Chapter 1
Bottom-Up Estimation and Top-Down Prediction: Solar Energy Prediction Combining Information from Multiple Sources

Youngdeok Hwang, Siyuan Lu and Jae-Kwang Kim

Abstract 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 consider a new approach of analyzing large-scale multilevel models for computational efficiency. 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 US Department of Energy's SunShot Initiative.

1.1 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 Contiguous United States by 2030, and 27% by 2050,

Y. Hwang (✉)
Department of Statistics, Sungkyunkwan University, Seoul, Korea
e-mail: yhwang@skku.edu

S. Lu
IBM Thomas. J. Watson Research Center, Yorktown Heights, NY, USA
e-mail: lus@us.ibm.com

J.-K. Kim
Department of Statistics, Iowa State University, Ames, IA, USA
e-mail: jkim@iastate.edu

D. Choi et al. (eds.), *Proceedings of the Pacific Rim Statistical Conference for Production Engineering*, ICSA Book Series in Statistics,
https://doi.org/10.1007/978-981-10-8168-2_1

respectively (Margolis et al. 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 et al. 2011). Moreover, the uncertainty of solar power often obliges 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. 2015; Zhang et al. 2015). Applying statistical methods on the forecasts from these numerical models can significantly improve the forecasting accuracy (Mathiesen and Kleissl 2011; Pelland et al. 2013).

Computer models have advanced beyond scientific research to become an essential part of industrial applications. Such expansions need a different methodological focus. To take advantage of the availability of such computer models, matching the model output with the historical observations is essential. This task is closely related to model calibration (Gramacy et al. 2015; Wong et al. 2016) to choose the optimal parameters for the computer model.

In this work, we consider a general 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 advantage over the existing Bayesian method in computation for parameter estimation, because it does not rely on the Markov chain Monte Carlo (MCMC) method. Our approach is a frequentist based on the Expectation-Maximization (EM) algorithm.

1.2 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. GHI forecast is of main interest of the participants in the electricity market.

To monitor the GHI, sensors are located over CONUS. The collected observations are obtained from the sensor locations marked on Fig. 1.1. The GHI readings are recorded at 1,528 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 are separately stored in the database indexed by the site location. The readings are obtained from various kinds of sensors, which may cause some potential variability among different locations. In our application, we consider two models to

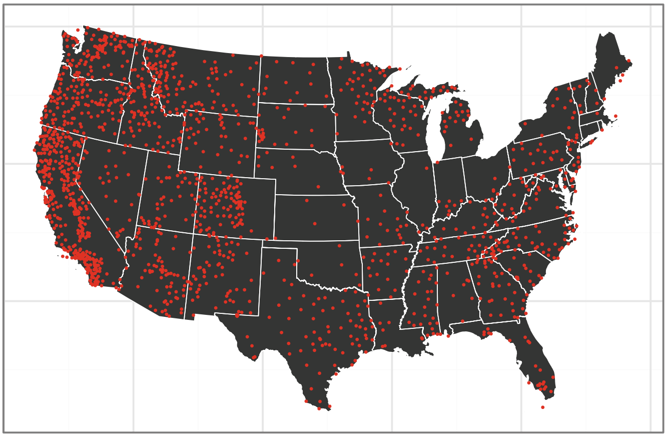

Fig. 1.1 The map of the 1,528 monitoring network locations, marked by dots

forecast GHI: Short-Range Ensemble Forecast (SREF, Du and Tracton 2001) and North American Mesoscale Forecast System (NAM, Skamarock et al. 2008). 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 model output is stored at every hour, but can be matched with 15-min interval measurement data after post-processing.

1.3 Model

In this section, we present the basic setup and our proposed method. A model with three levels is considered in this paper, but the number of levels can be arbitrary.

