solar generation prediction using the arma model in a laboratory-level micro-grid. Paper presented at: 2012 IEEE Third International Conference on Smart Grid Communications (SmartGridComm) (pp. 528–533). 10.1109/SmartGridComm.2012.6486039.

Jimenez, P. A., Hacker, J. P., Dudhia, J., Haupt, S. E., Ruiz-Arias, J. A., Gueymard, C. A., Thompson, G., Eidhammer, T., & Deng, A. (2016). WRF-solar: Description and clear-sky assessment of an augmented NWP model for solar power prediction. *Bulletin of the American Meteorological Society*, *97*(7), 1249–1264. https://doi.org/10.1175/BAMS-D-14-00279.1

Kleiber, W., Katz, R. W., & Rajagopalan, B. (2012). Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. *Water Resources Research*, *48*. https://doi.org/10.1029/2011WR011105

Kleiber, W., Sain, S., Madaus, L., & Harr, P. (2023). Stochastic tropical cyclone precipitation field generation. *Environmetrics*, *34*, e2766. https://doi.org/10.1002/env.2766

Larrañeta, M., Fernandez-Peruchena, C., Silva-Pérez, M. A., Lillo-bravo, I., Grantham, A., & Boland, J. (2019). Generation of synthetic solar datasets for risk analysis. *Solar Energy*, *187*, 212–225. https://doi.org/10.1016/j.solener.2019.05.042

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

Lew, D., Brinkman, G., Ibanez, E., Florita, A., Heaney, M., Hodge, B. M., Hummon, M., Stark, G., King, J., Lefton, S. A., Kumar, N., Agan, D., Jordan, G., & Venkataraman, S. (2013). *The Western Wind and Solar Integration Study Phase* (Vol. *2*. (NREL/TP-5500- 55588, 1220243)). Office of Scientific and Technical Information (OSTI). https://doi.org/10.2172/1220243

Nguyen, A., Velay, M., Schoene, J., Zheglov, V., Kurtz, B., Murray, K., Torre, B., & Kleissl, J. (2016). High PV penetration impacts on five local distribution networks using high resolution solar resource assessment with sky imager and quasi-steady state distribution system simulations. *Solar Energy*, *132*, 221–235. https://doi.org/10.1016/j.solener.2016.03.019

Prasad, A. A., & Kay, M. (2020). Assessment of simulated solar irradiance on days of high intermittency using WRF-solar. *Energies*, *13*(2), 385. https://doi.org/10.3390/en13020385 Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Pratt, J. W., & Gibbons, J. D. (1981). *Kolmogorov-smirnov two-sample tests*. In *Concepts of Nonparametric Theory* (pp. 318–344). Springer. https://doi.org/10.1007/978-1-4612-5931-2_7

Ramírez, A. F., Valencia, C. F., Cabrales, S., & Ramírez, C. G. (2021). Simulation of photo-voltaic power generation using copula autoregressive models for solar irradiance and air temperature time series. *Renewable Energy*, *175*, 44–67. https://doi.org/10.1016/j.renene.2021.04.115

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, *3*(4), 425–441. https://doi.org/10.1214/ss/1177012761 Publisher: Institute of Mathematical Statistics.

Reikard, G. (2009). Predicting solar radiation at high resolutions: A comparison of time series forecasts. *Solar Energy*, *83*(3), 342–349. https://doi.org/10.1016/j.solener.2008.08.007

Sengupta, M., & Andreas, A. (2010). Oahu solar measurement grid (1-year archive): 1-second solar irradiance; oahu, hawaii (data). https://doi.org/10.5439/1052451 Type: dataset

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. https://doi.org/10.1016/j.rser.2018.03.003

Smith, C. J., Bright, J. M., & Crook, R. (2017). Cloud cover effect of clear-sky index distributions and differences between human and automatic cloud observations. *Solar Energy*, *144*, 10–21. https://doi.org/10.1016/j.solener.2016.12.055

Sun, F., Gramacy, R. B., Haaland, B., Lu, S., & Hwang, Y. (2019). Synthesizing simulation and field data of solar irradiance. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *12*(4), 311–324. https://doi.org/10.1002/sam.11414

Sun, Y., & Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, *20*(2), 316–334. https://doi.org/10.1198/jcgs.2011.09224 Publisher: Taylor & Francis.

