The Annals of Applied Statistics
2018, Vol. 12, No. 4, 2096–2120
https://doi.org/10.1214/18-AOAS1145

BOTTOM-UP ESTIMATION AND TOP-DOWN PREDICTION: SOLAR ENERGY PREDICTION COMBINING INFORMATION FROM MULTIPLE SOURCES

BY YOUNGDEOK HWANG[*,1,2], SIYUAN LU[†,1] AND JAE-KWANG KIM[‡,§,3]

Sungkyunkwan University[*], *IBM Thomas J. Watson Research Center*[†], *Iowa State University*[‡] *and KAIST*[§]

Accurately forecasting solar power using the data from multiple sources is an important but challenging problem. Our goal is to combine two different physics model forecasting outputs with real measurements from an automated monitoring network so as to better predict solar power in a timely manner. To this end, we propose a new approach of analyzing large-scale multilevel models with great computational efficiency requiring minimum monitoring and intervention. This approach features a division of the large scale data set into smaller ones with manageable sizes, based on their physical locations, and fit a local model in each area. The local model estimates are then combined sequentially from the specified multilevel models using our novel bottom-up approach for parameter estimation. The prediction, on the other hand, is implemented in a top-down matter. The proposed method is applied to the solar energy prediction problem for the U.S. Department of Energy's SunShot Initiative.

1. Introduction. Solar energy's contribution to the total energy mix is rapidly increasing. As the most abundant form of renewable energy resource, solar electricity is projected to supply 14% of the total demand of the contiguous U.S. by 2030, and 27% by 2050, respectively [Margolis, Coggeshall and Zuboy (2012)]. Having a high proportion of solar energy in the electric grid, however, poses significant challenges because solar power generation has inherent variability and uncertainty due to varying weather conditions [Denholm and Margolis (2007), Ela, Milligan and Kirby (2011)]. For instance, the variability can result in steep ramps of solar power being injected into the grid causing system reliability issues. Moreover, the uncertainty of solar power often oblige system operators to hold extra reserves of conventional power generation at significant cost. Accurate forecasting of solar power can improve system reliability and reduce reserve cost [Orwig et al.

Received December 2016; revised January 2018.

[1] Supported by U.S. Department of Energy Contract DE-EE0006017.

[2] Supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2017R1C1B5075436).

[3] Supported in part by U.S. NSF Grant MMS-1733572.

Key words and phrases. Big data prediction, large-scale monitoring data, multilevel model, physics models, solar power generation.

(2015), Zhang et al. (2015a)]. Variation in solar power generation is primarily associated with cloud movement, formation, and dissipation. Depending on whether the goal is minute-, hour-, or day-ahead forecast, sky imagery [Marquez and Coimbra (2013), Chu et al. (2014)], satellite imagery [Hammer et al. (1999), Perez et al. (2002)], and numerical weather predictions [Perez et al. (2013), Mathiesen, Collier and Kleissl (2013)] are used, respectively. Instead of the direct use of these predictions, applying statistical methods on the forecasts from these numerical models can significantly improve the forecasting accuracy [Mathiesen and Kleissl (2011), Pelland, Galanis and Kallos (2013)].

Computer models have long been used to study and simulate complex physical systems. With increasing computational capacities, these models have become an essential part of various research areas [Welch et al. (1992), Santner, Williams and Notz (2003), Wu (2015)]. These computer models have been extensively used in the service industry, in particular, to predict or simulate the environmental variables in various scales. For example, Jiang et al. (2015) considered a computer model to simulate the data center thermal system, and Klein et al. (2015) discussed a general platform to take advantage of various environmental models to develop predictive tools in industry.

Such application requires a specific methodological focus. Statistical analysis of such models typically involve three modules [Qian and Wu (2008), Liu, Bayarri and Berger (2009)]: (1) the computer model itself; (2) real measurement data; (3) the discrepancy between the computer model and the real process. A key issue in many service industry applications is leveraging the second and third module to make an accurate prediction; in other words, the goal is to develop a framework to make predictions by matching the computer model outputs with the historical observations from the sensor network.

This task is closely related to model calibration to choose the optimal parameters for the computer model [Bayarri et al. (2007)]. However, our focus is on building an additional layer of an empirical model and calibrating its parameters to improve the prediction rather than optimizing the computer model itself, because iterative model running is not possible. In service industry applications, the computer model is a small part of the larger system, hence a reasonable parameter setting is set and minimally changed for maintaining the system reliability [Klein et al. (2015)]. The main focus of the statistical methodology is how to efficiently exploit the model outputs and field data to make accurate prediction and inference, while accommodating the structure, scale, and availability of the data. In this line, Gramacy et al. (2015) and Wong, Storlie and Lee (2017) proposed approaches to avoid expensive computation in a computer model context, while flexibly incorporating data sets in a large scale.

Using the outputs of complex physical models in real applications, however, poses significant challenges in statistical modeling. The main problems are generally three-fold. First, in industrial applications, the amount of data is often very large because measurements are obtained through automated systems. For example, in our application in Section 4, 1522 sensors collect monitoring data every 15 minutes. Hence, the method's capacity to incorporate a massive amount of data is indispensable. Second, the computation method must be expeditious yet able to take advantage of large scale data from physical models. Both efficiency and reliability are essential to avoid intensive computing and system collapse. Third, the available data is often complex. A hierarchical structure often exists and the information varies at each level. This requires a modeling approach that can naturally handle such complexity.

In this work, we propose a framework to exploit the abundance of physical model forecasting outputs and real measurements from an automated monitoring network, using multilevel models. Our method addresses the aforementioned challenges for large scale industrial applications. The proposed bottom-up approach has a computational convenience for parameter estimation when implemented in an automated system, because it does not rely on the Markov chain Monte Carlo (MCMC) method. From a modeling point of view, our approach is a Bayesian hierarchical model, whose inference are obtained by the Expectation-Maximization (EM) algorithm. Using EM, the convergence of algorithm is easy to check and the computation can be implemented in an automated fashion. Other approaches such as mixed-effects modeling have been proposed for multilevel models [e.g., Bates and Pinheiro (1998)]. However, to apply those methods, one may need to assemble a single data file from different types of data structures, which can cause reliability issues. Instead, our approach uses a summary version of data calculated from the granular level, the storage of data, use of computer memory, and communication during the computation, which are convenient for large-scale data. It is similar to the split-and-conquer approach [Chen and Xie (2014)], but ours exploits the data set's inherent partitioned hierarchical structure. It naturally pairs well with a distributed storage system and computing resources; there is no need to assemble a single data file for the entire data set, hence it works very well for a large scale prediction problem.

The idea of dividing a large data set and then combining the results from each to obtain a global inference is not new [Pratola et al. (2014), Scott et al. (2016)]. When each partition is a random sample of the data, then the likelihood becomes easier to work with. On the contrary, our approach focuses on exploiting the already partitioned data structure to minimize the management overhead. We start a simple modeling process at a distributed level and incorporate such structure in the proceeding steps so that the modeling can incorporate the partition structure. Gelman et al. (2014) also considers the bottom-up approach for inference using expectation propagation (EP) based on local tilted distribution to make the inference more affordable when the data set is large.

A multilevel model is a powerful tool which allows for model heterogeneity across areas but simultaneously borrows strength from other areas [Gelman (2006)]. It has been particularly popular in areas where data are often structured hierarchically, such as education [Goldstein (1986)] or epidemiology [Wong and Mason (1985)]. While these traditional applications of multilevel models deal with the lack of independence between measurements and model heterogeneity, our approach is intended to incorporate the large scale data in a flexible and efficient manner.

The remainder of the paper is organized as follows. Section 2 describes the overall problem and background. Section 3 delineates the proposed model and methodology. Section 4 demonstrates the application of the proposed method to large-scale solar monitoring data. We conclude with some remarks and discussions in Section 5.

2. Global Horizontal Irradiance. In this section we describe our solar energy application and the overall problem. Our goal is to improve Global Horizontal Irradiance (GHI) prediction over the contiguous United States (CONUS). GHI is the total amount of shortwave radiation received by a surface horizontal to the ground, which is the sum of Direct Normal Irradiance (DNI, the amount of solar radiation received by a surface perpendicular to the rays that come from the direction of the sun), Diffuse Horizontal Irradiance (DHI, the amount received by a surface that has been diffused by the atmosphere), and ground-reflected radiation. The amount of solar electricity produced from photovoltaic systems (i.e., solar panels) is directly associated with GHI; a day with high GHI means a high solar electricity production. Hence, GHI forecast is of main interest of the participants in the electricity market.

To make a solar electricity prediction for a site where a solar panel is located, prediction of GHI is made first. Then it is fed to a irradiance-to-power conversion model [Soto, Klein and Beckman (2006)]—which takes the panel specification as its inputs—to make a forecast of the amount of solar electricity generated. Among the different sources of error for solar power forecasting, the error of GHI forecast dominates. Thus it is critical to improve GHI forecasting to obtain more accurate forecasting and inference of solar power.

To monitor the GHI, sensors are located over CONUS. The collected observations are obtained from the sensor locations marked on Figure 1. The GHI readings are recorded at 1522 locations in 15-min intervals. Hence, the data size grows very quickly; every day, thousands of additional observations are added. The data from each site is separately stored in the database indexed by the site location. The readings are obtained from various types of sensors, which may cause some potential variability among different locations. In our application, we consider two models to forecast GHI: Short-Range Ensemble Forecast [SREF, Du and Tracton (2001)] and North American Mesoscale Forecast System [NAM, Skamarock et al. (2008)]. A vis-á-vis comparison of the outputs from the two models is presented in Figure 2. They share a common overall trend, however there are certain discrepancies between the two model outputs. The model outputs are available at any location in a pre-specified computational domain, which covers the entire CONUS. The

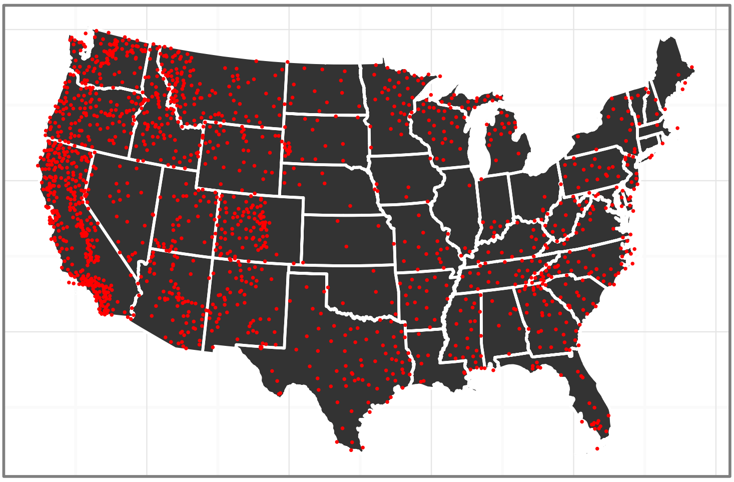


FIG. 1. *The map of the* 1522 *monitoring network locations*, *marked by dots*.

model output is stored at every hour, but can be matched with 15-minute interval measurement data after post-processing.