1.3.1 Multilevel Model

Assume that the sensors are divided into H exhaustive and non-overlapping groups. For group h, measurements are collected at n_h sensors. From the ith sensor in group h, the measurements y_{hij} are available, as well as the output from computer models as the covariates $\mathbf{x}_{hij}$, for $j = 1, \ldots, n_{hi}$. Information at sensor or group level, $\mathbf{c}_h$ and $\mathbf{c}_{hi}$, is also available. Note that the covariates $\mathbf{x}$ are often more widely available than y_{hij}'s; in our application in Sect. 1.4, 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

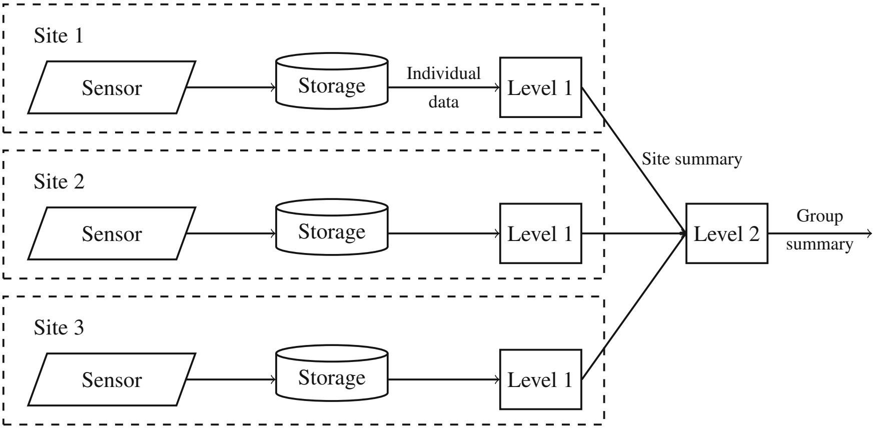

Fig. 1.2 Overall description of the data storage and modeling structure, where the data are stored separately for each site

the existing sensors and taking additional measurements from them usually do not cost much, while deploying new monitoring sensors often causes considerable cost.

Figure 1.2 shows the overall data storage and modeling structure of our proposed method to achieve these goals. Our so-called bottom-up 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. 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.

1.3.2 Bottom-Up Estimation

In this section, we give a detailed description of the estimation procedure. First, consider the level one and level two models,

$$\mathbf{y}_{hi} \sim f_1(\mathbf{y}_{hi}|\mathbf{x}_{hi};\boldsymbol{\theta}_{hi}), \tag{1.1}$$

$$\boldsymbol{\theta}_{hi} \sim f_2(\boldsymbol{\theta}_{hi}|\boldsymbol{c}_{hi};\boldsymbol{\zeta}_h), \tag{1.2}$$

where $\mathbf{y}_{hi} = (y_{hi1}, \ldots, y_{hin_{hi}})^\top$ and $\mathbf{x}_{hi} = (\mathbf{x}_{hi1}^\top, \ldots, \mathbf{x}_{hin_{hi}}^\top)^\top$ 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 (1.2), $\boldsymbol{\theta}_{hi}$ 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 (1.2), we use the three-step approach discussed in Sect. 2.1. 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 (1.2) 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

$$\sum_{i}^{n_h} \log \int g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi}) f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h) d\boldsymbol{\theta}_{hi} \tag{1.3}$$

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 (1.3) can be obtained by

$$\hat{\boldsymbol{\zeta}}_h = \arg\max_{\boldsymbol{\zeta}_h} \sum_{i=1}^{n_h} \mathbb{E}\left[\log\{f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)\} \mid \hat{\boldsymbol{\theta}}_{hi}; \boldsymbol{\zeta}_h\right]. \tag{1.4}$$

Note that $\boldsymbol{\zeta}_h$ is the parameter associated with the level two distribution, and (1.4) aggregates the information associated with $\hat{\boldsymbol{\theta}}_{hi}$ to estimate $\boldsymbol{\zeta}_h$.

