Probabilistic Solar Power Forecasting Using Bayesian Model Averaging

Kate Doubleday, *Graduate Student Member, IEEE,* Stephen Jascourt, William Kleiber, and Bri-Mathias Hodge, *Senior Member, IEEE*

***Abstract*—There is rising interest in probabilistic forecasting to mitigate risks from solar power uncertainty, but the numerical weather prediction (NWP) ensembles readily available to system operators are often biased and underdispersed. We propose a Bayesian model averaging (BMA) post-processing method suitable for forecasting power from utility-scale photovoltaic (PV) plants at multiple time horizons up to at least the day-ahead timescale. BMA is a kernel dressing technique for NWP ensembles in which the forecast is a weighted sum of member-specific probability density functions. We tailor BMA for utility-scale PV forecasting by modeling power clipping at the AC inverter rating and advance the theory of BMA with a new beta kernel parameterization that accommodates theoretical constraints not previously addressed. BMA is demonstrated for a case study of 11 utility-scale PV plants in Texas, forecasting at hourly resolution for the complete year 2018. BMA's mixture-model approach mitigates underdispersion of the raw ensemble to significantly improve forecast calibration, while consistently outperforming an ensemble model output statistics (EMOS) parametric approach from the literature. At 4-hour lead time, the BMA post-processing achieves continuous ranked probability skill scores of 2–36% over the raw ensemble, with consistent performance at multiple lead times suitable for power system operations.**

***Index Terms*—solar power forecasting, probabilistic forecasting, Bayesian model averaging, beta distribution, solar power clipping**

NOMENCLATURE

Input Parameters and Sets

s	Solar plant index
$\mathcal{P}^{(s)}$	AC power rating of solar plant at site s
k	Ensemble member index
K	Number of ensemble members
t	Forecast time index
t_l	Forecast lead time
λ	Clipping threshold

K. Doubleday, W. Kleiber, and B.-M. Hodge are with the University of Colorado Boulder, Boulder, CO, 80309 USA. K. Doubleday and B.-M. Hodge are also with the National Renewable Energy Laboratory (NREL), Golden, CO 80401 USA. S. Jascourt is with Maxar, Gaithersburg, MD 20878 USA.

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement Number 33505 and the National Science Foundation under Award No. 1923062. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Forecast Variables and Function Definitions

$y_t^{(s)}$	Power at site s at time t
$f_{k,(t)}^{(s)}$	NWP forecast from member k for site s at time t
$h_{k,(t)}^{(s)}$	Conditional PDF from member k for site s at time t
p	Probability density function
P	Discrete probability
z	Dummy variable
ϕ	Standard normal PDF
Φ	Standard normal CDF
Γ	Gamma function
g_k	Member-specific beta or normal PDF
G_k	Member-specific beta or normal CDF
α, β	Standard beta distribution parameters
γ	Beta shape parameter
μ	Mean
σ	Standard deviation

Bayesian Model Averaging Parameters

a_{0k}, a_{1k}	Logistic regression coefficients for member k
b_k	Linear bias correction coefficient for member k
c_k	Variance height coefficient for member k
w_k	Weight of member k
τ_h	Sliding window width
τ_d	Time-of-day window width

Forecast Evaluation and Metrics

T	Evaluation period
F_t	Predictive CDF at time t
$\overline{\delta}$	Average interval width
ρ	Central interval
(w)$\overline{\text{CRPS}}$	(Weighted) mean Continuous Ranked Probability Score
(w)QS$_\xi$	(Weighted) Quantile Score at level ξ
ξ	Level in (0,1)
w_l, w_c, w_r	Left-, center-, and right-weighting functions
SS	CRPS skill score

Acronyms

NWP	Numerical weather prediction
BMA	Bayesian model averaging
EMOS	Ensemble model output statistics
CDF	Cumulative distribution function
PDF	Probability density function
SLI	Sliding
TOD	Time-of-day

I. Introduction

A. Motivation

As the penetration of renewable resources such as wind and solar photovoltaics (PV) increases in the power system, there is also a need for improved forecasting techniques [1]. Accurate and reliable forecasting can improve the utilization of variable and uncertain renewable generators, while mitigating the associated risks. Historically, efforts have focused on deterministic or "point" forecasts; however, there has been recent movement toward probabilistic forecasting to more fully capture forecast uncertainty [2], [3]. In operational practice, the state of the art is still to use deterministic forecasts, but a few system operators such as the Hawaiian Electric Company have begun to experiment with probabilistic forecasts [4]. As they become more widely available, probabilistic forecasts can be valuable for both power system operators and market participants [5], informing adaptive reserve algorithms [6], robust and/or stochastic unit commitment and economic dispatch models [7], [8], and market bidding strategies [9].

When considering different solar forecasting horizons, there are broadly two categories of techniques: statistical and time-series methods, which are generally applicable for "very short-term" forecasting (minutes to 6 hours ahead), and physics-based models, such as numerical weather prediction (NWP), which are more accurate for "short-term" hours- to days-ahead forecasting [2]. While recent literature has developed machine learning and time series methods for very short-term probabilistic solar forecasting [10]–[12], these methods have not yet translated into operations. We restrict our view to state-of-the-art short-term NWP-based approaches currently used by system operators for, e.g., the day-ahead unit commitment.

Post-processing NWP forecasts to develop probabilistic solar forecasts is a recent area of interest. One class of techniques post-processes a single deterministic NWP prediction to generate a probabilistic forecast [13]–[16], while a second group uses an ensemble of NWP forecasts, by collecting a variety of NWP models or perturbing their initial conditions [2], [17]. These "NWP ensembles" usually require post-processing to address weaknesses and to smooth the ensemble from a discrete set of points to a full cumulative distribution function (CDF). These weaknesses usually include a sunny bias and ensemble underdispersion—that is, a tendency to underestimate the uncertainty in the forecast [17]. This paper addresses this area of interest: post-processing NWP ensembles for solar power applications.

B. Background and Related Works

Recently, NWP ensemble post-processing methods from the meteorology and forecasting fields that address bias and underdispersion have been investigated for solar applications. [18] demonstrated two NWP post-processing techniques for solar power forecasting: variance deficit and ensemble model output statistics (EMOS) [19], which fits a parametric truncated normal distribution to the ensemble. EMOS is a state-of-the-art approach recently applied to wind speed [20] and electricity price [21] forecasting. [22] also investigated EMOS as well as Bayesian model averaging (BMA) for accumulated irradiance forecasting. BMA is a common approach from the meteorology field [23], which has been successfully applied to other weather variables, including precipitation [24], wind speed [25], and visibility [26], but has not been explored for the specific challenges of solar power forecasting.

BMA is a "kernel-dressing" method, in which each ensemble member is dressed with a probability density function (PDF) based on its historical performance. It is a relative of non-parametric methods like kernel density estimation (KDE), applied in [27], [28] for wind and [16], [29] for solar applications. In KDE, the overall probability distribution is the normalized sum of kernels (PDFs) centered at each data point. All kernels typically have the same bandwidth, but there is not a standard bandwidth selection method. Options range from rules-of-thumb to plug-in methods that require prior information [30].

BMA improves upon classic KDE by offering added customization. First, BMA includes a bias correction step, so that each kernel is not necessarily centered at the raw NWP data point. Additionally, the bandwidth and relative weight of each kernel are individually determined from that NWP member's historical performance, giving higher weight to more reliable members and ensuring the spread of uncertainty is adequately captured. With this added customization, BMA shapes and weights each ensemble member's kernel based on its historical performance to generate a mixture model that combats ensemble bias and underdispersion.

BMA is also distinguished from similarly named "Bayesian methods" [31], [32] that use historical observations as inputs to a time-series ARIMA model to fit a single parametric distribution, e.g., a beta [32]. In contrast, BMA uses physics-based NWP models, which are more accurate at longer time-horizons, as inputs to a mixture-model that cannot be described with a single parametric distribution.

Given that potential forecast users are system operators and plant owners, this method produces usable solar power forecasts, rather than irradiance forecasts. However, forecasting power from utility-scale PV plants does come with specific challenges. Training data quality can be a concern, due to plant maintenance and partial outages, as well as system conditions, like forced curtailment due to transmission constraints. Additionally, PV power output is determined by a plant's technical specifications including panel type, axis tracking configuration, and DC and AC power ratings. In particular, if the DC side is oversized compared to the AC inverter rating, the plant power might be "clipped" at its AC rating—nonlinear behavior that should be taken into account [33]. Besides common-sense data quality control, this paper addresses these challenges by directly incorporating clipping into its methods.