Our goal is to develop an approach for parameter estimation and prediction with the following three considerations:

Computational efficiency: uses little memory and communication during computation.

Applicability: readily handles a variety of complex data.

Practicality: runs with a deterministic algorithm so that convergence is simple to check and the method can be implemented easily in practice.

The methodology focuses on the solar energy application for a practical implementation. Similar problems are easily found in industrial applications, as the problem of prediction using real measurements at monitoring sensors and computer model outputs covering the entire domain is prevalent. For those problems, the overall

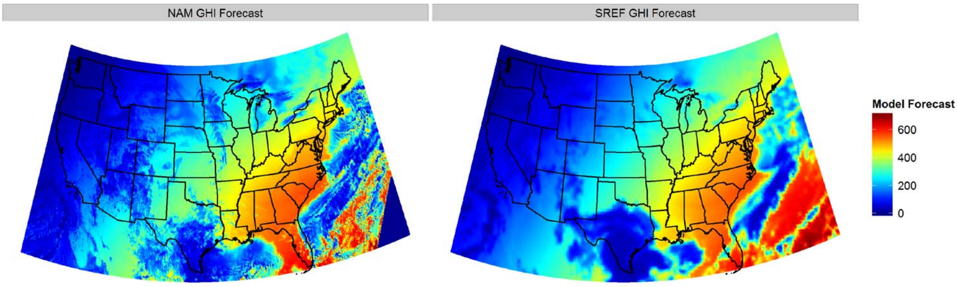


FIG. 2. *A vis-á-vis comparisons of two computational model outputs at one time point*. *Left panel shows NAM output*, *while right panel SREF*.

goal is often managing a large complex system, such as micro-climate within a facility [Jiang et al. (2015)] or environmental monitoring [Liu et al. (2016)].

In practice, the raw data related to individual data set is often too large to be moved from the servers to a centralized location for a combined modeling. For example, when working with satellite images or high temporal resolution data for a large number of individual sites, it is difficult to gather the data to one place. In contrast, our approach conducts the very first step of data summarization in the cloud, which gives two benefits: (1) avoid the burden of transferring the data, (2) the computation for individual sites is naturally distributed over many servers. Our approach stems from these considerations.

3. Methodology.

In this section we introduce our modeling approach within the context of our application. Estimation and prediction are described in more general terms to be applied to a wider array of problems.

3.1. *Multilevel model for GHI.*

Assume that the sensors are partitioned into H exhaustive and non-overlapping groups. Group h consists of n_h sensors, where the ith sensor in group h collects the measurements y_{hij} for $j = 1, \dots, n_{hi}$. The predicted GHI outputs from the two models are available at the site location as scalar values and used as covariates, $\mathbf{x}_{hij}$. Information at sensor or group level, $\mathbf{c}_h$ and $\mathbf{c}_{hi}$ are also available. Note that the covariates $\mathbf{x}$ are often more widely available than y_{hij}'s; in our application in Section 4, the computer model output is available not only at monitoring sites but also everywhere in the spatial domain of interest. We assume that n_h can be relatively small while n_{hi} is usually large, because managing the existing sensors and taking additional measurements from them usually does not cost much, while deploying new monitoring sensors often causes considerable cost.

We assume that the GHI measurement y_{hij} follows

$$y_{hij} = \mathbf{x}_{hij}\boldsymbol{\theta}_{hi} + e_{hij}, \tag{3.1}$$

with a latent site-specific parameter $\boldsymbol{\theta}_{hi}$, where the covariates $\mathbf{x}_{hij}$ has NAM and SREF model output as predictors including an intercept term, and $e_{hij} \sim t(0, \sigma^2_{hi}, \nu_{hi})$, where σ^2_{hi} is scale parameter and ν_{hi} are the degree of freedom [Lange, Little and Taylor (1989)]. The coefficients for two predictors operate as the weight parameters for two computer models.

We assume that the level two model follows

$$\boldsymbol{\theta}_{hi} \sim N(\boldsymbol{\beta}_h, \boldsymbol{\Sigma}_h) \tag{3.2}$$

for some group-specific parameters $\boldsymbol{\beta}_h = (\beta_{h1}, \dots, \beta_{hp})$ and $\boldsymbol{\Sigma}_h$. For further presentation, define the length H vector of jth coefficients of $\boldsymbol{\beta}_h$ concatenated over H groups

$$\boldsymbol{\beta}_{(j)} = (\beta_{1j}, \dots, \beta_{Hj}),$$

and similarly define $\hat{\boldsymbol{\beta}}_{(j)}$. The subscript j is omitted hereinafter as we model each parameter separately but in the same manner. To incorporate the spatial dependence that may exist in the data, we assume that the level three model follows

$$\boldsymbol{\beta} \sim N(\boldsymbol{F}\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3.3}$$

where $\boldsymbol{F}$ is a pre-specified H by q model matrix, and $\boldsymbol{\mu}$ is the mean parameter of length q. In the analysis in Section 4.2, $\boldsymbol{F}$ is chosen to be $\mathbf{1}$, length H vector of 1's and hence μ is a scalar. The spatial covariance $\boldsymbol{\Sigma}$ has its (k, l)th element

$$\Sigma_{kl} = \operatorname{cov}(\beta_k, \beta_l) = \tau^2 \exp(-\rho d_{kl}),$$

where d_{kl} is the distance between the groups. The distance between two groups is defined to be the distance between the centroids of groups. Thus, the level one model is to describe the structure of the measurements and the computer model outputs, the level two model is for the intra-group structure, and the level three model is for the inter-group structure.

Note that a group is formed by collapsing several neighboring sites, hence the number of groups is less than that of sites. This also reduces the computational burden because the main computation in our spatial model relies on the number of spatial locations. Hence it is helpful to introduce the spatial components in the group level instead of the sensor level to provide computational benefit. As a trend over the spatial domain exists, including spatial component in the model improves the overall model accuracy, as discussed in Section 4.

Multilevel models are developed to incorporate the three considerations in Section 2. The overall data storage and modeling structure of our proposed model is summarized in Figure 3. Our approach builds up a hierarchy with the measurements by taking the following three steps. The first step is *summarization*. There is no direct measurement for the kth level model ($k \geq 2$), so we use the observations from the lower level model to obtain a "measurement" and construct an appropriate measurement model. The second step is *combination*; we combine the measurement model and structural model to build a prediction model using Bayes theorem. The third step is *learning*, in which we estimate the parameters by using the EM algorithm. Using EM instead of Gibbs sampling not only greatly facilitates computation but also results in an algorithm that can be run with minimal monitoring and intervention. In the bottom-up approach, the computation for each step uses a summary version to ease the storage of data and spare the use of computer memory despite the large amount of data. In the subsection below, we describe each step in detail.

3.2. *Bottom-up estimation.* In this section we give a detailed description of the estimation procedure. Since our modeling approach can be generally applied,

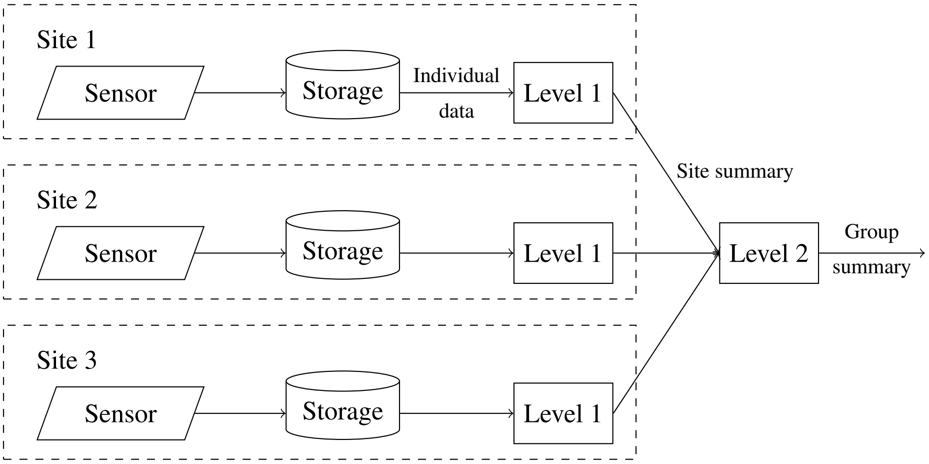


FIG. 3. *Overall description of the data storage and modeling structure, where the data is stored separately for each site.*

we present it with a generic model. We provide detailed steps for our model in Section 3.1 in the Appendix.

We first consider the level one and level two models,

$$\mathbf{y}_{hi} \sim f_1(\mathbf{y}_{hi}|\mathbf{x}_{hi};\boldsymbol{\theta}_{hi}), \tag{3.4}$$

$$\boldsymbol{\theta}_{hi} \sim f_2(\boldsymbol{\theta}_{hi}|\boldsymbol{c}_{hi};\boldsymbol{\zeta}_h), \tag{3.5}$$

where $\mathbf{y}_{hi} = (y_{hi1}, \dots, y_{hin_{hi}})^\top$ and $\mathbf{x}_{hi} = (\mathbf{x}_{hi1}^\top, \dots, \mathbf{x}_{hin_{hi}}^\top)^\top$ are the observations and covariates associated with the ith sensor in the hth group for the level one model, respectively, and $\boldsymbol{\theta}_{hi}$ is the parameter in the level one model. In (3.5), $\boldsymbol{\theta}_{hi}$ is treated as a random variable and linked to the unit-specific covariate $\boldsymbol{c}_{hi}$ and parameter $\boldsymbol{\zeta}_h$ in the level two model.