To evaluate the conditional expectation in (1.4), we derive

$$p_2(\boldsymbol{\theta}_{hi} \mid \hat{\boldsymbol{\theta}}_{hi}; \boldsymbol{\zeta}_h) = \frac{g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi}) f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)}{\int g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi}) f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h) d\boldsymbol{\theta}_{hi}}. \tag{1.5}$$

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

$$\hat{\boldsymbol{\zeta}}_h^{(t)} = \arg\max_{\boldsymbol{\zeta}_h} \sum_{i=1}^{n_h} \mathbb{E}\left[\log\{f_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{c}_{hi}; \boldsymbol{\zeta}_h)\} \mid \hat{\boldsymbol{\theta}}_{hi}; \boldsymbol{\zeta}_h = \hat{\boldsymbol{\zeta}}_h^{(t-1)}\right], \tag{1.6}$$

where the conditional expectation is with respect to the prediction model in (1.5) 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 $g_1(\hat{\boldsymbol{\theta}}_{hi} \mid \boldsymbol{\theta}_{hi})$. Asymptotically, $\hat{\boldsymbol{\theta}}_{hi}$ is a sufficient statistic for $\boldsymbol{\theta}_{hi}$ and normally distributed with mean $\boldsymbol{\theta}_{hi}$ and the estimated variance $\{I_{1hi}(\boldsymbol{\theta}_{hi})\}^{-1}$, where $\{I_{1hi}(\boldsymbol{\theta}_{hi})\}^{-1}$ is the observed Fisher information derived from g_1.

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

$$\boldsymbol{\zeta}_h \sim f_3\left(\boldsymbol{\zeta}_h | \boldsymbol{c}_h; \boldsymbol{\xi}\right), \tag{1.7}$$

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 (1.3), we maximize

$$\sum_{h=1}^{H} \log \int g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h) f_3\left(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \boldsymbol{\xi}\right) d\boldsymbol{\zeta}_h \tag{1.8}$$

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

$$\hat{\boldsymbol{\xi}}^{(t)} = \arg\max_{\boldsymbol{\xi}} \sum_{h=1}^{H} \mathbb{E}\left[\log\left\{f_3\left(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \boldsymbol{\xi}\right)\right\} \mid \hat{\boldsymbol{\zeta}}_h; \boldsymbol{\xi} = \hat{\boldsymbol{\xi}}^{(t-1)}\right], \tag{1.9}$$

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

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

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.

1.3.3 Top-Down Prediction

In this section, we describe the prediction procedure. In contrast to the bottom-up approach of Sect. 1.3.2, the prediction is made in a top-down fashion.

To describe the top-down approach to prediction, consider the three-level models in (1.1), (1.2), and (1.7). The bottom-up estimation in Sect. 1.3.2 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

$$p(y_{hij} \mid \mathbf{x}_{hij}; \hat{\boldsymbol{\theta}}_{hi}, \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) = \frac{\int\int f_1(y_{hij} \mid \mathbf{x}_{hij}; \boldsymbol{\theta}_{hi}) p_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{\zeta}_h, \hat{\boldsymbol{\theta}}_{hi}, \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) p_3(\boldsymbol{\zeta}_h \mid \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) d\boldsymbol{\zeta}_h d\boldsymbol{\theta}_{hi}}{\int\int\int f_1(y_{hij} \mid \mathbf{x}_{hij}; \boldsymbol{\theta}_{hi}) p_2(\boldsymbol{\theta}_{hi} \mid \boldsymbol{\zeta}_{hi}, \hat{\boldsymbol{\theta}}_{hi}, \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) p_3(\boldsymbol{\zeta}_h \mid \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) d\boldsymbol{\zeta}_h d\boldsymbol{\theta}_{hi} dy_{hij}} \tag{1.10}$$