C. Contributions

To the best of the authors' knowledge, this is the first demonstration of BMA post-processing for NWP ensembles for solar power forecasting. Using a "raw" ensemble of NWP forecasts that have been individually preprocessed into hourly PV plant power, each ensemble member is dressed in a

two-part mixture model that explicitly accounts for clipping. In addition, we advance the field of BMA by addressing theoretical constraints on applying a beta kernel that were not addressed in previous literature [26]. Improvement from BMA post-processing is quantified with probabilistic metrics relative to three benchmarks: a persistence ensemble [15], the raw NWP ensemble, and a state-of-the-art parametric EMOS post-processed forecast from the literature [18]. The methods comparison is replicated for 11 utility-scale (∼5–100 MW) PV plants in Texas to illustrate the effectiveness over multiple locations and plant specifications, including thin film and regular silicon technologies and fixed, 1-axis, and 2-axis tracking configurations. At each site, BMA outperforms the three benchmarks at multiple lead times suitable for intra-day and day-ahead forecasting. These case studies use actual data for the complete year 2018 from power plants spanning hundreds of acres, rather than irradiance point measurements or kW-scale rooftop systems, as is common in the literature [10], [13], [32], [34].

D. Organization

The rest of the paper is organized as follows: Section II outlines the BMA post-processing model; Section III introduces how the model is fit to historical data; Section IV describes forecast benchmarks and metrics; Section V introduces the case study data; Section VI shows sensitivities on how to train the BMA models; Section VII presents the final post-processed forecast performance for 1 year; and Section VIII concludes.

II. BMA Post-Processing Method

In BMA, the PDF of the quantity of interest y_t^s for each location s at time t is determined as a mixture of conditional PDFs, $h_{k,t}^s(y_t^s|f_{k,t}^s)$, one for each forecast $f_{k,t}^s$ in an ensemble of K members. For brevity, the indices s and t are omitted. Each PDF $h_k(y|f_k)$ can be understood to be a PDF for y, conditional on member k. Based on that member's relative performance in the historical training period, each conditional PDF is assigned a nonnegative weight w_k, such that the sum of the weights is 1. The predictive PDF determined through BMA is then:

$$p(y|f_1,...,f_K) = \sum_{k=1}^{K} w_k h_k(y|f_k). \tag{1}$$

The selection of an appropriate kernel for $h_k(y|f_k)$ depends on the application. Popular choices include the Gaussian distribution for continuous variables [23] and the gamma or truncated normal distributions for non-negative quantities [25], [35]. In the context of solar power, the forecasted power should obey a lower bound of zero and an upper bound of the AC rating of the PV plant, $\mathcal{P}$. A doubly truncated normal kernel is one parsimonious option. A beta kernel is another flexible choice bounded on the interval [0,1]; power values can be easily translated onto this interval through normalization.

As an added complication, there could be a discrete probability that the plant is being clipped. When a PV plant has a DC power rating higher than its AC inverter rating, clipping

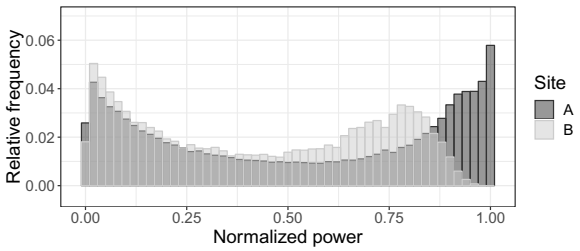


Fig. 1. Histogram of power output over 2+ years of plant operation, normalized by the AC power rating, for two sites used in this case study. Site A exhibits regular clipping at its AC power rating, but site B does not.

can be observed when the plant's output is restricted to the AC power rating. For utility-scale solar power plants, it is common to see DC-to-AC ratios of 1.2 or more [33]. For example, Fig. 1 illustrates the historical normalized power from two of the utility-scale plants used in this study, one of which exhibits regular clipping, but the other does not. Clipping can complicate probabilistic forecasting because it implies a high density or point mass in the PDF near the site's AC rating.

To explicitly handle clipping, we model each conditional PDF $h_k(y|f_k)$ as a discrete-continuous mixture with two parts. In the historical data, clipping results in some small power fluctuations slightly under the plant's AC rating, so clipping is qualified here by a threshold λ at 99.5% of the AC rating. The probability of clipping $P(y \geq \lambda\mathcal{P}|f_k)$ is estimated through logistic regression on the forecasted power:

$$\begin{aligned}\text{logit}\left(P(y \geq \lambda\mathcal{P}|f_k)\right) &\equiv \log\frac{P(y \geq \lambda\mathcal{P}|f_k)}{P(y < \lambda\mathcal{P}|f_k)} \\ &= a_{0k} + a_{1k}f_k\end{aligned} \tag{2}$$

Here, $P(y < \lambda\mathcal{P}|f_k)$ is the conditional probability that the solar power is not clipped, if f_k is the best ensemble member forecast at that time. In a forecast, this discrete component is re-distributed evenly over the top 0.5% of plant AC rating.

The second part of the mixture is a continuous kernel that models the amount of power, subject to not clipping. Both a beta and truncated normal kernel are considered here.

Since plant power is limited by the AC power rating, a truncated normal kernel can be defined on the interval $0 \leq z \leq \mathcal{P}$, using the PDF of the standard normal distribution, ϕ, and CDF, Φ:

$$p_\phi(z,\mu,\sigma) = \frac{\phi\left(\frac{z-\mu}{\sigma}\right)}{\sigma\left(\Phi\left(\frac{\mathcal{P}-\mu}{\sigma}\right) - \Phi\left(\frac{0-\mu}{\sigma}\right)\right)}. \tag{3}$$

Analogously, the PDF of a beta kernel with parameters $\alpha > 0$ and $\beta > 0$ for $0 \leq z \leq 1$ is given as: $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{(\alpha-1)}(1-z)^{(\beta-1)}$, where the mean is $\alpha/(\alpha+\beta)$, and the variance is $\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$. To allow a clearer interpretation of the parameters, we follow [26] by using the alternate formulation in [36]. By defining $\mu \equiv \alpha/(\alpha+\beta)$ for $0 < \mu < 1$ and $\gamma \equiv \alpha + \beta$ for $\gamma > 0$ (i.e., $\alpha = \mu\gamma$ and $\beta = \gamma(1-\mu)$), the beta PDF can be defined as:

$$p_\beta(z,\mu,\gamma) = \frac{\Gamma(\gamma)}{\Gamma(\mu\gamma)\Gamma\left(\gamma(1-\mu)\right)} z^{\mu\gamma-1}(1-z)^{(1-\mu)\gamma-1} \tag{4}$$

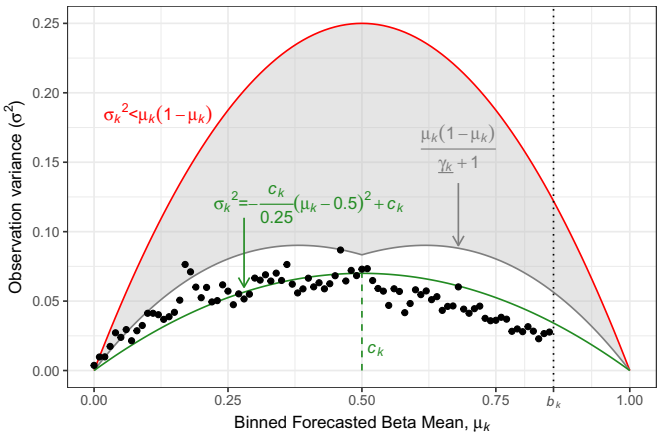


Fig. 2. For one ensemble member, trends in variance are illustrated by binning 3 months of estimated beta means into increments of 0.01; these fall within $[0, b_k]$ because of the linear correction in (5). The variance of the associated observations is modeled with a quadratic equation, shown in green. This model must obey the theoretical limit, shown in red. Additionally, the grey area indicates variances that result in ∪-shaped beta distributions.

These two parameters can be interpreted as a location parameter (μ) associated with the forecast and a shape parameter (γ) associated with its uncertainty.