Tankov, P. (2003). *Financial Modelling with Jump Processes* (1st ed.). Chapman and Hall/CRC.

Weiss, A., & Hays, C. J. (2004). Simulation of daily solar irradiance. *Agricultural and Forest Meteorology*, *123*(3), 187–199. https://doi.org/10.1016/j.agrformet.2003.12.002

Wilks, D. S., & Wilby, R. L. (1999). The weather generation game: a review of stochastic weather models. *Progress in Physical Geography*, *23*, 329–357.

Wuertz, D., Setz, T., Chalabi, Y., Boudt, C., Chausse, P., & Miklovac, M. (2020). *fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling*. R package version 3042.83.2.

Yang, D., Alessandrini, S., Antonanzas, J., Antonanzas-Torres, F., Badescu, V., Beyer, H. G., Blaga, R., Boland, J., Bright, J. M., Coimbra, C. F. M., David, M., Frimane, Â., Gueymard, C. A., Hong, T., Kay, M. J., Killinger, S., Kleissl, J., Lauret, P., Lorenz, E., … Zhang, J. (2020). Verification of deterministic solar forecasts. *Solar Energy*, *210*, 20–37. https://doi.org/10.1016/j.solener.2020.04.019

Zhang, W., Kleiber, W., Florita, A. R., Hodge, B.-M., & Mather, B. (2019). Modeling and simulation of high-frequency solar irradiance. *IEEE Journal of Photovoltaics*, *9*(1), 124–131. https://doi.org/10.1109/JPHOTOV.2018.2879756 Conference Name: IEEE Journal of Photovoltaics.

Zhang, W., Kleiber, W., Hodge, B.-M., & Mather, B. (2022). A nonstationary and non-gaussian moving average model for solar irradiance. *Environmetrics*, *33*(3), e2712. https://doi.org/10.1002/env.2712

How to cite this article: Berry, C. M., Kleiber, W., & Hodge, B. -M. (2023). Subordinated Gaussian processes for solar irradiance. *Environmetrics*, e2800. https://doi.org/10.1002/env.2800

APPENDIX

A.1 Definitions and formulae

Matérn Covariance

$$\frac{\pi^{1/2}\phi\rho^{2\nu}}{2^{\nu-1}\Gamma(\nu+1/2)}\left(\frac{1}{\rho}|t|\right)^{\nu}\mathcal{K}_{\nu}\left(\frac{1}{\rho}|t|\right) \tag{A1}$$

where $\mathcal{K}_\nu$ is the modified Bessel function of the second kind, Γ is the gamma function, ρ is a range parameter, and ν is a smoothness parameter.

Censored Normal Density:

$$f(x) = \Phi((\epsilon-\mu)/\sigma)\mathbb{I}_{\{x\leq\epsilon\}} + (1-\Phi((x-\mu)/\sigma))\mathbb{I}_{\{x>\epsilon\}} \tag{A2}$$

Pareto Density:

$$\frac{\theta\eta^{\theta}}{x^{\theta+1}}, \quad \eta > 0, \quad \theta > 0, \quad x \geq \eta \tag{A3}$$

Gamma Density:

$$\frac{(x-c)^{\alpha-1}\exp(\frac{-(x-c)}{\beta})}{\beta^{\alpha}\Gamma(\alpha)}, \quad \alpha > 0, \quad \beta > 0, \quad x \geq c \tag{A4}$$

A.2 Lévy versus spline-subordinators

We give a short explanation as to the motivation of the choice for the spline-subordinator over a true Lévy subordinator. For example, let $S_L(t)$ be a Gamma subordinator. This means, by definition, that $S_L(t) > 0$, for $t_2 > t_1$ $S_L(t_2) \geq S_L(t_2)$, and $S_L(t_2) - S_L(t_1) \overset{iid}{\sim} \Gamma(\alpha, \beta)$.