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

$$p_3(\boldsymbol{\zeta}_h \mid \hat{\boldsymbol{\zeta}}_h, \hat{\boldsymbol{\xi}}) = \frac{g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h) f_3(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \hat{\boldsymbol{\xi}})}{\int g_2(\hat{\boldsymbol{\zeta}}_h \mid \boldsymbol{\zeta}_h) f_3(\boldsymbol{\zeta}_h \mid \boldsymbol{c}_h; \hat{\boldsymbol{\xi}}) d\boldsymbol{\zeta}_h}, \tag{1.11}$$

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 (1.11), the predicted values of $\boldsymbol{\theta}_{hi}$ are generated by (1.5). The best prediction for y_{hij} is

$$\hat{y}_{hij}^* = \mathbb{E}_3 \left[\mathbb{E}_2 \left\{ \mathbb{E}_1(y_{hij} \mid \mathbf{x}_{hij}, \boldsymbol{\theta}_{hi}) \mid \hat{\boldsymbol{\theta}}_{hi}; \boldsymbol{\zeta}_h \right\} \mid \hat{\boldsymbol{\zeta}}_h; \hat{\boldsymbol{\xi}} \right] \tag{1.12}$$

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.

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

Once the bootstrap samples of $\mathbf{Y}^{*(b)} = \{y_{hij}^{*(b)}; h = 1, 2, \ldots, H; i = 1, \ldots, n_h; j = 1, \ldots, m_{hi}\}$ 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.

1.4 Prediction of Global Horizontal Irradiance

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

1.4.1 Data Description

We use 15 days of data for our analysis (12/01/2014–12/15/2014). There are 1528 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 are homogeneous. The number of sites in each group, n_h, varies between 10 and 59. Depending on the goal, one can use other grouping schemes such as the distribution zone described in (Zhang et al. 2015). Calculated irradiance is available at every 0.1 degree and is matched to the monitoring site location.

Since we are interested in the amount of irradiance, we first exclude zeros from both observed measurements and computer model outputs for the analysis. 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.

1.4.2 Model

This section presents the model that we used in the data analysis in detail. Let y_{hij} be the jth measurement for the ith sensor in the hth group. Following the multilevel modeling approach described in Sect. 1.3, we first assume that the measurement y_{hij} follows

$$y_{hij} = \mathbf{x}_{hij}\boldsymbol{\theta}_{hi} + e_{hij}, \tag{1.13}$$

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 et al. 1989).

The degrees of freedom are assumed to be five in the analysis, but it also can assumed to be unknown and estimated by the method of (Lange et al. 1989). Assume that the level two model follows

$$\boldsymbol{\theta}_{hi} \sim N(\boldsymbol{\beta}_h, \boldsymbol{\Sigma}_h), \tag{1.14}$$

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

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

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

$$\boldsymbol{\beta} \sim N(\boldsymbol{F}\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1.15}$$

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 Sect. 1.4.3, $\boldsymbol{F}$ is chosen to be $\mathbf{1}$, length H vector of 1's and a scalar μ. The spatial covariance $\boldsymbol{\Sigma}$ has its (k, l) th element

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

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. The estimated spatial effect for two coefficients is depicted in Fig. 1.3. 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 is associated with 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.

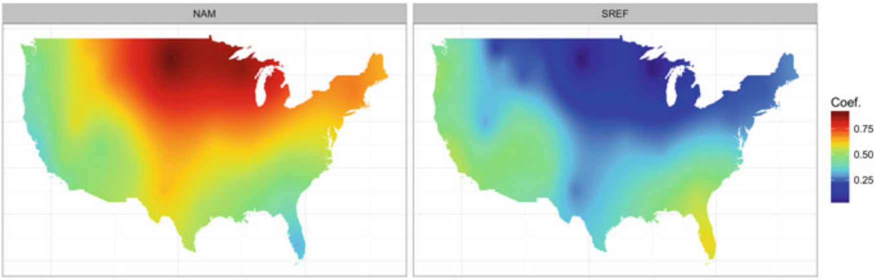

Fig. 1.3 Spatial variation of the group-level coefficients from the second level for two computer models, where the left panel shows the NAM model and the right panel the SREF model

1.4.3 Results

This section presents the data analysis result. Under the linear regression model in (1.13), the best prediction is $\hat{y}^*_{hij}$ in (1.12). We compared the multilevel approach with two other modeling 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. To evaluate the prediction accuracy, we conducted tenfold cross-validation. The data set 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.