The parameters μ and σ or γ for the two kernels are estimated similarly. The historical data suggest that the observed power has a linear relationship with the NWP forecast—note that the "raw" NWP forecast here is already in units of power, following the preprocessing described in the Appendix. After preprocessing, the power forecasts can still retain some sunny bias from the NWP model. The kernel mean is estimated from the forecast for member k through a scaling factor, b_k, which corrects this bias from the NWP forecast:

$$\mu_k = \begin{cases} b_k \frac{f_k}{\mathcal{P}}, & \text{if beta kernel} \\ b_k f_k, & \text{if truncated normal kernel} \end{cases} \tag{5}$$

A constant bias correction is not included, so the mean tends to zero as the forecast does. Previous work considered fitting to powers of the forecast, such as the square root for visibility [26] or cubed root for precipitation [24]. For this application, however, a linear relationship results in an acceptable fit.

The shape parameter, γ_k or σ_k, is determined by the distribution's standard deviation. Previous BMA implementations have estimated standard deviation through simple power relationships, such as $\sigma_k = c_{0k} + c_{1k} f_k^{1/2}$ [24]–[26]. When applying a beta kernel, however, the domain of the standard deviation is restricted, given the restricted domain of the distribution itself. In other words, to ensure α and β are positive, the variance must be limited to $\sigma^2 < \mu(1-\mu)$, a quadratic domain with a maximum value of 0.25, as shown in Fig. 2. An investigation of the historical data shows that the variance of the observations follows a quadratic trend within this domain, suggesting a quadratic model with height c_k:

$$\sigma_k^2 = -\frac{c_k}{0.25}\left(\mu_k - 0.5\right)^2 + c_k \tag{6}$$

A large c_k indicates a larger spread of uncertainty, while small c_k indicates high confidence. Although a similar discrete-beta model was proposed in [26] for visibility, the authors

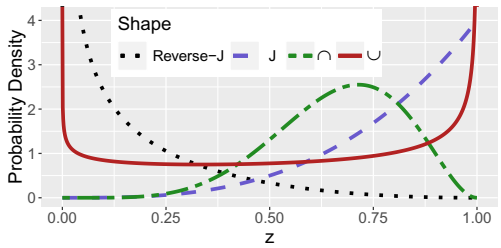


Fig. 3. Some typical beta distribution shapes.

suggested a linear relationship with the standard deviation, which could violate the theoretical limit on σ^2. Restricting the height parameter to $0 < c_k < 0.25$ in (6) resolves this issue and ensures the limit on σ^2 is satisfied for the beta kernel, while also generating a σ^2 estimate that is appropriate for the truncated normal distribution.

With the standard deviation model in (6), the truncated normal kernel can be calculated directly, while a few more steps are required for the beta kernel. The beta's shape parameter, γ_k, has a defined relationship with the standard deviation, $\sigma_k = \sqrt{\mu_k(1-\mu_k)/(\gamma_k+1)}$, which can be substituted into (6) and simplified to yield:

$$\gamma_k = \frac{0.25 - c_k}{c_k} \tag{7}$$

To further simplify and reduce computation time, the number of parameters is reduced by holding the height parameter constant across the ensemble members, as suggested in [24]. That is, $c_1 = \ldots = c_K = c$, and therefore $\gamma_1 = \ldots = \gamma_K = \gamma$.

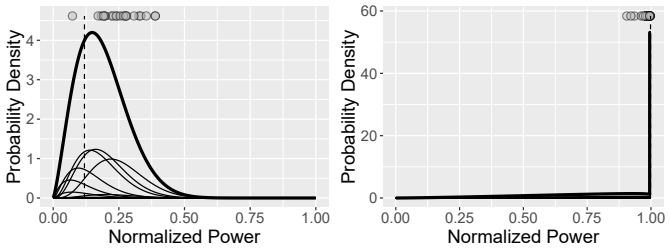
After the a, b, and c coefficients are fit as discussed in Section III, a final adjustment to γ is applied during the forecasting step. For different μ and γ values, the beta kernel can take various shapes, illustrated in Figure 3. In particular, "∪-shape" distributions emphasize the likelihoods of both zero and maximum power at the same time, which conflicts with forecaster intuition for a solar power application. ∪-shapes occur when $\alpha < 1$ and $\beta < 1$ and are the result of high variance (high c) in this model. A preliminary analysis indicated that this is not generally a major issue; a handful of the 11 sites had <0.5% ∪-shaped member forecasts, but one site had up to 7%. As mitigation, the beta variance estimate is truncated during forecasting to reduce ∪ distributions to J- or reverse-J-shapes. That is, (7) is truncated by a minimum value, $\underline{\gamma_k}$, based on the forecasted distribution mean:

$$\underline{\gamma_k} = \begin{cases} \frac{1}{1-\mu_k} & \text{if } \mu_k \le 0.5 \\ \frac{1}{\mu_k} & \text{if } \mu_k > 0.5 \end{cases} \tag{8}$$

This is illustrated in Fig. 2 by collapsing variance values that fall in the shaded grey area onto the line at its boundary.

With these elements, the conditional PDF from each ensemble member f_k, assuming f_k is the best forecast is:

$$\begin{aligned} h_k(y|f_k) = &\frac{P(y \ge \lambda\mathcal{P}|f_k)}{(1-\lambda)\mathcal{P}} \mathbb{1}[y \ge \lambda\mathcal{P}] \\ &+ \frac{P(y < \lambda\mathcal{P}|f_k)}{G_k\left(y|f_k\right)|_\lambda} g_k\left(y|f_k\right) \mathbb{1}[y < \lambda\mathcal{P}], \end{aligned} \tag{9}$$



(a) (b)

Fig. 4. Example BMA forecasts, (a) without and (b) with a high likelihood of clipping. The shaded circles show the NWP member forecasts, the thick line shows the BMA forecast, the thin lines show the component forecasts ($w_k h_k$), and the vertical dashed line shows the power at this instance.

where $\mathbb{1}$ is the indicator function and $g_k\left(y|f_k\right)$ is the member specific beta or truncated normal kernel:

$$g_k\left(y|f_k\right) = \begin{cases} \frac{p_\beta\left(\frac{y}{\mathcal{P}},\mu_k,\gamma_k\right)}{\mathcal{P}}, & \text{if beta kernel} \\ p_\phi\left(y,\mu_k,\sigma_k\right), & \text{if truncated normal} \end{cases} \tag{10}$$

$G_k(y|f_k)$ is the corresponding CDF—this minor scaling adjustment is added to counterbalance estimating the discrete component continuously over a nonzero width, $(1-\lambda)\mathcal{P}$. With these conditional member PDFs, the complete weighted model can be evaluated in (1).

The end result is illustrated in Fig. 4 for two forecast times from the case study, using a beta kernel. Fig. 4(a) shows a time when the majority of ensemble members overestimate power, which BMA addresses by shifting the bulk of the probability downwards. Fig. 4(b) shows a time when the discrete probability of clipping is high (27%), and the actual power was indeed clipped at 99.8% of the AC rating.

The complete BMA parameter fitting, forecasting, and evaluation process is summarized in Fig. 5, including key steps described further in Sections III, IV and VI.

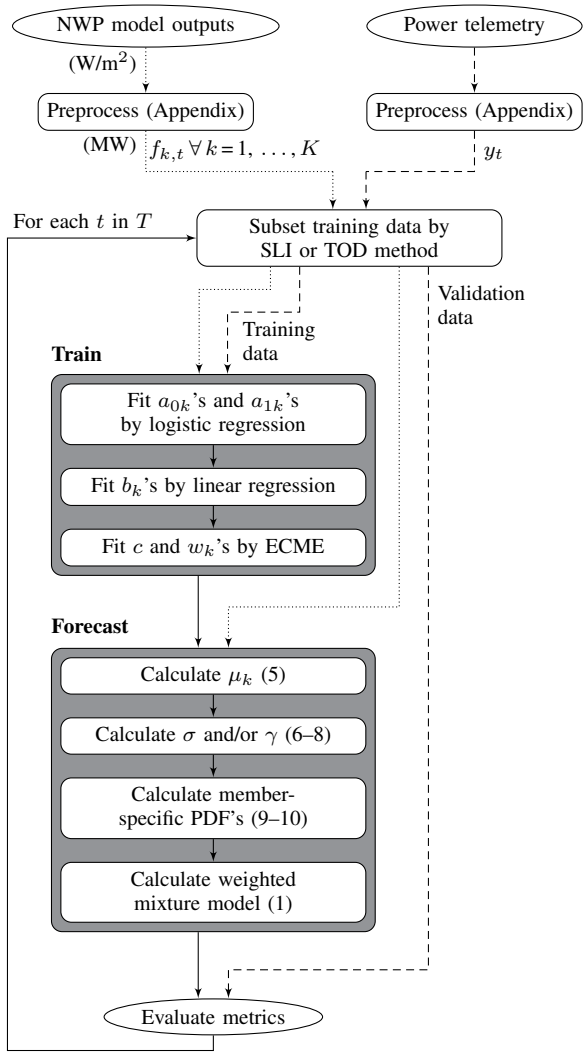


Fig. 5. Schematic of the parameter fitting, forecasting, and evaluation process.