To estimate $\boldsymbol{\zeta}_h$ in (3.5), we use the three-step approach discussed in Section 2. In the summarization step, for each sensor, we treat $(\mathbf{x}_{hi}, \mathbf{y}_{hi})$ as a single data set to obtain the best estimator $\hat{\boldsymbol{\theta}}_{hi}$ of $\boldsymbol{\theta}_{hi}$, a fixed parameter. Define $g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi})$ to be the density of the sampling distribution of $\hat{\boldsymbol{\theta}}_{hi}$. This sampling distribution is used to build a measurement error model, where $\hat{\boldsymbol{\theta}}_{hi}$ is a measurement for the latent variable $\boldsymbol{\theta}_{hi}$, while (3.5) is a structural error model for $\boldsymbol{\theta}_{hi}$.

The sampling distribution $g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi})$ is combined with the level two model f_2 to obtain the marginal distribution of $\hat{\boldsymbol{\theta}}_{hi}$. Thus, the MLE of the level two parameter $\boldsymbol{\zeta}_h$ can be obtained by maximizing the log-likelihood derived from the marginal density of $\hat{\boldsymbol{\theta}}_{hi}$. That is, we maximize

$$\sum_{i}^{n_h} \log \int g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi}) f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)\, d\boldsymbol{\theta}_{hi} \tag{3.6}$$

with respect to $\boldsymbol{\zeta}_h$, *combining* $g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi})$ with $f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)$. The maximizer of (3.6) can be obtained by

$$\hat{\boldsymbol{\zeta}}_h = \arg\max_{\boldsymbol{\zeta}_h} \sum_{i=1}^{n_h} \mathbb{E}[\log\{f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)\} \mid \hat{\boldsymbol{\theta}}_{hi}; \boldsymbol{\zeta}_h]. \tag{3.7}$$

Note that $\boldsymbol{\zeta}_h$ is the parameter associated with the level two distribution, and (3.7) aggregates the information associated with $\hat{\boldsymbol{\theta}}_{hi}$ to estimate $\boldsymbol{\zeta}_h$.

To evaluate the conditional expectation in (3.7), we derive

$$p_2(\boldsymbol{\theta}_{hi} \mid \hat{\boldsymbol{\theta}}_{hi}; \boldsymbol{\zeta}_h) = \frac{g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi}) f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)}{\int g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi}) f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)\, d\boldsymbol{\theta}_{hi}}. \tag{3.8}$$

The level two model can be *learned* by the EM algorithm. Specifically, at the tth iteration of EM, we update $\boldsymbol{\zeta}_h$ by

$$\hat{\boldsymbol{\zeta}}_h^{(t)} = \arg\max_{\boldsymbol{\zeta}_h} \sum_{i=1}^{n_h} \mathbb{E}[\log\{f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)\} \mid \hat{\boldsymbol{\theta}}_{hi}; \boldsymbol{\zeta}_h = \hat{\boldsymbol{\zeta}}_h^{(t-1)}], \tag{3.9}$$

where the conditional expectation is with respect to the prediction model in (3.8) evaluated at $\hat{\boldsymbol{\zeta}}_h^{(t-1)}$, which is obtained from the previous iteration of the EM algorithm.

When $\hat{\boldsymbol{\theta}}_{hi}$ is the maximum likelihood estimator, we may use a normal approximation for $g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi})$. To see this, note that we use the following score equation:

$$\frac{\partial}{\partial \boldsymbol{\theta}_{hi}} \log f_1(\mathbf{y}_{hi} \mid \mathbf{x}_{hi}; \boldsymbol{\theta}_{hi}) = 0 \tag{3.10}$$

with respect to $\boldsymbol{\theta}_{hi}$ to obtain $\hat{\boldsymbol{\theta}}_{hi}$. Letting $l_{1hi}(\boldsymbol{\theta}_{hi}) = \log\{f_1(\mathbf{y}_{hi} \mid \mathbf{x}_{hi}; \boldsymbol{\theta}_{hi})\}$, Taylor expansion gives

$$\begin{aligned} f_1(\mathbf{y}_{hi} \mid \boldsymbol{\theta}_{hi}) &= \exp\{l_{1hi}(\boldsymbol{\theta}_{hi})\} \\ &\cong \exp\{l_{1hi}(\hat{\boldsymbol{\theta}}_{hi}) + l_{1hi}^{(1)}(\hat{\boldsymbol{\theta}}_{hi})(\boldsymbol{\theta}_{hi} - \hat{\boldsymbol{\theta}}_{hi}) \\ &\quad + 0.5(\boldsymbol{\theta}_{hi} - \hat{\boldsymbol{\theta}}_{hi})^\top l_{1hi}^{(2)}(\hat{\boldsymbol{\theta}}_{hi})(\boldsymbol{\theta}_{hi} - \hat{\boldsymbol{\theta}}_{hi})\}, \end{aligned}$$

where $l_{1hi}^{(1)}(\boldsymbol{\theta}_{hi})$ and $l_{1hi}^{(2)}(\boldsymbol{\theta}_{hi})$ are the first and the second order partial derivatives of $l_{1hi}(\boldsymbol{\theta}_{hi})$, respectively. By (3.10), we have $l_{1hi}^{(1)}(\hat{\boldsymbol{\theta}}_{hi}) = 0$ and hence the above expansion becomes

$$f_1(\mathbf{y}_{hi} \mid \boldsymbol{\theta}_{hi}) \cong \exp\{l_{1hi}(\hat{\boldsymbol{\theta}}_{hi})\} \exp\{-0.5(\boldsymbol{\theta}_{hi} - \hat{\boldsymbol{\theta}}_{hi})^\top I_{1hi}(\hat{\boldsymbol{\theta}}_{hi})(\boldsymbol{\theta}_{hi} - \hat{\boldsymbol{\theta}}_{hi})\}, \tag{3.11}$$

where $I_{1hi}(\boldsymbol{\theta}_{hi}) = E\{-l_{1hi}^{(2)}(\boldsymbol{\theta}_{hi})\}$ is the Fisher information of $\boldsymbol{\theta}_{hi}$. Thus, asymptotically, $\hat{\boldsymbol{\theta}}_{hi}$ is a sufficient statistic for $\boldsymbol{\theta}_{hi}$ and normally distributed with mean $\boldsymbol{\theta}_{hi}$

and variance $\{I_{1hi}(\boldsymbol{\theta}_{hi})\}^{-1}$. Hence, the prediction model (3.8) is approximately equal to

$$p_2(\boldsymbol{\theta}_{hi} \mid \mathbf{y}_{hi}; \boldsymbol{\zeta}) = \frac{f_1(\mathbf{y}_{hi} \mid \mathbf{x}_{hi}; \boldsymbol{\theta}_{hi}) f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)}{\int f_1(\mathbf{y}_{hi} \mid \mathbf{x}_{hi}; \boldsymbol{\theta}_{hi}) f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)\, d\boldsymbol{\theta}_{hi}},$$

because, by (3.11), $f_1(\mathbf{y}_{hi} \mid \mathbf{x}_{hi}; \boldsymbol{\theta}_{hi}) \cong K(\mathbf{x}_{hi}, \mathbf{y}_{hi}) g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi})$ holds for some $K(\mathbf{x}_{hi}, \mathbf{y}_{hi})$ which does not depend on $\boldsymbol{\theta}_{hi}$.

The EM algorithm approach is a convenient way of solving (3.7) iteratively. If f_2 is the density of the normal distribution with mean $\mu_{2hi}(\boldsymbol{\zeta}_h) = \mathbb{E}(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)$ and variance $V_{2hi}(\boldsymbol{\zeta}_h) = \mathrm{var}(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)$, the conditional distribution in (3.7) is also normal with mean $\mu^*_{hi}(\boldsymbol{\zeta})$ and variance $V^*_{hi}(\boldsymbol{\zeta})$, where

$$\mu^*_{hi}(\boldsymbol{\zeta}_h) = \frac{V_{2hi}(\boldsymbol{\zeta})\hat{\boldsymbol{\theta}}_{hi} + V_{1hi}\mu_{2hi}(\boldsymbol{\zeta}_h)}{V_{2hi}(\boldsymbol{\zeta}_h) + V_{1hi}}$$

and

$$V^*_{hi}(\boldsymbol{\zeta}_h) = \frac{V_{2hi}(\boldsymbol{\zeta}_h)V_{1hi}}{V_{2hi}(\boldsymbol{\zeta}_h) + V_{1hi}},$$

with $V_{1hi} = \{I_{1hi}(\hat{\boldsymbol{\theta}}_{hi})\}^{-1}$. In the E-step, the conditional expectation in (3.9) is taken with respect to the normal distribution with mean $\mu^*_{hi}(\boldsymbol{\zeta}_h^{(t)})$ and variance $V^*_{hi}(\boldsymbol{\zeta}_h^{(t)})$. The M-step updates parameter $\boldsymbol{\zeta}_h$ by solving (3.9) with respect to $\boldsymbol{\zeta}_h$ where the conditional expectation is evaluated from the E-step.

If f_2 is not normal, then the marginal density in (3.8) may not have a known closed form and hence the mean score equation in (3.9) is difficult to solve. In this case, we may use the Monte Carlo EM algorithm to estimate the parameters. Instead of MCMC algorithms such as Metropolis–Hastings, we can use the parametric fractional imputation (PFI) of Kim (2011) to simplify the computation in the E-step of the EM algorithm. The PFI method uses the importance sampling in the E-step and the normalized importance weights in computing the mean score function in the M-step. Specifically, we use

$$\boldsymbol{\theta}^{*(t)}_{hi,m} \sim f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \hat{\boldsymbol{\zeta}}_h^{(t)}), \qquad m = 1, \dots, M$$

to generate M Monte Carlo samples of $\boldsymbol{\theta}_{hi}$. Then compute the fractional weights

$$w^{*(t)}_{hi,m} \propto g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}^{*(t)}_{hi,m})$$

with $\sum_{m=1}^{M} w^{*(t)}_{hi,m} = 1$. In the M-step, we update $\boldsymbol{\zeta}_h$ by maximizing

$$l_w(\boldsymbol{\zeta}_h) = \sum_{i=1}^{n_{hi}} \sum_{m=1}^{M} w^{*(t)}_{hi,m} \log f_2(\boldsymbol{\theta}^{*(t)}_{hi,m} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)$$

with respect to $\boldsymbol{\zeta}_h$.