We considered two scenarios: (a) prediction made at observed sites and (b) prediction made at new sites. For scenario (a), we partitioned the time point into ten subperiods, while for (b) the sites into ten subregions.

We compare 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 1.1 presents the overall summary statistics for the accuracy of each method, calculated from cross-validation. The standard deviation calculated over the subsamples is in parentheses.

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.

Table 1.1 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)

1.5 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.

Acknowledgements This report was prepared as an account of work sponsored by an agency of the US government. Neither the US 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 US government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the US government or any agency thereof.

References

Chatterjee, S., Lahiri, P., & Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *The Annals of Statistics*, *36*, 1221–1245. (06)

Denholm, P., & Margolis, R. M. (2007). Evaluating the limits of solar photovoltaics (pv) in traditional electric power systems. *Energy Policy*, *35*, 2852–2861.

Du, J., & Tracton, M. S. (2001). Implementation of a real-time shortrange ensemble forecasting system at ncep: An update. In *Ninth Conference on Mesoscale Processes*, Preprints, Ninth Conference on Mesoscale Processes, Fort Lauderdale, FL., American Meteorological Society.

Ela, E., Milligan, M., & Kirby, B. (2011). Operating Reserves and Variable Generation. NREL/TP-5500-51978. http://www.nrel.gov/docs/fy11osti/51978.pdf.

Gramacy, R. B., Bingham, D., Holloway, J. P., Grosskopf, M. J., Kuranz, C. C., Rutter, E., et al. (2015). Calibrating a large computer experiment simulating radiative shock hydrodynamics. *The Annals of Applied Statistics*, *9*(3), 1141–1168.

Hall, P., & Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*, 221–238.

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modeling using the T distribution. *Journal of the American Statistical Association*, *84*, 881–896.

Margolis, R., Coggeshall, C., & Zuboy, J. (2012). Integration of solar into the U.S. electric power system. In *SunShot vision study*. Washington, DC: U.S. Department of Energy.

Mathiesen, P., & Kleissl, J. (2011). Evaluation of numerical weather prediction for intra-day solar forecasting in the continental united states. *Solar Energy*, *85*, 967–977.

Orwig, K., Ahlstrom, M., Banunarayanan, V., Sharp, J., Wilczak, J., Freedman, J., et al. (2015). Recent trends in variable generation forecasting and its value to the power system. *IEEE Transactions on Sustainable Energy*, *99*, 1–10.

Pelland, S., Galanis, G., & Kallos, G. (2013). Solar and photovoltaic forecasting through post-processing of the global environmental multiscale numerical weather prediction model. *Progress in Photovoltaics: Research and Applications*, *21*, 284–296.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., et al. (2008). A description of the advanced research WRF version 3. NCAR TECHNICAL NOTE: NCAR/TNñ475+STR. National Center for Atmospheric Research, Boulder, Colorado, USA.

Wong, R. K. W., Storlie, C. B., & Lee, T. C. M. (2016). A frequentist approach to computer model calibration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. (To appear)

Zhang, J., Florita, A., Hodge, B.-M., Siyuan, L., Hamann, H. F., Banunarayanan, V., et al. (2015). A suite of metrics for assessing the performance of solar power forecasting. *Solar Energy*, *111*, 157–175.

Zhang, J., Hodge, B.-M., Siyuan, L., Hamann, H. F., Lehman, B., Simmons, J., et al. (2015). Baseline and target values for regional and point PV power forecasts: Toward improved solar forecasting. *Solar Energy*, *122*, 804–819.