III. BMA Parameter Fitting

Given a training data set for a given forecast, the a, b, c, and w parameters of the BMA model are estimated based on a previously published approach in [25]. The a coefficients in (2) are estimated by maximum likelihood of a logistic regression model. The member forecast is the predictor variable, whereas the dependent variable is the binary observation of clipping/no clipping. Because of minor deviations in the telemetry, clipping is determined by a threshold at 99.5% of the plant's AC power limit. This value was selected based on a review of the case study data, but it can be customized based on a plant's specifications. For this application, there is a high incidence of complete or quasi-complete separation in the logistic regression—that is, when the predictor at or above a constant is associated with only one of the binary outcomes. For example, training data for a sunrise forecast might contain no history of clipping, or training data for an afternoon forecast might show clipping only in some instances when the forecast was exactly at the plant's rated power. Therefore, two modifications are implemented: if there is no clipping in the training data, the discrete component is assumed to be 0, while in the general case, penalized logistic regression is used instead of basic logistic regression to handle instances of quasi-complete separation [37].

The linear bias correction slope b_k in (5) is determined by linear regression. Only time points in the training set when the power observations are not clipped (i.e., $y < 0.995 \times \mathcal{P}$) are used. The member forecast is the predictor, whereas the observed power is the dependent variable.

Finally, the member weights, w_k, and the variance coefficient c in (6) are found by maximum log-likelihood estimation, given the a and b coefficients found in the steps above. The maximum log-likelihood is estimated numerically using the Expectation Conditional Maximization Either (ECME) algorithm, which iterates between expectation (E) and conditional maximization (CM) steps [38]. Briefly, the algorithm introduces unobserved variables for each ensemble member, which can be interpreted as the probability of the member being the most skillful for the given time. In the E step, these latent variables are estimated from the current weight and variance coefficient estimates. In the first conditional maximization step CM-1, new estimates of the weights are developed from a maximization of the complete data log-

likelihood. This is followed by a second step, CM-2, in which the mixture log-likelihood is numerically maximized as a function of the variance coefficient c, given the estimated weights from CM-1. As suggested in [25], the CM-2 step is performed only once per 50 iterations of the E and CM-1 steps, which significantly reduces computation with very similar resulting parameter estimates. These steps are iterated until the changes in parameters are very small ($<10^{-5}$). For starting values, the members are weighted equally, though missing members are assigned 0 weight.

IV. Forecast Evaluation

To evaluate the improvement from the BMA method, the post-processed forecasts are compared to two benchmark forecasts using appropriate metrics, as introduced here.

A. Benchmark Forecasts

The proposed method is compared to three benchmark probabilistic forecasts: The first benchmark is a persistence ensemble (PeEn), a commonly used benchmark that only relies on historical observations [10], [11], [15], [18]. Following [15], a PeEn forecast for a given time is defined as the empirical CDF of the last 20 available measurements at the same hour of the day, which captures weather-related variations in solar power over the past three weeks. Comparing to a PeEn can illustrate the value of using NWP models that integrate more information than only the historical observations.

The second benchmark is the raw NWP ensemble itself, used to illustrate the added value from post-processing. The quantiles of both the PeEn and raw ensemble are interpolated using the "Classical" empirical CDF approach described in [39]. Third, an alternative NWP ensemble post-processing technique called ensemble model output statistics (EMOS) is implemented to benchmark BMA against a competing method [18]. Unlike BMA which results in a more complex mixture model, EMOS uses the NWP ensemble to fit a single parametric distribution. In [18], EMOS was applied to fit a non-negative, truncated normal distribution to solar irradiance. Here, we modify that method to fit a doubly-truncated normal distribution for solar power. The EMOS forecast is defined using the truncated normal equation in (3) as:

$$p(y|f_1,...,f_K) = p_\phi(y, a+b_1f_1+...+b_Kf_K, c+dS^2), \quad (11)$$

where $a + b_1f_1 + ... + b_Kf_K$ is the bias-corrected mean of the ensemble members and $c+dS^2$ is a linear function of the ensemble variance, S^2. The EMOS a, b, c, and d coefficients, which are distinct from the BMA a, b, and c coefficients, are fit by minimizing CRPS over the training data (Section VII) [18], using the robust Nelder-Mead algorithm in R's `optim` function and the truncated normal CRPS calculation from the `scoringRules` package.

B. Probabilistic Forecast Metrics

Several probabilistic metrics and diagnostic techniques are used here to compare forecast performance. A probabilistic forecaster intends to maximize forecast sharpness, subject to calibration [40]. Sharpness measures how concentrated the forecast is, whereas calibration is the statistical consistency between the forecasts and observations. Sharpness can be assessed over an evaluation period T by the average interval width, $\overline{\delta}$, of a central $(1-\rho)\times 100\%$ interval of interest [39]:

$$\overline{\delta} = \frac{1}{T}\sum_{t=1}^{T} F_t^{-1}\left(1-\frac{\rho}{2}\right) - F_t^{-1}\left(\frac{\rho}{2}\right), \quad (12)$$

where F_t is the forecast CDF at time t. Calibration can be assessed visually using a reliability diagram [39].

The Continuous Ranked Probability Score (CRPS) captures both sharpness and reliability in one metric [41]. Average CRPS ($\overline{\text{CRPS}}$) over period T can be decomposed as the integral of the Quantile Score (QS) over all quantiles [42]:

$$\overline{\text{CRPS}} = \int_0^1 \frac{1}{T}\sum_{t=1}^{T} \text{QS}_\xi\left(F_t^{-1}(\xi), y_t\right) d\xi, \quad (13)$$

where QS of the forecast $F_t^{-1}(\xi)$ at the level $\xi \in (0,1)$ is:

$$\text{QS}_\xi = 2\left(\mathbb{1}\left\{y_t \le F_t^{-1}(\xi)\right\} - \xi\right)\left(F_t^{-1}(\xi) - y_t\right) \quad (14)$$

$\overline{\text{CRPS}}$ can be weighted to focus on areas of interest within the distribution [42]. From a utility perspective, the lower tail of the distribution is of particular interest. The lower tail corresponds to times when solar power is uncommonly low, which is more likely to impact system reliability. To focus on different areas of the distribution, three quantile-weighting functions are applied [42]: a left tail weight function, $w_l(\xi) = (1-\xi)^2$; a center weight function, $w_c(\xi) = \xi(1-\xi)$; or a right tail function $w_r(\xi) = \xi^2$, for $\xi \in (0,1)$. A weighted QS of the form $\text{wQS}_\xi = w(\xi)\text{QS}_\xi$ can be evaluated and substituted into (13) to calculate a weighted average CRPS, $\overline{\text{wCRPS}}$.

Finally, to compare the forecast performance to a reference benchmark, a CRPS skill score can be evaluated:

$$\text{SS} = \frac{\overline{\text{CRPS}} - \overline{\text{CRPS}}_{ref}}{\overline{\text{CRPS}}_{ideal} - \overline{\text{CRPS}}_{ref}} = 1 - \frac{\overline{\text{CRPS}}}{\overline{\text{CRPS}}_{ref}}, \quad (15)$$

where $\overline{\text{CRPS}}_{ref}$ and $\overline{\text{CRPS}}_{ideal}$ (i.e., 0) are the average CRPS values for a reference and an ideal forecast, respectively [41]. A forecast with negative SS is worse than the reference, a SS of 0 is on par with the reference, and a SS of 1 is ideal.

V. Texas Case Study Data

Using the evaluation framework above, we demonstrate the value of BMA post-processing for a case study using actual historical forecasts and observation data. Historical 5-minute power data are gathered from 11 utility-scale PV plants in Texas, with ratings of ~5–100 MW. Two-plus years of data are used: November 2016 to December 31, 2017 data are available for training, and January 1, 2018 to December 31, 2018 is the test period. The Appendix details several data preprocessing steps, including quality control for suspect values, curtailment, and partial plant outages, and hourly averaging to match the available hourly forecasts. For data privacy concerns of the solar plant owners and to allow relative comparison among sites, power data and the forecast metrics are shown normalized by the plant rating, $\mathcal{P}$.