Once each $\hat{\boldsymbol{\zeta}}_h$ is obtained, we can use $\{\hat{\boldsymbol{\zeta}}_h; h = 1, \ldots, H\}$ as the summary of observations to estimate the parameters in the level three model. Let the level three model be expressed as

$$\boldsymbol{\zeta}_h \sim f_3(\boldsymbol{\zeta}_h | \boldsymbol{c}_h; \boldsymbol{\xi}), \tag{3.12}$$

where $\boldsymbol{c}_h$ are the covariates associated with group h and $\boldsymbol{\xi}$ is the parameter associated with the level three model. Estimation can be done in a similar fashion to the level two parameters. However, $\boldsymbol{\zeta}_h$ is now treated as a latent variable, and $\hat{\boldsymbol{\zeta}}_h$ as a measurement. Similar to (3.6), we maximize

$$\sum_{h=1}^{H} \log \int g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h) f_3(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \boldsymbol{\xi})\, d\boldsymbol{\zeta}_h \tag{3.13}$$

with respect to $\boldsymbol{\xi}$ to obtain $\hat{\boldsymbol{\xi}}$, where $g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h)$ is the sampling distribution of $\hat{\boldsymbol{\zeta}}_h$, which is assumed to be normal. The EM algorithm can be applied by iteratively solving

$$\hat{\boldsymbol{\xi}}^{(t)} = \arg\max_{\boldsymbol{\xi}} \sum_{h=1}^{H} \mathbb{E}[\log\{f_3(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \boldsymbol{\xi})\} \mid \hat{\boldsymbol{\zeta}}_h; \boldsymbol{\xi} = \hat{\boldsymbol{\xi}}^{(t-1)}], \tag{3.14}$$

where the conditional distribution is with respect to the distribution with density

$$p_3(\boldsymbol{\zeta}_h \mid \hat{\boldsymbol{\zeta}}_h; \boldsymbol{\xi}) = \frac{g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h) f_3(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \boldsymbol{\xi})}{\int g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h) f_3(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \boldsymbol{\xi})\, d\boldsymbol{\zeta}_h}$$

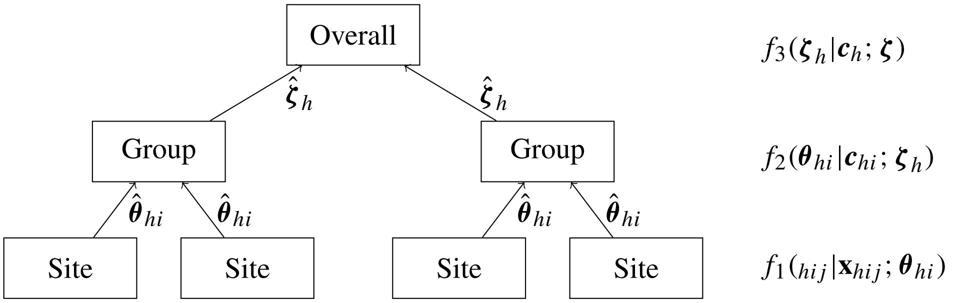
evaluated at $\boldsymbol{\xi} = \hat{\boldsymbol{\xi}}^{(t-1)}$. The level three model can be chosen flexibly depending on the usage, as it was in the lower levels. It can incorporate the hierarchical nature inherent to the data structure, such as utility distribution zone or state border line, or a pragmatic grouping to potentially adjust the computational burden. Figure 4 summarizes the bottom-up approach at three levels.

3.3. *Top-down prediction.* In this section we describe the prediction procedure. In contrast to the bottom-up approach of Section 3.2, the prediction is made in a top-down fashion.

To describe the top-down approach to prediction, consider the three-level models in (3.4), (3.5), and (3.12). The bottom-up estimation in Section 3.2 provides a way of estimating the parameters, $\boldsymbol{\theta}_{hi}$, $\boldsymbol{\zeta}_h$, and $\boldsymbol{\xi}$ by $\hat{\boldsymbol{\theta}}_{hi}$, $\hat{\boldsymbol{\zeta}}_h$, and $\hat{\boldsymbol{\xi}}$, respectively, using EM algorithm or maximizing the marginal likelihood.

Our goal is to predict unobserved y_{hij} values from the above models using the parameter estimates. The goal is to generate Monte Carlo samples of y_{hij} from

$$\begin{aligned} &p(y_{hij} \mid \mathbf{x}_{hij}; \hat{\boldsymbol{\theta}}_{hi}, \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) \\ &\quad = \frac{\int\int f_1(y_{hij} \mid \mathbf{x}_{hij}; \boldsymbol{\theta}_{hi}) p_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{\zeta}_h, \hat{\boldsymbol{\theta}}_{hi}, \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) p_3(\boldsymbol{\zeta}_h \mid \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}})\, d\boldsymbol{\zeta}_h\, d\boldsymbol{\theta}_{hi}}{\int\int\int f_1(y_{hij} \mid \mathbf{x}_{hij}; \boldsymbol{\theta}_{hi}) p_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{\zeta}_{hi}, \hat{\boldsymbol{\theta}}_{hi}, \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) p_3(\boldsymbol{\zeta}_h \mid \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}})\, d\boldsymbol{\zeta}_h\, d\boldsymbol{\theta}_{hi}\, dy_{hij}}, \end{aligned} \tag{3.15}$$


Overall
$\hat{\boldsymbol{\zeta}}_h$
$\hat{\boldsymbol{\zeta}}_h$
Group
Group
$\hat{\boldsymbol{\theta}}_{hi}$
$\hat{\boldsymbol{\theta}}_{hi}$
$\hat{\boldsymbol{\theta}}_{hi}$
$\hat{\boldsymbol{\theta}}_{hi}$
Site
Site
Site
Site
$f_3(\boldsymbol{\zeta}_h|\boldsymbol{c}_h;\boldsymbol{\zeta})$
$f_2(\boldsymbol{\theta}_{hi}|\boldsymbol{c}_{hi};\boldsymbol{\zeta}_h)$
$f_1(_{hij}|\mathbf{x}_{hij};\boldsymbol{\theta}_{hi})$

$$\hat{\boldsymbol{\xi}} = \arg\max_{\boldsymbol{\xi}} \textstyle\sum_{h=1}^{H} \log \int g_2(\hat{\boldsymbol{\zeta}}_h|\boldsymbol{\zeta}_h) f_3(\boldsymbol{\zeta}_h|\boldsymbol{c}_h;\boldsymbol{\xi})\, d\boldsymbol{\zeta}_h$$

$$\hat{\boldsymbol{\zeta}}_h = \arg\max_{\boldsymbol{\zeta}_h} \textstyle\sum_{i=1}^{n_h} \log \int g_1(\hat{\boldsymbol{\theta}}_{hi}|\boldsymbol{\theta}_{hi}) f_2(\boldsymbol{\theta}_{hi}|\boldsymbol{c}_{hi};\boldsymbol{\zeta}_h)\, d\boldsymbol{\theta}_{hi}$$

$$\hat{\boldsymbol{\theta}}_{hi} = \arg\max_{\boldsymbol{\theta}_{hi}} \textstyle\sum_{j=1}^{n_{hi}} \log f_1(y_{hij}|\mathbf{x}_{hij};\boldsymbol{\theta}_{hi})$$

FIG. 4. *The summary of bottom-up approach at three levels.*

where $p_2(\boldsymbol{\theta}_{hi} \mid \hat{\boldsymbol{\theta}}_{hi}, \boldsymbol{\zeta}_h, \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) = p_2(\boldsymbol{\theta}_{hi} \mid \hat{\boldsymbol{\theta}}_{hi}, \boldsymbol{\zeta}_h)$ and $p_3(\boldsymbol{\zeta}_h \mid \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}})$ are the predictive distribution of $\boldsymbol{\theta}_{hi}$ and $\boldsymbol{\zeta}_h$, respectively.

To generate Monte Carlo samples from (3.15), we use the top-down approach. We first compute the predicted values of $\boldsymbol{\zeta}_h$ from the level three model,

$$p_3(\boldsymbol{\zeta}_h \mid \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) = \frac{g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h) f_3(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \hat{\boldsymbol{\xi}})}{\int g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h) f_3(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \hat{\boldsymbol{\xi}})\, d\boldsymbol{\zeta}_h}, \tag{3.16}$$

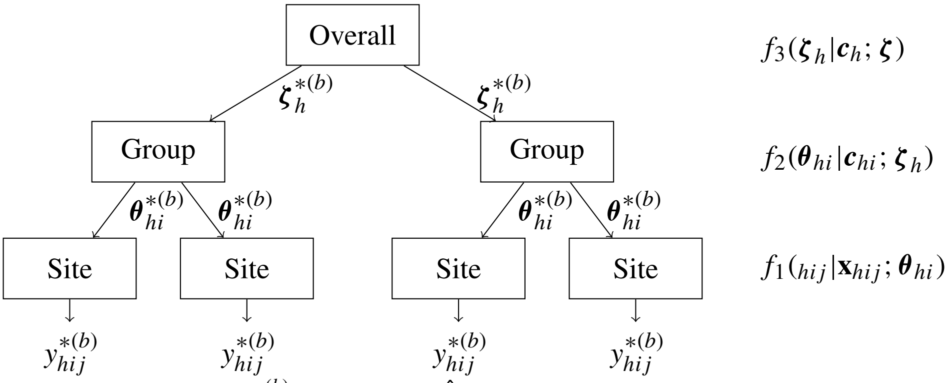
where $g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h)$ is the sampling distribution of $\hat{\boldsymbol{\zeta}}_h$. Also, given the Monte Carlo sample $\boldsymbol{\zeta}_h^*$ obtained from (3.16), the predicted values of $\boldsymbol{\theta}_{hi}$ are generated by (3.8). The best prediction for y_{hij} is

$$\hat{y}_{hij}^* = \mathbb{E}_3[\mathbb{E}_2\{\mathbb{E}_1(y_{hij} \mid \mathbf{x}_{hij}, \boldsymbol{\theta}_{hi}) \mid \hat{\boldsymbol{\theta}}_{hi}; \boldsymbol{\zeta}_h\} \mid \hat{\boldsymbol{\zeta}}_h; \hat{\boldsymbol{\xi}}], \tag{3.17}$$

where subscripts 3, 2, and 1 denote the expectation with respect to p_3, p_2, and f_1, respectively. Thus, while the bottom-up approach to parameter estimation starts with taking the conditional expectation with respect to p_1 and then moves on to p_2, the top-down approach to prediction starts with the generation of Monte Carlo samples from p_2 and then moves on to p_1 and f_1. Figure 5 summarizes the top-down approach, which contrasts with the bottom-up approach illustrated in Figure 4.