We select 1 day of GHI data from which we have estimated parameters for the spline-subordinator, call it $S_S(t)$ for the purposes of this section, using the Estimation and Model Specifics outlined in this paper. Additionally, for a selection of $(\alpha, \beta) = (0.1, 10)$, a realization of $S_L(t)$ is shown (Figure A1).

Figure A2 shows a comparison of a day of real GHI 1 s data to simulations using $S_L(t)$ and using $S_S(t)$. This illustrates that using $S_L(t)$ doesn't afford the same mix of clear and variable time periods that one sees in real GHI data and that the spline-subordinators do a better job of capturing this behavior.

A.3 Discussion of independence of β_k

Here we discuss the choice of treating the β_k as independent in the model for the α_ℓ, which are the coefficients of the spline-subordinator. In this work, we treat the β_k for all k, as independent as opposed to using a 25 dimensional joint

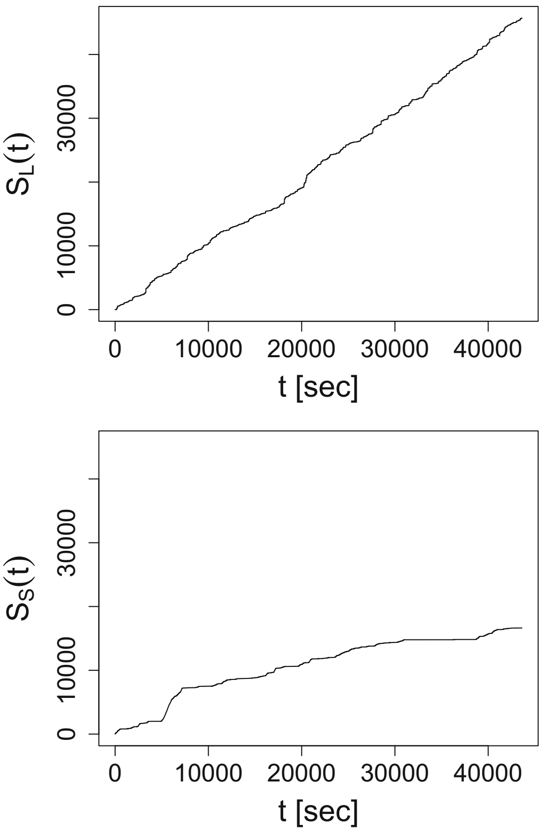

FIGURE A1 A realization of $S_L(t)$ with $(\alpha, \beta) = (0.01,100)$ (top) and $S_S(t)$, the spline-subordinator (bottom)