TABLE I
NUMERICAL WEATHER PREDICTION MODEL DATA USED

Model	Output Scale	Output Horizon	Output Interval	Frequency
NOAA GFS	0.25° (28 km)	16 days	3 hours	6 hours
NOAA NAM	3 km	60 hours	1 hour	6 hours
NOAA HRRR	3 km	18 hours	1 hour	1 hour
ECMWF HRES	0.125° (14 km)	10 days	3 hours	12 hours

For the corresponding historical forecasts, a base ensemble of four NWP models is used: the National Oceanic and Atmospheric Administration's (NOAA) Global Forecast System (GFS), NOAA's North American Mesoscale high-resolution nest (NAM), NOAA's High Resolution Rapid Refresh hourly (HRRR), and the European Centre for Medium-Range Weather Forecasts's (ECMWF) High Resolution (HRES) models [43], [44]. Table I gives details on each model used, including its output resolution, the forecast horizon, the output time interval of the forecast, and how frequently the forecast is issued. Each ensemble member is preprocessed to generate hourly-resolution forecasts of PV power from the weather variables; details are available in the Appendix. These ensemble members are the inputs for a BMA post-processed power forecast, at a given time t with lead time of t_l.

In addition to the base ensemble, a time-lagged ensemble of 21 members is considered by including the two previous model runs for each of the GFS, NAM, and HRES models, as well as the previous 11 runs of the HRRR model. For example, if the forecast includes a GFS run issued at time $t-1$ with a lead time of $t_l = 1$, the lagged ensemble would also include the GFS run issued at time $t-7$ with a lead time of $t_l = 7$ and the run issued at $t-13$ with a lead time of $t_l = 13$. In a similar vein to [45], this time-lagged alternative investigates the value of a more diverse ensemble that includes older forecast runs, which have different initial conditions.

VI. BMA TRAINING SENSITIVITIES

Before presenting the full results, this section presents a few sensitivities to tune the data inputs for this case study.

A. Training Data Selection

The last step in the BMA post-processing algorithm is to select training data to fit the BMA parameters and weights (Section III). For other weather variables, BMA models have typically been trained with a sliding window of data to capture recent weather conditions [24]–[26]. However, solar power also has known seasonal and diurnal trends, and other solar uncertainty models have been trained with data from similar seasons and/or times of day, e.g., [15].

To investigate these diurnal and weather impacts on model training, two data selection methods are explored: a "sliding" (SLI) window of τ_h hours (i.e., forecasts with the same lead time and resulting observations for times $t-\tau_h, ..., t-1$) or a "time-of-day" (TOD) window of τ_d days. The TOD window includes data for the same hour of the day from the past τ_d days, plus a centered window of $2\tau_d+1$ days around the same date in the previous year. Only time points in the training set when both the telemetry and forecast data are available are used, based on the quality control preprocessing.

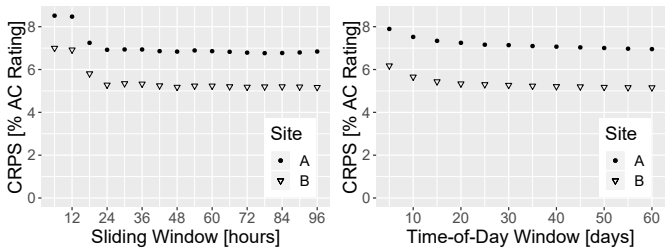


Fig. 6. Sensitivity of $\overline{\text{CRPS}}$ to (left) SLI and (right) TOD training windows widths used to select historical data for BMA coefficient fitting.

The SLI and TOD selection methods are compared through sensitivities on their window widths, τ_h and τ_d, respectively. A SLI window up to 4 days and a TOD window up to 60 days are considered. In the training step, the a, b, and c coefficients are fit using the SLI or TOD training data. In the forecasting step, the forecast of the power output 4 hours in the future is generated by BMA post-processing of the 4-hour-ahead NWP ensemble using those trained coefficients. This sensitivity is repeated for two of the case study sites (A and B), using the 21-member NWP ensemble at rolling 4-hour lead time over 2018 and a beta distribution for the BMA kernel. Figs. 6 and 7 illustrate the sensitivity of $\overline{\text{CRPS}}$ and central 90% interval width (i.e., sharpness) to the training data. First, small amounts of training data (i.e., low τ), typically result in sharp intervals but high $\overline{\text{CRPS}}$—that implies false sharpness at the expense of reliability. Using at least the past 24 hours of training data with the sliding window is enough achieve a flattening of $\overline{\text{CRPS}}$, though $\overline{\text{CRPS}}$ values show a slight 24-hour cycle with longer training windows. With the TOD window, $\overline{\text{CRPS}}$ tapers for windows extending back past 30 days.

While the $\overline{\text{CRPS}}$ values are similar in magnitude, the single metric obscures different forecast characteristics. When looking at the central 90% interval width, the SLI window is clearly sharper than the TOD window. Given the similar $\overline{\text{CRPS}}$, this implies the SLI approach is sharper, but the TOD window may be more reliable. There is a tendency for the TOD method to err on the side of over-dispersion and broader intervals, while the SLI method errs on the side of underdispersion and narrower intervals. One interpretation is that the sliding window's reliance on recent conditions results in a smaller standard deviation estimate and a tighter beta kernel. However, the TOD approach is likely better suited to estimating the clipping coefficients in (2). Additional sensitivities on the clipping threshold and training approach that best balance the continuous kernel and clipped components are left for future work. Hybrid schemes could also be considered. In the next two subsections, a 72-hour SLI window and 60-day TOD window are selected for further analysis. These windows achieve minimum $\overline{\text{CRPS}}$, though most of the benefits could likely be achieved with lower computation time using a 24-hour SLI or 30-day TOD window.

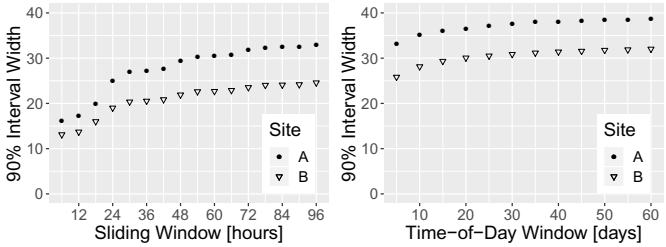


Fig. 7. Impact of (left) SLI and (right) TOD training window widths on central 90% interval width, as a percentage of plant AC rating.

TABLE II
$\overline{\text{CRPS}}$ (%$\mathcal{P}$) OF 4- VS. 21-MEMBER ENSEMBLE FORECASTS. EVALUATED BY ROLLING 4-HOUR AHEAD FORECAST OVER 2018.

	Site A			Site B		
Size	Raw	SLI	TOD	Raw	SLI	TOD
4	9.27	7.52	7.48	7.50	5.75	5.58
21	7.37	6.79	6.95	5.93	5.18	5.16
SS[1]	20.5%	9.78%	7.10%	20.8%	9.82%	7.50%

[1] SS of 21-member with 4-member ensemble as reference

B. Value of Time-Lagged NWP Members

Next, the performance of the base ensemble with 4 members is compared to the larger 21-member ensemble with time-lagged members. The comparison is conducted for the raw ensemble as well as BMA with the two training windows selected above to produce a 4-hour-ahead rolling forecast for sites A and B. Results in Table II show that the 21-member forecast significantly improves the raw ensemble $\overline{\text{CRPS}}$, with skill scores of ∼20% compared to the 4-member ensemble. The increased diversity of the larger ensemble improves the BMA post-processed approaches as well, with $\overline{\text{CRPS}}$ SS's of 7-10%. The larger ensemble also increases computation time commensurately, but the total time to train and post-process a single TOD BMA forecast takes on average 25-70 seconds on an Intel Xeon E5-2697 v4 2.30GHz processor, which is not prohibitive for an hour to day-ahead forecast. Therefore, the benefits from contributing members to the more diverse 21-member ensemble by reusing older forecasts is deemed worthwhile, and it is applied in all following analyses.

C. Beta vs. Truncated Normal Kernel

Finally, the performance of BMA with the beta kernel is compared to its performance with a truncated normal kernel. Similar to the sensitivities above, this sensitivity post-processes the 21-member ensemble to generate a rolling 4-hour-ahead forecast over 2018, replicated for sites A and B. The two BMA kernels (beta and truncated normal) are combined with the two training data windows from Section VI-A (72-hour SLI or 60-day TOD) to compare four variants on BMA post-processing.