To estimate the mean squared prediction error of $\hat{y}_{hij}^*$ given by $M_{hij} = \mathbb{E}\{(\hat{y}_{hij}^* - y_{hij})^2\}$, we can use the parametric bootstrap approach [Hall and Maiti (2006), Chatterjee, Lahiri and Li (2008)]. In the parametric bootstrap approach, we first generate bootstrap samples of y_{hij} using the three-level model as follows:

1. Generate $\boldsymbol{\zeta}_h^{*(b)}$ from $f_3(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \hat{\boldsymbol{\xi}})$, for $b = 1, 2, \ldots, B$.


Overall
$f_3(\boldsymbol{\zeta}_h|\boldsymbol{c}_h;\boldsymbol{\zeta})$
$\boldsymbol{\zeta}_h^{*(b)}$
$\boldsymbol{\zeta}_h^{*(b)}$
Group
Group
$f_2(\boldsymbol{\theta}_{hi}|\boldsymbol{c}_{hi};\boldsymbol{\zeta}_h)$
$\boldsymbol{\theta}_{hi}^{*(b)}$
$\boldsymbol{\theta}_{hi}^{*(b)}$
$\boldsymbol{\theta}_{hi}^{*(b)}$
$\boldsymbol{\theta}_{hi}^{*(b)}$
Site
Site
Site
Site
$f_1(_{hij}|\mathbf{x}_{hij};\boldsymbol{\theta}_{hi})$
$y_{hij}^{*(b)}$
$y_{hij}^{*(b)}$
$y_{hij}^{*(b)}$
$y_{hij}^{*(b)}$

$$\boldsymbol{\zeta}_h^{*(b)} \sim f_3(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \hat{\boldsymbol{\xi}})$$

$$\boldsymbol{\theta}_{hi}^{*(b)} \sim f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h^{*(b)})$$

$$y_{hij}^{*(b)} \sim f_1(y_{hij} \mid \mathbf{x}_{hij}; \boldsymbol{\theta}_{hi}^{*(b)})$$

FIG. 5. *The summary of top-down approach at three levels.*

2. Generate $\boldsymbol{\theta}_{hi}^{*(b)}$ from $f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h^{*(b)})$, for $b = 1, 2, \dots, B$.
3. Generate $y_{hij}^{*(b)}$ from $f_1(y_{hij} \mid \mathbf{x}_{hij}; \boldsymbol{\theta}_{hi}^{*(b)})$, for $b = 1, 2, \dots, B$.

Once the bootstrap samples of $\mathbf{Y}^{*(b)} = \{y_{hij}^{*(b)}; h = 1, 2, \dots, H; i = 1, \dots, n_h; j = 1, \dots, m_{hi}\}$ are obtained, we can treat them as the original samples and apply the same estimation and prediction method to obtain the best predictor of y_{hij}. The mean squared prediction error (MSPE) M_{hij} can also be computed from the bootstrap sample. That is, we use

$$\hat{M}_{hij} = \mathbb{E}_* \{ (\hat{y}_{hij}^* - y_{hij})^2 \}$$

to estimate M_{hij}, where $\mathbb{E}_*$ denote the expectation with respect to the bootstrapping mechanism.

4. Prediction of Global Horizontal Irradiance.

In this section we give a detailed description of the available data. We also apply the proposed model and compare results to those of the comparators.

4.1. *Data description.*

We use 15 days of data for our analysis (from December 01, 2014 to December 15, 2014). There are 1522 sites to monitor GHI, where the number of available data varies between 12 and 517 observations, and the total number of observations is 557,284. To borrow strength from neighboring sites, we formed 50 groups that are spatially clustered by applying the K-means algorithm on the geographic coordinates. We assume the sites belonging to the same group

TABLE 1
A sample data file from one site in the analysis

Date time	GHI Meas	NAM GHI	SREF GHI
2014-12-01 15:15:00	10.16	8.85	7.79
2014-12-01 15:30:00	38.66	33.67	29.64
2014-12-01 15:45:00	73.30	63.83	56.20
2014-12-01 16:00:00	110.52	96.24	84.74
2014-12-01 16:15:00	148.53	108.02	119.85
2014-12-01 16:30:00	186.31	108.75	157.83

are homogeneous. The number of sites in each group, n_h, varies between 10 to 59. Depending on the goal, one can use other grouping schemes such as the distribution zone described in Zhang et al. (2015b). Calculated irradiance is available at every 0.1 degree, and is matched to the monitoring site location. A sample data file from one site is shown in Table 1.

In model (3.1), the degrees of freedom are assumed to be five in the analysis, but can be assumed unknown and estimated by the method of Lange, Little and Taylor (1989). The estimated spatial effect for two coefficients in (3.3) is depicted in Figure 6.

Since we are interested in the amount of irradiance, we first exclude zeros from both observed measurements and computer model outputs for the analysis. This operation does not cause much loss of information because the physics related to zero irradiance is well understood; before sunrise and after sunset, there is no solar irradiance. Thus, all values are positive and skewed to the right, and we used the logarithm transformation for both predictors and responses. Hereinafter, all variables are assumed to be log-transformed.

4.2. *Results.* Under the linear regression model in (3.1), the best prediction is $\hat{y}^*_{hij}$ in (3.17). We compared the multilevel approach with several other modeling

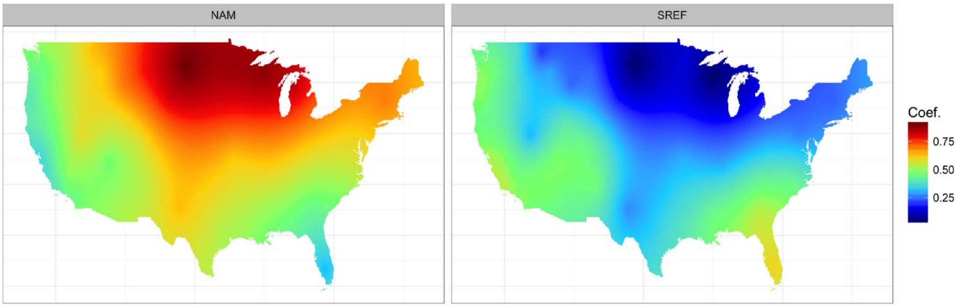


FIG. 6. *Spatial variation of the group-level coefficients from the second level for two computer models, where the left panel shows the NAM model and the right panel the SREF model.*

approaches. They can be categorized into non-spatial models and spatial models. We consider these alternatives as the applications are spatial in nature. For non-spatial models, we consider two methods: (1) Site-by-site model: fit a separate model for each individual site; (2) Global model: fit a single model for all sensor locations using the aggregate data combining all sensors. These are natural choices for the practitioner if simplicity is the first priority in implementation, but inherently limited as they do not incorporate the spatial nature of the data. For spatial models, we consider *Localized Approximate GP* [laGP, Gramacy and Apley (2015), Gramacy (2016)]. For spatial models, computation should be an important consideration because it can be prohibitively slow. laGP is specifically designed to handle a large scale spatial data, by actively choosing the subset of design points based on the prediction target. We let laGP search the space considering it's along rays emanating from the predictive location of interest [Gramacy and Haaland (2016)].

To evaluate the prediction accuracy, we conducted 10-fold cross-validation. The dataset is randomly partitioned into 10 subsamples. Of these 10 subsamples, one subsample was held out for validation, while the remaining nine subsamples are used to fit the model and obtain predicted values. The cross-validation process is repeated for each fold. The prediction from the fitted model was compared with the observed measurements in the log scale.

We considered two scenarios: (a) prediction made at observed sites, (b) prediction made at new sites. For scenario (a), we partitioned the time point into 10 sub-periods, while for (b) the sites into 10 sub-regions.

We compared the accuracy of different methods by the root mean squared prediction error (RMSPE), $\{N^{-1}\sum_j (y_{hij} - \hat{y}_{hij})^2\}^{1/2}$, with N being the size of the total data set. Table 2 presents the overall summary statistics for the accuracy of each method, calculated from cross validation. The standard deviation calculated over the subsamples are in parenthesis.

The rightmost column shows the overall accuracy. The global model suffers because it cannot incorporate the site-specific variation. On the contrary, the site model suffers from reliability issues for some sites because it does not use the information from neighboring sites. The multilevel approach strikes a fine balance between flexibility and stability. For a comprehensive comparison of each method, we evaluate the accuracy measure divided by the number of available data points for each site. As noted earlier, some stations may suffer from the data reliability problem. As such, the available sample size can vary from station to station, which affects the site-by-site model. When the prediction is made based on few available samples due to the data reliability issues, the inference can be unstable, affecting the accuracy of the prediction. The multilevel method can utilize information from other sites belonging to the same group, so it is particularly beneficial for locations with smaller sample sizes.

We then compare the coverage of the prediction intervals of each methods evaluated at different confidence levels. We obtain the prediction interval using (3.15)

TABLE 2
Root mean squared prediction error comparison of the different modeling methods, divided by the size of the training sample and overall

	Training sample size		
Method	**<200**	**≥200**	**Overall**
Multilevel	0.678 (0.129)	0.591 (0.052)	0.594 (0.055)
Site	1.344 (0.764)	0.593 (0.073)	0.632 (0.133)
Global	0.646 (0.038)	0.639 (0.009)	0.639 (0.009)
laGP	0.502 (0.027)	0.513 (0.011)	0.513 (0.012)

for confidence levels $0.8, 0.9, 0.95, 0.99$, and calculate the proportion of prediction intervals that contain the observed GHI for each site. The prediction intervals are computed using 5000 Monte Carlo samples. Table 3 shows the mean coverage for each level. Standard deviations of the root mean squared error from the target coverage are computed over subsamples, and presented in parentheses. It can be seen that laGP performs the best, but the performance of multilevel models is more robust than site-by-site and global models.