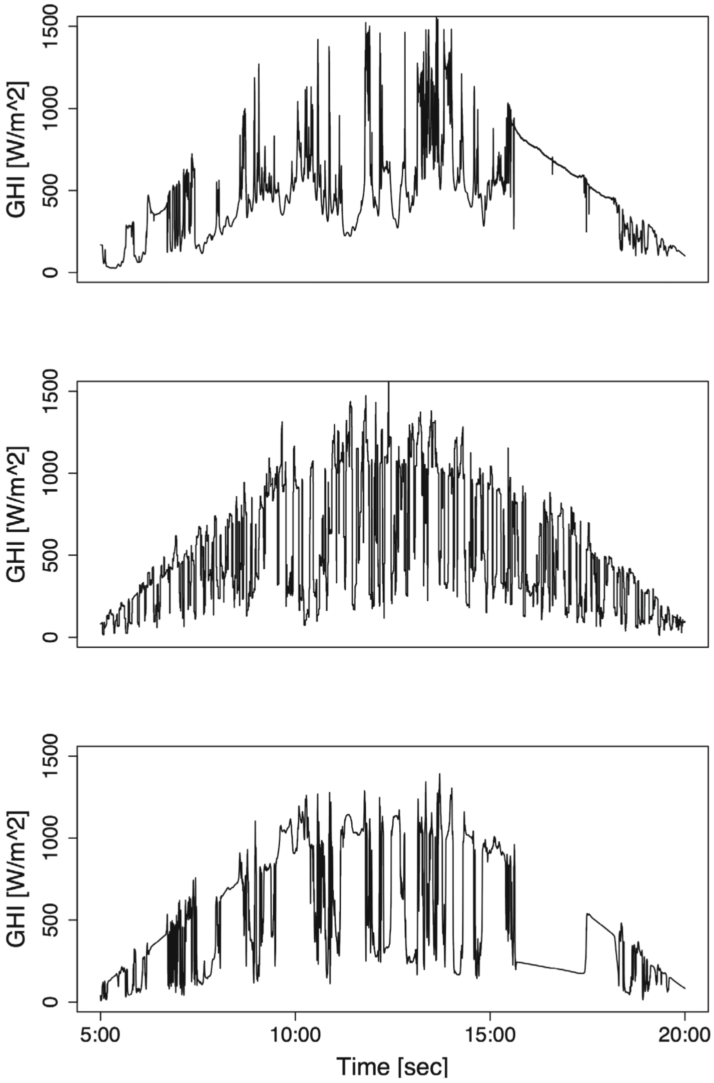

FIGURE A2 A day of real GHI 1 s data (top), a simulation of this day using $S_L(t)$ (middle), and a simulation of this day using $S_S(t)$ (bottom).

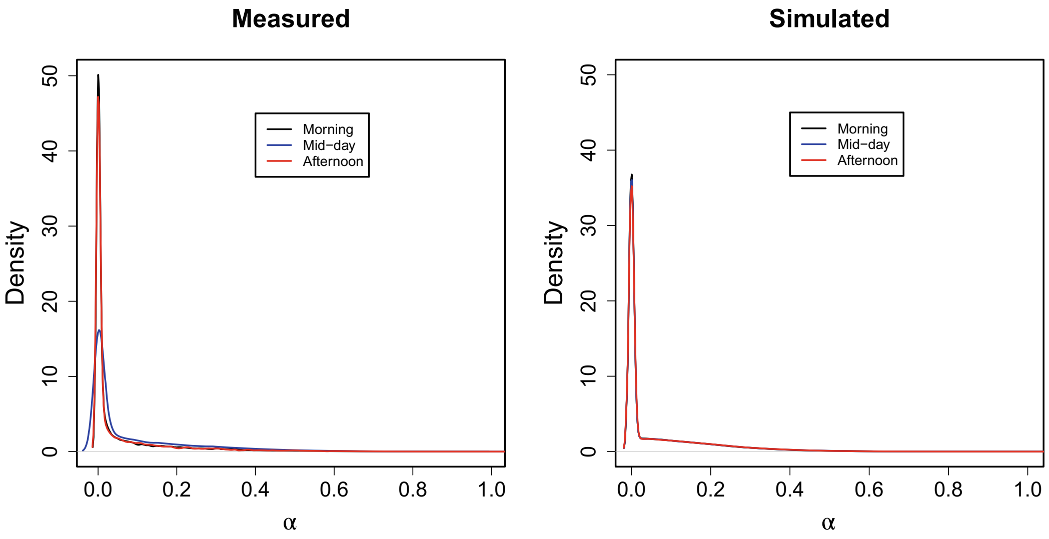

FIGURE A3 Distribution of estimated α_ℓ parameters from measured data (left) compared to distribution of 1000 days of simulated α_ℓ (right) grouped by time of day.

distribution for each $\boldsymbol{\beta}$. There are drawbacks to this assumption, for instance, the left plot of Figure A3 depicts a clear difference in how the α_ℓ are distributed during the middle of the day versus the beginning and end of the day. This does not support the assumption that β_k are independent; however, we found that treating the β_k as independent allows for more varied simulations and we still achieves good results.