This comparison investigates if the beta kernel's flexible shape lends any benefits over the truncated normal distribution, which is slightly simpler to implement. For instance, the beta kernel's "reverse-J" or "J" shape as the distribution nears one of the boundaries (0 or $\mathcal{P}$) might result in different tail behavior. To investigate these impacts on different areas of the forecast distribution, Table III shows the weighted $\overline{\text{CRPS}}$ skill scores for the four variants, shaded in order of performance with the best skill shaded the darkest. To calculate these skill scores, each $\overline{\text{CRPS}}$ value in (15) is replaced with $\overline{\text{wCRPS}}$ using the left, center, or right weighting functions, then compared to the unweighted original ($w = 1$).

TABLE III
WEIGHTED CRPS SS (%) OF 4 BMA VARIANTS COMPARED TO THE RAW NWP ENSEMBLE FOR A ROLLING 4-HOUR-AHEAD FORECAST OVER 2018. FOR THESE COMBINATIONS OF THE 2 KERNELS AND 2 TRAINING DATA SELECTION METHODS, β INDICATES A BETA KERNEL AND ϕ INDICATES A TRUNCATED NORMAL KERNEL.

	Site A				Site B			
	$w=1$	w_l	w_c	w_r	$w=1$	w_l	w_c	w_r
β-SLI	7.87	11.4	8.35	3.25	12.7	16.2	12.0	9.73
β-TOD	5.61	8.64	5.20	2.74	13.1	16.4	12.4	10.4
ϕ-SLI	6.59	10.8	6.84	1.47	12.0	15.8	11.2	8.83
ϕ-TOD	3.15	7.47	2.61	-1.00	-1.95	4.85	-2.32	-8.85

Table III confirms the more flexible beta kernel consistently outperforms the truncated normal kernel in all regions of the forecast distribution. The beta kernel achieves a modest 1%–4% increase in unweighted and left-, center-, and right-weighted CRPS skill scores over the truncated normal kernel for site A, and a 1%–19% increase for site B. These results also reinforce the importance of training data selection. With a truncated normal kernel, BMA outperforms the raw ensemble with a 12.0% skill score using SLI window training data, but it performs 1.95% worse with the TOD approach. The broadness of the truncated normal kernel exacerbates the tendency of TOD selection towards over-dispersion, worsening performance. Due to the consistent performance of BMA using the beta kernel and 72-hour SLI window training data (β-SLI), this variant is selected for the final methods comparison replicated for each of the 11 sites in the next section.

VII. BMA POST-PROCESSING PERFORMANCE

The performance of BMA post-processing is compared to the 3 benchmark methods for the validation year, 2018. The PeEn benchmark is compared to the three NWP-based methods using the 21-member ensemble: the raw ensemble benchmark; the EMOS post-processed benchmark, and the proposed BMA post-processed forecast. Based on the results in Section VI, the "BMA" method in this section refers to the β-SLI variant. For consistency, the 72-hour SLI training data selection is also used to train the EMOS model coefficients. For each of the four methods, forecasts are generated individually for each of the 11 case study PV plants in Texas at rolling lead times of 1, 4, 12, and 24 hours for 2018.

For a given site and lead time, each hourly probabilistic forecast is validated with a single observation of the site's average hourly power. For example, Fig. 8 compares the forecasts issued at 4-hour lead time over 2 days with mixed clouds to the actual power, demonstrating the differences among these four methods. Since both the PeEn and raw NWP ensemble estimate the forecast uncertainty from 20 or

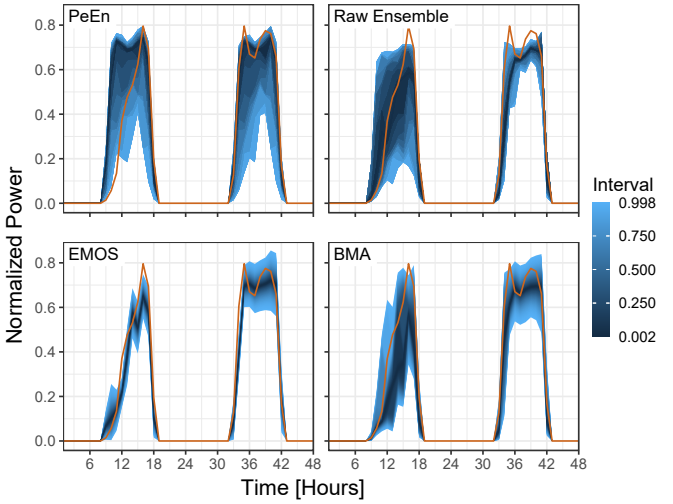


Fig. 8. Forecasts from the (clock-wise) PeEn, raw ensemble, EMOS, and BMA methods for a 1-axis tracking plant over two days with mixed cloudiness. The fan plot shows the intervals predicted 4 hours ahead, from 0.2% to 99.8% nominal coverage. The orange line shows the observed power.

21 discrete values, the resulting distribution is quite coarse. The PeEn forecast, which uses previous observations at the same time of day, generates a very broad and slowly changing forecast that nonetheless captures the diurnal trend in power. The raw ensemble shows large variance on the first cloudy day, while it is clustered and underdispersed on the second day, failing to provide adequate coverage for the peaks on both days. Recall that the "raw" ensemble here already accounts for the irradiance-to-power transform of this 1-axis tracking plant (see the Appendix). While the EMOS model adjusts for recent bias in the raw ensemble to improve coverage on the second day, it provides false confidence on the first day. EMOS's single parametric model does not fully capture the disagreement in the raw ensemble, resulting in a very sharp forecast with worse coverage compared to the raw ensemble. In contrast, BMA's mixture model captures more of the uncertainty in the raw ensemble, while providing sharper confidence intervals. Like EMOS, it also improves coverage on the second day.

To compare average performance of the rolling 4-hour ahead forecast over the entire validation year, Table IV compares $\overline{\text{CRPS}}$ for the four methods, replicated for each site. Each site has a different subset ($T < 8760$) of forecasts that can be validated for 2018 because of data quality control, ranging from ∼2200–4300 time-points (see the Appendix). Sites F–I, for example, have restricted data availability because of frequent curtailment. The best scores are in bold. First, note that the raw NWP ensemble alone has significant skill over the PeEn benchmark, with SS's of 14–45%; the post-processed forecasts have skill scores of 27–49% over the PeEn benchmark. Looking at the raw ensemble as the reference, post-processing with EMOS achieves skill scores of 28% for 2 sites, but skill scores ≤6% are more common, while four sites have negative skill scores, showing worse performance than the raw ensemble. In contrast, BMA improves over the raw ensemble for all sites and performs as well as or better than the EMOS technique for each site. Six sites show SS's

TABLE IV
$\overline{\text{CRPS}}$ & SS OF ROLLING 4-HOUR AHEAD FORECASTS OVER 2018, REPLICATED OVER THE 11 SITES. SS_{PeEn} IS SS WITH PEEN AS THE REFERENCE FORECAST; SS_{RAW} IS REFERENCED TO THE RAW ENSEMBLE.

Site	$\overline{\text{CRPS}}$ (%$\mathcal{P}$)				SS_{PeEn} (%)		SS_{raw} (%)	
	PeEn	Raw	EMOS	BMA	EMOS	BMA	EMOS	BMA
A	13.0	7.37	7.94	**6.79**	38.9	**47.8**	-7.78	**7.87**
B	9.55	5.93	5.61	**5.18**	41.3	**45.7**	5.47	**12.7**
C	9.74	8.40	**6.00**	**6.00**	**38.4**	**38.4**	**28.5**	**28.5**
D	10.2	6.74	6.70	**6.46**	34.4	**36.8**	0.52	**4.13**
E	13.5	7.53	8.03	**7.37**	40.7	**45.5**	-6.62	**2.11**
F	9.86	6.93	6.51	**6.09**	34.0	**38.2**	6.11	**12.2**
G	11.9	10.0	7.19	**6.39**	39.7	**46.4**	28.2	**36.3**
H	11.0	8.17	7.99	**6.96**	27.5	**36.8**	2.27	**14.8**
I	10.8	8.07	7.93	**6.90**	26.4	**35.8**	1.72	**14.4**
J	12.3	7.72	8.84	**7.47**	28.2	**39.3**	-14.5	**3.17**
K	12.6	6.91	7.34	**6.44**	41.9	**49.0**	-6.16	**6.89**

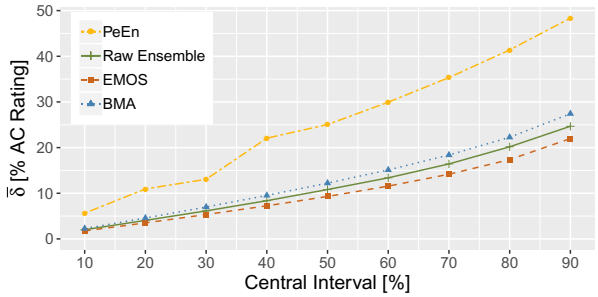


Fig. 9. Average widths of the 10% to 90% central intervals for site C.

of at least 12%, with one site achieving a 36% SS.