When predictive distribution is available, the Continuous Rank Probability Score [CRPS, Hersbach (2000), Krüger et al. (2016)] can be used for verification metric. For a realization y_{hij}, the estimated cumulative distribution function F_{hij} is available from $p(y_{hij} \mid \mathbf{x}_{hij}; \hat{\boldsymbol{\theta}}_{hi}, \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}})$ in (3.15). CRPS of F_{hij} is defined

TABLE 3
Coverage results of the four methods with standard deviation of the root mean squared error from the target coverage in parenthesis

	Confidence level			
Method	**0.80**	**0.90**	**0.95**	**0.99**
Multilevel	0.788 (0.019)	0.886 (0.018)	0.935 (0.017)	0.978 (0.013)
Site	0.766 (0.069)	0.866 (0.062)	0.916 (0.054)	0.964 (0.034)
Global	0.846 (0.046)	0.914 (0.015)	0.947 (0.004)	0.978 (0.012)
laGP	0.801 (0.007)	0.897 (0.006)	0.943 (0.008)	0.983 (0.007)

TABLE 4
CRPS comparison of the different modeling methods, divided by the size of the training sample and overall

	Training sample size		
Method	**<200**	**≥200**	**Overall**
Multilevel	0.346 (0.015)	0.345 (0.005)	0.345 (0.005)
Site	0.450 (0.181)	0.333 (0.049)	0.337 (0.054)
Global	0.341 (0.012)	0.350 (0.005)	0.349 (0.005)
laGP	0.260 (0.011)	0.291 (0.006)	0.290 (0.006)

by

$$\text{CRPS}(F_{hij}, y_{hij}) = \int_{\mathbb{R}} \{F_{hij}(z) - \mathbf{1}(z \geq y_{hij})\}^2 \, dz. \tag{4.1}$$

We can compare the performance of the different methods by calculating (4.1) for each of them. Table 4 presents the overall summary statistics for CRPS of each method. The standard deviation calculated over the subsamples are in parenthesis. Overall, the laGP stands out, followed by the site-by-site and multilevel model in aggregated results. Similar to the RMSPE, the site-by-site model suffers more from the data reliability issues. Also, the variation is much larger for the site-by-site model as seen in the standard deviation.

The prediction at an unobserved site also can be of interest depending on the application. When a prediction for a new site is needed, the site level model is not able to make the prediction. As such, the site level model is excluded in the comparison. The results are summarized in Table 5. The standard deviation in parenthesis is calculated over 10 subsamples. Although multilevel may have some advantages over the other two methods, it is not apparently much different than the others given the large uncertainty. An interesting observation is that, laGP's prediction performance in this metric is not as good as other categories, which suggests that laGP heavily uses geographic information when choosing the subdesign.

Overall, the laGP model performs the best in the comparison—at the cost of more computational overhead. It exploits the data from the neighboring sites, and decides the subset to use when making prediction; which requires some level of data writing process. So it should be viewed as a reference accuracy assuming

TABLE 5
Root mean squared prediction error comparison for out-of-sample sites

	Multilevel	**Global**	**laGP**
RMSPE	0.619 (0.395)	0.639 (0.411)	0.676 (0.008)

that more resources are available rather than a comparator. Compared to the site-by-site level model, there are extra benefits of using the multilevel model besides improved accuracy. First, there is no clear way to make a prediction for an unobserved site from the site-by-site model, while the group level inference can be used from the multilevel approach. The global model approach allows a prediction for this case but at the cost of accuracy. Second, the site-by-site level model cannot incorporate non-varying site-specific information, such as geographic or climate information. Site specific information can be useful when different sites are compared.

Lastly, it is also beneficial to compare the computation between different methods. Since it is difficult to conduct a controlled experiment to compare methodologies in our setting due to many sources that affect the process, such as database access and writing processes, we provide some metric as a reference. As site-by-site and global models have little appeal, we focus on the comparison between laGP and multilevel models.

There are two major differences. First, laGP model requires the data be transferred to form a single file, while our multilevel model only transfers the summary version of the data. As such, in the case study, laGP transfers the total of 12.929 Megabyte, while multilevel transfers 0.113 Megabyte for both bottom-up and top-down direction. Second, computation for laGP model depends on the number of predictions. For making 1000 predictions, laGP takes 70 seconds with four OpenMP threads. On the contrary, estimation for multilevel model takes up to three seconds at level one and two, respectively, and 20 seconds for level three, while the prediction takes one second for each site. Computational demand of multilevel estimation and prediction depends on the number of sites and groups, while that of laGP depends on the number of total data points.

5. Conclusion.

With the advances in remote sensing and storage technology, data are now collected over automated monitoring networks at an unprecedented scale. A simple yet efficient modeling approach that can reliably handle such data is of great need.

In this paper, we have developed a general framework using a multilevel modeling approach, which utilizes monitoring data collected to manage a large-scale system. It is presented with a solar energy application, although it can be flexibly modified to incorporate the data structure or overall goal. The computation can be automated with deterministic criteria, and be easily distributed. It has been shown that the method can provide improved inference compared to naive approaches. Our methodology can also be extended to incorporate discrete measurements.

We would like to conclude with a remark on potential topics for future research. First, the proposed method is illustrated using a parametric regression model approach. Extension of the method to more general models is also possible, with a modification on the assumptions of the multilevel structure. Second, a data-driven clustering method that can establish a group structure combined with our multilevel model approach will have potential impact in many industrial applications. Third, we believe that, expectation propagation [Gelman et al. (2014)] is closely related to our approach and can be used widely in settings similar to ours. Developing an easy-to-implement framework for EP will have an impact in industrial applications.

APPENDIX

We present the details of parameter estimation and prediction under the model setup in Section 4.

A. Parameter estimation. Let $\boldsymbol{\zeta}_h = (\boldsymbol{\beta}_h, \boldsymbol{\Sigma}_h)$ and $\boldsymbol{\xi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. From (3.11), we may assume $\hat{\boldsymbol{\theta}}_{hi} | \boldsymbol{\theta}_{hi} \sim N(\boldsymbol{\theta}_{hi}, \boldsymbol{V}_{hi})$, where $\boldsymbol{V}_{hi} = \{\boldsymbol{I}(\boldsymbol{\theta}_{hi})\}^{-1} = (\hat{\nu}_{hi} + 3)/(\hat{\nu}_{hi} + 1)(\boldsymbol{X}_{hi}^{\top}\boldsymbol{X}_{hi})^{-1}\hat{\sigma}_{hi}^2$. For the details of the parameter estimation in level one, see Lange, Little and Taylor (1989). Then

$$\boldsymbol{\theta}_{hi} | \hat{\boldsymbol{\theta}}_{hi} \sim N(\boldsymbol{\theta}_{hi}^*, \boldsymbol{V}_{hi}^*), \tag{A.1}$$

where

$$\begin{aligned}\boldsymbol{\theta}_{hi}^* &= [\{\boldsymbol{V}(\hat{\boldsymbol{\theta}}_{hi})\}^{-1} + \boldsymbol{\Sigma}_h^{-1}]^{-1}[\{\boldsymbol{V}(\hat{\boldsymbol{\theta}}_{hi})\}^{-1}\hat{\boldsymbol{\theta}}_{hi} + \boldsymbol{\Sigma}_h^{-1}\hat{\boldsymbol{\beta}}_h],\\ \boldsymbol{V}_{hi}^* &= [\{\boldsymbol{V}(\hat{\boldsymbol{\theta}}_{hi})\}^{-1} + \boldsymbol{\Sigma}_h^{-1}]^{-1}.\end{aligned}$$

Hence, with the given $\boldsymbol{\xi}_h^{(t)}$, E-step of the level one model gives

$$\begin{aligned}\boldsymbol{Q}_{1,hi}^{(t)} &\triangleq \mathbb{E}\{\boldsymbol{\theta}_{hi} \mid \hat{\boldsymbol{\theta}}_{hi}; \hat{\boldsymbol{\beta}}_h^{(t)}, \hat{\boldsymbol{\Sigma}}_h^{(t)}\} = \boldsymbol{\theta}_{hi}^*(\hat{\boldsymbol{\beta}}_h^{(t)}, \hat{\boldsymbol{\Sigma}}_h^{(t)}),\\ \boldsymbol{Q}_{2,hi}^{(t)} &\triangleq \mathbb{E}\{\boldsymbol{\theta}_{hi}\boldsymbol{\theta}_{hi}^{\top} \mid \hat{\boldsymbol{\theta}}_h; \hat{\boldsymbol{\beta}}_h^{(t)}, \hat{\boldsymbol{\Sigma}}_h^{(t)}\} = \boldsymbol{V}_{hi}^*(\hat{\boldsymbol{\Sigma}}_h^{(t)}) + \boldsymbol{Q}_{1,hi}^{(t)}(\boldsymbol{Q}_{1,hi}^{(t)})^{\top}.\end{aligned}$$

Then the M-step is

$$\begin{aligned}\hat{\boldsymbol{\beta}}_h^{(t+1)} &= \frac{1}{m_h}\sum_{i=1}^{m_h}\boldsymbol{Q}_{1,hi}^{(t)},\\ \hat{\boldsymbol{\Sigma}}_h^{(t+1)} &= \frac{1}{m_h}\sum_{i=1}^{m_h}\boldsymbol{Q}_{2,hi}^{(t)} - \hat{\boldsymbol{\beta}}_h^{(t+1)}(\hat{\boldsymbol{\beta}}_h^{(t+1)})^{\top}.\end{aligned}$$