To investigate the tension between sharpness and calibration that can be obscured by an aggregate metric like $\overline{\text{CRPS}}$, Figs. 9 and 10 show the interval width and reliability diagrams for site C, which is the site where the EMOS and BMA techniques both achieved 28.5% skill scores over the raw ensemble. Consistent with the snapshot shown in Fig. 8, the PeEn benchmark has poor sharpness with consistently large intervals, while the intervals of the raw NWP ensemble are on average half the width. In contrast, Fig. 10 shows the PeEn benchmark has decent, though coarse, reliability, while the raw NWP ensemble is unreliable by regularly overestimating power, though this effect is exceptionally pronounced for site C. In fact, this site suffered from degradation of its 2-axis trackers over the evaluation period, which the data pre-processing did not adequately capture. Both post-processing techniques were able to correct for this changing behavior and improve the forecast calibration, while smoothing the stepped behavior of the raw ensemble. The EMOS forecast is somewhat sharper than the raw ensemble while the BMA forecast is somewhat broader. However, the EMOS benchmark errs on the side of sharpness, but this is sometimes a false sharpness that sacrifices reliability. The BMA forecast, in contrast, provides somewhat broader forecast intervals that provide better reliability, particularly on the lower tail of the distribution. While these methods result in the same $\overline{\text{CRPS}}$ for this site, this trend is consistent across the sites: by providing better coverage than the raw ensemble, BMA outperforms both it and the EMOS benchmark.

It is important to note that BMA does not evenly improve

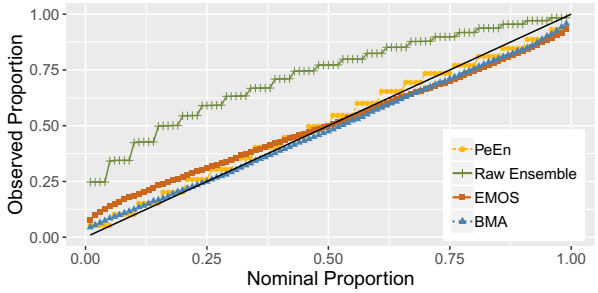


Fig. 10. Reliability diagram of the 1st to 99th forecast percentiles for site C. The black line shows ideal calibration.

TABLE V
WEIGHTED CRPS SS (%) OF EMOS OR BMA POST-PROCESSING COMPARED TO THE RAW ENSEMBLE AT ROLLING 4-HOUR LEAD-TIME.

	EMOS				BMA			
Site	$w=1$	w_l	w_c	w_r	$w=1$	w_l	w_c	w_r
A	-7.78	-3.40	-3.61	-18.4	7.87	11.4	8.35	3.25
B	5.47	7.62	7.27	0.77	12.7	16.2	12.0	9.73
C	28.5	36.9	29.0	16.6	28.5	38.7	27.6	16.2
D	0.52	4.32	2.32	-7.09	4.13	8.37	3.39	-0.37
E	-6.62	-5.07	-4.03	-11.8	2.11	5.70	1.92	-1.56
F	6.11	14.2	8.25	-7.91	12.2	20.7	11.9	1.01
G	28.2	39.6	28.8	9.49	36.3	47.0	34.6	21.8
H	2.27	0.45	7.02	-2.15	14.8	19.6	15.6	8.21
I	1.72	2.69	6.66	-6.30	14.4	21.2	15.5	4.83
J	-14.5	-6.00	-10.3	-31.5	3.17	8.30	2.47	-2.37
K	-6.16	-0.82	-2.38	-18.1	6.89	11.0	7.13	1.46

the forecasts: the lower tail benefits most. This effect is pronounced in Fig. 10 and is also evident in left-, center-, and right-weighted CRPS SS's. Across the sites, the left tail has the highest relative improvement with BMA, followed by the distribution center and then the right tail. Benefits from EMOS post-processing also skew towards the left tail or the distribution center, but for all sites, BMA's improvement of the left/lower tail out outweighs that from EMOS. Underestimation of lower-tail risk is concerning to system operators because it is likely to result in the highest cost and reliability impacts. BMA strongly improves tail risk estimation with left-weighted SS's of 6–47%, and improves the right tail estimation for most sites as well with right-weighted SS's as high as 22%.

To verify BMA improvements extend to other lead times as well, Table VI reports unweighted SS's over the raw NWP ensemble at four rolling lead times up to 24-hours ahead. In general, the SS's are maintained or even increased as the lead time increases, which is valuable as the number of available time-lagged members reduces from 21 to 14 for the 12-hour-ahead and 9 for the 24-hour-ahead forecast. BMA outperforms EMOS at all lead times for all sites, except for a few lead times at site C where the skill scores are very similar. The consistent performance demonstrates BMA's applicability at multiple time horizons, from the intra-day to the day-ahead.

VIII. SUMMARY AND CONCLUSIONS

This paper proposed a new Bayesian model averaging approach to post-process NWP ensemble estimates of utility-scale PV power, applicable for forecasting up to the day ahead timescales. This method uses a kernel-based mixture

TABLE VI
$\overline{\text{CRPS}}$ SS (%) OF THE BMA OR EMOS POST-PROCESSING METHODS OVER THE RAW ENSEMBLE AT VARYING LEAD-TIMES

	1-Hour 21 members		4-Hour 21 members		12-Hour 14 members		24-Hour 9 members	
Site	EMOS	BMA	EMOS	BMA	EMOS	BMA	EMOS	BMA
A	-12.8	**7.48**	-7.78	**7.87**	-7.81	**6.03**	0.33	**10.3**
B	5.15	**12.1**	5.47	**12.7**	8.82	**14.1**	12.0	**16.7**
C	27.8	**28.4**	**28.5**	**28.5**	**28.0**	26.2	**28.9**	28.1
D	1.93	**5.33**	0.52	**4.13**	-1.55	**1.84**	-0.79	**4.16**
E	-9.42	**0.68**	-6.62	**2.11**	-2.18	**2.56**	1.64	**5.54**
F	8.28	**14.1**	6.11	**12.2**	7.31	**12.2**	13.0	**17.3**
G	30.0	**38.4**	28.2	**36.3**	29.6	**35.7**	32.4	**35.8**
H	0.61	**14.8**	2.27	**14.8**	3.96	**14.8**	11.4	**18.5**
I	-1.47	**14.6**	1.72	**14.4**	1.96	**14.2**	10.8	**18.9**
J	-19.8	**3.36**	-14.5	**3.17**	-20.2	**3.77**	-16.9	**7.56**
K	-7.12	**5.92**	-6.16	**6.89**	-11.6	**3.57**	-5.20	**7.05**

model combining a discrete component for power clipped at the inverter rating and a continuous portion for lower output. To use beta kernels within the mixture model, a new parametrization was developed that accommodates the beta distribution's theoretical constraints. In developing this mixture model, a large variety of comparisons were examined to justify the selections made.

For a given forecast time, the parameters of the BMA model are trained using historical forecasts and observations. The length of this historical data training window was selected based on the asymptotic behavior shown in Figures 6 and 7. Training with the sliding window (SLI) approach showed overall slightly better results than with the time of day (TOD) approach (Table III), thus was used for all further comparisons. Including older NWP runs to make a time-lagged ensemble produced superior results to using only the latest runs (Table II). Using beta functions in the mixture model was superior to using normal distributions truncated on two sides (Table III).