When the EM algorithm of the level two model converges, we proceed to the level three. Let $\boldsymbol{\psi} = (\boldsymbol{\mu}, \tau, \rho)$ collectively denote the parameters associated with the group 3 model. From (3.11), we can derive

$$\hat{\boldsymbol{\beta}} \mid \boldsymbol{\beta} \sim N(\boldsymbol{\beta}, \boldsymbol{V}),$$

where $\boldsymbol{V}$ is a $H \times H$ is covariance matrix from the sampling distribution $g_2(\hat{\boldsymbol{\beta}} \mid \boldsymbol{\beta})$. With given $\boldsymbol{\psi}^{(t)}$,

$$(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}; \boldsymbol{\psi}^{(t)}) \sim N(\tilde{\boldsymbol{\beta}}^{(t)}, \tilde{\boldsymbol{V}}^{(t)}), \tag{A.2}$$

where $\tilde{\boldsymbol{\beta}}^{(t)} = \boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}} + (\boldsymbol{\Sigma}^{(t)})^{-1}\boldsymbol{F}\boldsymbol{\mu}^{(t)}$ and $\tilde{\boldsymbol{V}} = (\boldsymbol{V}^{-1} + (\boldsymbol{\Sigma}^{(t)})^{-1})^{-1}$. In the E step, the conditional expectation of the log likelihood with respect to $\boldsymbol{\psi}$ is

$$\begin{aligned}(\text{A.3})\quad &\mathbb{E}\{\log f_3(\boldsymbol{\psi}; \boldsymbol{\beta}) \mid \hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\psi}}^{(t)}\} \\ &\quad = -\{\operatorname{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\psi}}^{-1}\tilde{\boldsymbol{V}}) + (\tilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{F}\boldsymbol{\mu}^{(t)})^{\top}\boldsymbol{\Sigma}_{\boldsymbol{\psi}}^{-1}(\tilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{F}\boldsymbol{\mu}^{(t)} + \log|\boldsymbol{\Sigma}_{\boldsymbol{\psi}}|)\}/2.\end{aligned}$$

Then M step first finds the estimates for $\boldsymbol{\mu}$ and τ^2 in a closed form,

$$\begin{aligned}\hat{\boldsymbol{\mu}}^{(t+1)} &\leftarrow (\boldsymbol{F}^{\top}(\hat{\boldsymbol{\Sigma}}^{(t)})^{-1}\boldsymbol{F})^{-1}(\boldsymbol{F}^{\top}(\hat{\boldsymbol{\Sigma}}^{(t)})^{-1}\tilde{\boldsymbol{\beta}}^{(t+1)}), \\ \hat{\tau^2}^{(t+1)} &\leftarrow (\tilde{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{F}\hat{\boldsymbol{\mu}}^{(t+1)})^{\top}(\hat{\boldsymbol{\Sigma}}^{(t)})^{-1}(\tilde{\boldsymbol{\beta}}^{(t+1)} - \boldsymbol{F}\hat{\boldsymbol{\mu}}^{(t+1)})/H.\end{aligned}$$

Then the estimate for ρ can be found by minimizing

$$\begin{aligned}\hat{\rho}^{(t+1)} \leftarrow \arg\min_{\rho} &\operatorname{Tr}(\boldsymbol{\Sigma}_{\rho}^{-1}\tilde{\boldsymbol{V}}) + (\tilde{\boldsymbol{\beta}} - \boldsymbol{F}\hat{\boldsymbol{\mu}}^{(t+1)})^{\top}\boldsymbol{\Sigma}_{\rho}^{-1}(\tilde{\boldsymbol{\beta}} - \boldsymbol{F}\hat{\boldsymbol{\mu}}^{(t+1)})/\hat{\tau^2}^{(t+1)} \\ &+ \log|\boldsymbol{\Sigma}_{\rho}|.\end{aligned}$$

Then the E-M steps iterate until convergence. The spatial covariance parameter ρ is chosen by evaluating (A.3) with values over a pre-specified range and selecting the maximizing value.

B. Prediction. The best prediction for $\boldsymbol{\theta}_{hi}$ is $\mathbb{E}(\boldsymbol{\theta}_{hi}|\hat{\boldsymbol{\theta}}_{hi}, \hat{\boldsymbol{\zeta}}_h; \hat{\boldsymbol{\xi}})$ given by (3.15). From (3.8) and (A.1), the best prediction for $\boldsymbol{\theta}_{hi}$ is

$$\begin{aligned}&\mathbb{E}(\boldsymbol{\theta}_{hi}|\hat{\boldsymbol{\theta}}_{hi}, \hat{\boldsymbol{\zeta}}_h; \hat{\boldsymbol{\xi}}) \\ &\quad = \mathbb{E}(\mathbb{E}(\boldsymbol{\theta}_{hi}|\hat{\boldsymbol{\theta}}_{hi}, \boldsymbol{\zeta}_h)|\hat{\boldsymbol{\zeta}}_h; \hat{\boldsymbol{\xi}}) \\ &\quad = (\boldsymbol{V}_{hi}^{-1} + \hat{\boldsymbol{\Sigma}}_h^{-1})^{-1}(\boldsymbol{V}_{hi}^{-1}\hat{\boldsymbol{\theta}}_{hi} + \hat{\boldsymbol{\Sigma}}_h^{-1}\tilde{\boldsymbol{\beta}}_h),\end{aligned}$$

where $\hat{\boldsymbol{\Sigma}}_h$, $\hat{\boldsymbol{\Sigma}}$ are the MLE obtained from the EM algorithm, and $\tilde{\boldsymbol{\beta}}_h$ is from (A.2). The prediction interval can be obtained using Monte Carlo samples.

C. Simulation study. In this section we conduct a simulation study to show validity and robustness of the proposed method. The main purpose of the simulation is to check the unbiasedness of the proposed estimators. We consider two simulation setups: (1) estimation under the correct model specification; (2) estimation under a modest departure from the correct model specification. The first setup is to show the proposed method works well under the ideal situation, while the second setup is to show whether the proposed method is robust to departures from the assumptions in the model. Note that since too many possibilities exist for model violation, we only consider a simple case.

To generate the data, we consider a three-level model. For the first setup, the level three model is $\boldsymbol{\beta}_h = (\beta_{h1}, \beta_{h2}, \beta_{h3}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $p = 3$, where $\boldsymbol{\mu} =$

$(\mu_1, \mu_2, \mu_3) = (0, 0, 0)$, $\boldsymbol{\Sigma} = \boldsymbol{I}$. The level two model is $\boldsymbol{\theta}_{hi} = (\theta_{hi1}, \theta_{hi2}, \theta_{hi3}) \sim N(\boldsymbol{\beta}_h, \boldsymbol{\Sigma}_h)$ and level one model is $y_{hij} \sim N(\mathbf{x}_{hij}\boldsymbol{\theta}_{hi}, \sigma^2)$, with $\boldsymbol{\Sigma}_h = \boldsymbol{I}$ and $\sigma^2 = 1$. For the second setup, the level three model is $\boldsymbol{\beta}_h \sim t_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the same as the first setup and t_ν represents the multivariate t distribution with degrees of freedom ν. The level two model is $\boldsymbol{\theta}_{hi} \sim t_5(\boldsymbol{\beta}_h, \boldsymbol{\Sigma}_h)$ and the level one model is $y_{hij} \sim N(\mathbf{x}_{hij}\boldsymbol{\theta}_{hi}, \sigma^2)$, with $\boldsymbol{\Sigma}_h = \boldsymbol{I}$ and $\sigma^2 = 1$.

The sample sizes of each level are chosen to be $H = 20$ and $n_h = 40$, $n_{hi} = 700$ for all h and i; that is, 560,000 data points are generated in total for each simulation run, and simulation is replicated for 500 times. Specifically, for the rth simulation under the first setup, the level two parameter $\boldsymbol{\beta}_h^{(r)} = (\beta_{h1}^{(r)}, \beta_{h2}^{(r)}, \beta_{h3}^{(r)})$ is generated from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the hth group. Then given $\boldsymbol{\beta}_h^{(r)}$, the level three parameters $\boldsymbol{\theta}_{hi}^{(r)} = (\theta_{hi1}^{(r)}, \theta_{hi2}^{(r)}, \theta_{hi3}^{(r)})$ are generated from $N(\boldsymbol{\beta}_h^{(r)}, \boldsymbol{\Sigma}_h)$. Given $\boldsymbol{\theta}_{hi}^{(r)}$ and $\mathbf{x}_{hij}$, the level one observations y_{hij} can be generated from $N(\mathbf{x}_{hij}\boldsymbol{\theta}_{hi}^{(r)}, \sigma^2)$. The simulation under the second setup is generated similarly.

Results from the simulation for two setups are summarized in Tables 6–7. Although only the level three parameters are fixed, the generated "true" parameters can be traced over the simulation, and compared with the estimates. First, define the error of the best predictor $\hat{\theta}_{hi1}^{(r)}$ for $\theta_{hi1}^{(r)}$, the first coefficient of the ith sensor in the hth group, to be

$$\mathrm{e}_{hi1}^{(r)} = \hat{\theta}_{hi1}^{(r)} - \theta_{hi1}^{(r)}.$$

Then the mean bias from the 500 simulations is calculated by

$$\bar{e} = N_T^{-1} \sum_{r=1}^{500} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \mathrm{e}_{hi1}^{(r)}$$

with $N_T = 500 \times \sum_{h=1}^{H} n_h$, and the standard deviation is calculated by

$$\left(\sum_{r=1}^{500} \sum_{h=1}^{H} \sum_{i=1}^{n_h} (\mathrm{e}_{hi1}^{(r)} - \bar{e})^2 / (N_T - 1) \right)^{1/2}.$$

The summary statistics at the other levels can be calculated similarly, but with different number of elements at each level, because more samples are generated for the lower level parameters for one simulation run.

We can see that the proposed estimators possess unbiasedness from Table 6. Table 7 shows that the estimators are reasonably robust.

Acknowledgments. We thank two reviewers, the Associate Editor and the Editor for their valuable comments in improving this work. We also thank Drs. J. N. K. Rao and Julie Novak for their helpful suggestions.

TABLE 6
Mean bias and standard deviation under correct model specification

Level 3 parameter	μ_1	μ_2	μ_3
Mean	0.001	−0.011	0.001
sd	0.468	0.465	0.453
Level 2 parameter	β_{h1}	β_{h2}	β_{h3}
Mean	−0.012	0.001	0.001
sd	0.813	0.834	0.818
Level 1 parameter	θ_{hi1}	θ_{hi2}	θ_{hi3}
Mean	−0.012	0.003	0.001
sd	1.182	1.195	1.188

This report was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represented that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or any agency thereof.

TABLE 7
Mean bias and standard deviation under incorrect model specification

Level 3 parameter	μ_1	μ_2	μ_3
Mean	−0.002	0.010	0.006
sd	0.590	0.595	0.568
Level 2 parameter	β_{h1}	β_{h2}	β_{h3}
Mean	0.001	0.002	0.005
sd	1.067	1.072	1.094
Level 1 parameter	θ_{hi1}	θ_{hi2}	θ_{hi3}
Mean	0.003	−0.002	0.004
sd	1.546	1.543	1.572

REFERENCES

BATES, D. M. and PINHEIRO, J. C. (1998). Computational methods for multilevel modelling. Univ. Wisconsin, Madison, WI.

BAYARRI, M. J., BERGER, J. O., PAULO, R., SACKS, J., CAFEO, J. A., CAVENDISH, J., LIN, C.-H. and TU, J. (2007). A framework for validation of computer models. *Technometrics* **49** 138–154. MR2380530

CHATTERJEE, S., LAHIRI, P. and LI, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Ann. Statist.* **36** 1221–1245. MR2418655

CHEN, X. and XIE, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statist. Sinica* **24** 1655–1684. MR3308656

CHU, Y., PEDRO, H. T. C., NONNENMACHER, L., INMAN, R. H., LIAO, Z. and COIMBRA, C. F. M. (2014). A smart image-based cloud detection system for intrahour solar irradiance forecasts. *J. Atmos. Ocean. Technol.* **31** 1995–2007.

DENHOLM, P. and MARGOLIS, R. M. (2007). Evaluating the limits of solar photovoltaics (PV) in traditional electric power systems. *Energy Policy* **35** 2852–2861.

DU, J. and TRACTON, M. S. (2001). Implementation of a real-time shortrange ensemble forecasting system at NCEP: An update. In *Ninth Conference on Mesoscale Processes*. American Meteorological Society, Fort Lauderdale, FL.

ELA, E., MILLIGAN, M. and KIRBY, B. (2011). Operating reserves and variable generation. NREL/TP-5500-51978. Available at http://www.nrel.gov/docs/fy11osti/51978.pdf.

GELMAN, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics* **48** 432–435. MR2252307

GELMAN, A., VEHTARI, A., JYLÄNKI, P., ROBERT, C., CHOPIN, N. and CUNNINGHAM, J. P. (2014). Expectation propagation as a way of life. Preprint. Available at arXiv:1412.4869.

GOLDSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73** 43–56. MR0836433

GRAMACY, R. B. (2016). laGP: Large-scale spatial modeling via local approximate Gaussian processes in R. *J. Stat. Softw.* **72** 1–46.

GRAMACY, R. B. and APLEY, D. W. (2015). Local Gaussian process approximation for large computer experiments. *J. Comput. Graph. Statist.* **24** 561–578. MR3357395

GRAMACY, R. B. and HAALAND, B. (2016). Speeding up neighborhood search in local Gaussian process prediction. *Technometrics* **58** 294–303. MR3520659

GRAMACY, R. B., BINGHAM, D., HOLLOWAY, J. P., GROSSKOPF, M. J., KURANZ, C. C., RUTTER, E., TRANTHAM, M. and DRAKE, P. R. (2015). Calibrating a large computer experiment simulating radiative shock hydrodynamics. *Ann. Appl. Stat.* **9** 1141–1168. MR3418718

HALL, P. and MAITI, T. (2006). On parametric bootstrap methods for small area prediction. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 221–238. MR2188983

HAMMER, A., HEINEMANN, D., LORENZ, E. and LUCKEHE, B. (1999). Short-term forecasting of solar radiation: A statistical approach using satellite data. *Sol. Energy* **67** 139–150.

HERSBACH, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15** 559–570.

JIANG, H., SCHÖRGENDORFER, A., HWANG, Y. and AMEMIYA, Y. (2015). A practical approach to spatio-temporal analysis. *Statist. Sinica* **25** 369–384. MR3328820

KIM, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika* **98** 119–132. MR2804214

KLEIN, L. J., MARIANNO, F. J., ALBRECHT, C. M., FREITAG, M., LU, S., HINDS, N., SHAO, X., RODRIGUEZ, S. B. and HAMANN, H. F. (2015). PAIRS: A scalable geo-spatial data analytics platform. In 2015 *IEEE International Conference on Big Data* (*Big Data*) 1290–1298.

KRÜGER, F., LERCH, S., THORARINSDOTTIR, T. L. and GNEITING, T. (2016). Probabilistic forecasting and comparative model assessment based on Markov chain Monte Carlo output. Preprint. Available at arXiv:1608.06802.

LANGE, K. L., LITTLE, R. J. A. and TAYLOR, J. M. G. (1989). Robust statistical modeling using the *t* distribution. *J. Amer. Statist. Assoc.* **84** 881–896. MR1134486

LIU, F., BAYARRI, M. J. and BERGER, J. O. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Anal.* **4** 119–150. MR2486241

LIU, X., YEO, K., HWANG, Y., SINGH, J. and KALAGNANAM, J. (2016). A statistical modeling approach for air quality data based on physical dispersion processes and its application to ozone modeling. *Ann. Appl. Stat.* **10** 756–785. MR3528359

MARGOLIS, R., COGGESHALL, C. and ZUBOY, J. (2012). Integration of solar into the U.S. electric power system. In *SunShot Vision Study* U.S. Department of Energy, Washington, DC.

MARQUEZ, R. and COIMBRA, C. F. M. (2013). Intra-hour DNI forecasting based on cloud tracking image analysis. *Sol. Energy* **91** 327–336.

MATHIESEN, P., COLLIER, C. and KLEISSL, J. (2013). A high-resolution, cloud-assimilating numerical weather prediction model for solar irradiance forecasting. *Sol. Energy* **92** 47–61.

MATHIESEN, P. and KLEISSL, J. (2011). Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. *Sol. Energy* **85** 967–977.

ORWIG, K., AHLSTROM, M., BANUNARAYANAN, V., SHARP, J., WILCZAK, J., FREEDMAN, J., HAUPT, S., CLINE, J., BARTHOLOMY, O., HAMANN, H., HODGE, B., FINLEY, C., NAKAFUJI, D., PETERSON, J., MAGGIO, D. and MARQUIS, M. (2015). Recent trends in variable generation forecasting and its value to the power system. *IEEE Trans. Sustain. Energy* **6** 924–933.

PELLAND, S., GALANIS, G. and KALLOS, G. (2013). Solar and photovoltaic forecasting through post-processing of the global environmental multiscale numerical weather prediction model. *Prog. Photovolt.* **21** 284–296.

PEREZ, R., INEICHEN, P., MOORE, K., KMIECIK, M., CHAIN, C., GEORGE, R. and VIGNOLA, F. (2002). A new operational model for satellite-derived irradiances: Description and validation. *Sol. Energy* **73** 307–317.

PEREZ, R., LORENZ, E., PELLAND, S., BEAUHARNOIS, M., KNOWE, G. V., HEMKER JR., K., HEINEMANN, D., REMUND, J., MULLER, S. C., TRAUNMULLER, W., STEINMAUER, G., POZO, D., RUIZ-ARIAS, J. A., LARA-FANEGO, V., RAMIREZ-SANTIGOSA, L., GASTON-ROMERO, M. and POMARES, L. M. (2013). Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. *Sol. Energy* **94** 305–326.

PRATOLA, M. T., CHIPMAN, H. A., GATTIKER, J. R., HIGDON, D. M., MCCULLOCH, R. and RUST, W. N. (2014). Parallel Bayesian additive regression trees. *J. Comput. Graph. Statist.* **23** 830–852. MR3224658

QIAN, P. Z. G. and WU, C. F. J. (2008). Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics* **50** 192–204. MR2439878

SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer, New York. MR2160708

SCOTT, S. L., BLOCKER, A. W., BONASSI, F. V., CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. *Int. J. Manag. Sci. Eng. Manag.* **11** 78–88.

SKAMAROCK, W. C., KLEMP, J. B., DUDHIA, J., GILL, D. O., BARKER, D. M., DUDA, M. G., HUANG, X.-Y., WANG, W. and POWERS, J. G. (2008). A description of the advanced research WRF version 3. NCAR Technical Note: NCAR/TN-475+STR, National Center for Atmospheric Research, Boulder, CO.

SOTO, W. D., KLEIN, S. A. and BECKMAN, W. A. (2006). Improvement and validation of a model for photovoltaic array performance. *Sol. Energy* **80** 78–88.

WELCH, W. J., BUCH, R. J., SACKS, J., WYNN, H. P., MITCHELL, T. J. and MORRIS, M. D. (1992). Screening, predicting and computer experiments. *Technometrics* **34** 15–25.

WONG, G. Y. and MASON, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *J. Amer. Statist. Assoc.* **80** 513–524.

WONG, R. K. W., STORLIE, C. B. and LEE, T. C. M. (2017). A frequentist approach to computer model calibration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 635–648. MR3611763

WU, C. F. J. (2015). Post-Fisherian experimentation: From physical to virtual. *J. Amer. Statist. Assoc.* **110** 612–620. MR3367251

ZHANG, J., HODGE, B.-M., LU, S., HAMANN, H. F., LEHMAN, B., SIMMONS, J., CAMPOS, E., BANUNARAYANAN, V., BLACK, J. and TEDESCO, J. (2015a). Baseline and target values for regional and point PV power forecasts: Toward improved solar forecasting. *Sol. Energy* **122** 804–819.

ZHANG, J., FLORITA, A., HODGE, B.-M., LU, S., HAMANN, H. F., BANUNARAYANAN, V. and BROCKWAY, A. M. (2015b). A suite of metrics for assessing the performance of solar power forecasting. *Sol. Energy* **111** 157–175.

Y. HWANG
DEPARTMENT OF STATISTICS
SUNGKYUNKWAN UNIVERSITY
SEOUL 03063
KOREA
E-MAIL: yhwang@skku.edu

S. LU
IBM THOMAS J. WATSON RESEARCH CENTER
YORKTOWN HEIGHTS, NEW YORK 10598
USA
E-MAIL: lus@us.ibm.com

J.-K. KIM
DEPARTMENT OF STATISTICS
IOWA STATE UNIVERSITY
AMES, IOWA 50011
USA
AND
DEPARTMENT OF MATHEMATICAL SCIENCE
KAIST
DAEJEON 34141
KOREA
E-MAIL: jkim@iastate.edu