Given those training and implementation selections, the BMA post-processing achieves skill scores of 2–36% over the raw ensemble. It generally improves forecast calibration while broadening the interval widths compared to the underdispersed raw ensemble, which is sharp but can be unreliable. The BMA mixture model was also demonstrably superior to the parametric EMOS post-processing method from the literature, which can sometimes sacrifice reliability by erring on the side of sharpness. (Tables IV, V, and VI). Skill benefits were shown for a variety of forecast time horizons relevant for power systems operations (Table VI). These comparisons were replicated for 11 utility-scale PV plants in Texas, demonstrating improvements over multiple locations and plant specifications.

The largest improvements demonstrated by the BMA mixture model were in the lower tail of the distribution (Table V), which is of greatest benefit to electric grid operators and solar plant managers. Underestimating the lower tail risk can result in high cost and potential power system reliability impacts.

Overall, the major findings of the article are:

- BMA post-processing contributed skill score improvements of up to 36% over the raw NWP ensemble.
- Using a beta kernel in the BMA mixture model was superior to using a doubly-truncated normal kernel.

- Skill score improvements were consistently shown over forecast time horizons of 1–24 hours relevant for power systems operations.
- Post-processing with BMA mixture models was superior to a state-of-the-art parametric EMOS treatment, by better capturing disagreements in the NWP ensemble.
- BMA contributed the largest improvement in the lower tail of the distribution, which is of greatest benefit due to the higher cost of managing situations of low output.

For future work, further sensitivities can investigate clipping behavior, including different irradiance-to-power preprocessing transforms, the clipping threshold, and whether multi-part kernels provide improvements over a single continuous kernel. Exploring additional hybrid training methods beyond the sliding and time-of-day windows could further improve performance by choosing the most appropriate data that capture both seasonal/diurnal cycles and different weather regimes. Future work will apply these improved probabilistic forecasts in power systems operational models, such economic dispatch and unit commitment models, to endogenously consider future generation risk through coherent risk measures such as conditional value-at-risk (CVaR) and entropic value-at-risk (EVaR).

APPENDIX
POWER TELEMETRY PREPROCESSING

The raw power data are screened for several data quality issues, including stuck loggers and nonsensical values (e.g., $< 0, > \mathcal{P}$). Suspect values are treated as missing. Times when the solar zenith angle indicates the sun is down are ignored. Also, times when all ensemble members forecast power less than 500 kW are ignored to remove spurious and erratic behavior caused by very small values around sunrise and sunset. This threshold is $< 10\%$ of the smallest plant size and $< 1\%$ of the largest plant size. Power data are also impacted by the operating conditions, including involuntary curtailment caused by transmission constraints and partial outages for maintenance. For known partial outages, the power is scaled to what it would have been without the outage. Curtailed data are treated as missing. Most sites have $< 10\%$ curtailment, but some are curtailed up to a third of the time.

Telemetry data are aggregated to hour-ending averages for consistency with the hourly-resolution NWP ensemble members. If any of the 12 5-minute values are missing, the entire hour is treated as missing. If the solar zenith angle indicates that the sun is down for part of the hour, it is assumed that those times contribute zero power.

NWP FORECAST PREPROCESSING

The hourly or three-hourly time-averaged Global Horizontal Irradiance (GHI) from each NWP model needs to be converted to hourly PV plant power in preparation for the BMA treatment. The preprocessing was done using the solar forecast system in [46]. It makes small statistical corrections to the shape and amplitude of the NWP model diurnal irradiance curve, then calculates clear sky irradiance every 1 minute assuming the time-averaged forecast clear sky index (CSI) from this corrected model is constant for the interval. Next, minute-by-minute clear sky direct and diffuse irradiance are calculated by prorating with this CSI and adding stochastic variations to account for the missed variability. These direct and diffuse irradiance data are transposed into plane-of-array irradiance for tilted sun-tracking PV panels, then used to empirically estimate power based on historical data. Finally, the data are averaged into hour-ending averages. This irradiance-to-power transform is similar to regression methods, such as that in [47]; a PV simulation tool like the System Advisor Model is an alternative [48]. Though not a "raw" output from an NWP model, we refer to these preprocessed, hourly PV plant power forecast members as the raw ensemble, which is then input into the BMA post-processing.

ACKNOWLEDGMENTS

The authors thank Christopher Cassidy of Maxar; Pengwei Du, Sandip Sharma, and Sean Chang of the Electric Reliability Council of Texas; and José Daniel Lara and Elina Spyrou of NREL for providing their data, insights, and suggestions.

REFERENCES

[1] B. Kroposki, B. Johnson, Y. Zhang, V. Gevorgian, P. Denholm, B.-M. Hodge, and B. Hannegan, "Achieving a 100% renewable grid: Operating electric power systems with extremely high levels of variable renewable energy," *IEEE Power and Energy Magazine*, vol. 15, no. 2, pp. 61–73, March 2017.

[2] D. van der Meer, J. Widén, and J. Munkhammar, "Review on probabilistic forecasting of photovoltaic power production and electricity consumption," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1484–1512, Jan. 2018.

[3] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond," *International Journal of Forecasting*, vol. 32, no. 3, pp. 896–913, 2016.

[4] D. Nakafuji and L. Gouveia, "Distributed resource energy analysis and management system (DREAMS) development for real-time grid operations," Hawaiian Electric Company, Honululu, HI (United States), Tech. Rep., 2016.

[5] R. J. Bessa, C. Möhrlen, V. Fundel, M. Siefert, J. Browell, S. Haglund El Gaidi, B.-M. Hodge, U. Cali, and G. Kariniotakis, "Towards improved understanding of the applicability of uncertainty forecasts in the electric power industry," *Energies*, vol. 10, no. 9, 2017.

[6] M. Bucksteeg, L. Niesen, and C. Weber, "Impacts of dynamic probabilistic reserve sizing techniques on reserve requirements and system costs," *IEEE Trans. Sustain. Energy*, vol. 7, no. 4, pp. 1408–1420, 2016.

[7] Á. Lorca and X. A. Sun, "Adaptive robust optimization with dynamic uncertainty sets for multi-period economic dispatch under significant wind," *IEEE Trans. Power Syst.*, vol. 30, no. 4, pp. 1702–1713, 2015.

[8] W. A. Bukhsh, C. Zhang, and P. Pinson, "An integrated multiperiod OPF model with demand response and renewable generation uncertainty," *IEEE Trans. Smart Grid*, vol. 7, no. 3, pp. 1495–1503, May 2016.

[9] C. J. Dent, J. W. Bialek, and B. F. Hobbs, "Opportunity cost bidding by wind generators in forward markets: Analytical results," *IEEE Trans. Power Syst.*, vol. 26, no. 3, pp. 1600–1608, Aug. 2011.

[10] M. David, F. Ramahatana, P. Trombe, and P. Lauret, "Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models," *Solar Energy*, vol. 133, pp. 55–72, 2016.

[11] H. T. Pedro, C. F. Coimbra, M. David, and P. Lauret, "Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts," *Renewable Energy*, vol. 123, pp. 191–203, 2018.

[12] J. Munkhammar, D. van der Meer, and J. Widén, "Probabilistic forecasting of high-resolution clear-sky index time-series using a markov-chain mixture distribution model," *Solar Energy*, vol. 184, pp. 688–695, 2019.

[13] E. Lorenz, J. Hurka, D. Heinemann, and H. G. Beyer, "Irradiance Forecasting for the Power Prediction of Grid-Connected Photovoltaic Systems," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 2, no. 1, pp. 2–10, March 2009.

[14] F. Golestaneh, H. B. Gooi, and P. Pinson, "Generation and evaluation of spacetime trajectories of photovoltaic power," *Appl. Energy*, vol. 176, pp. 80–91, Aug. 2016.
[15] S. Alessandrini, L. Delle Monache, S. Sperati, and G. Cervone, "An analog ensemble for short-term probabilistic solar power forecast," *Applied Energy*, vol. 157, pp. 95–110, 2015.
[16] Y. Zhang and J. Wang, "GEFCom2014 probabilistic solar power forecasting based on k-nearest neighbor and kernel density estimator," in *2015 IEEE Power Energy Society General Meeting*, July 2015, pp. 1–5.
[17] M. Leutbecher and T. Palmer, "Ensemble forecasting," *Journal of Computational Physics*, vol. 227, no. 7, pp. 3515–3539, 2008.
[18] S. Sperati, S. Alessandrini, and L. Delle Monache, "An application of the ECMWF ensemble prediction system for short-term solar power forecasting," *Solar Energy*, vol. 133, pp. 437–450, 2016.
[19] T. Gneiting, A. E. Raftery, A. H. Westveld, and T. Goldman, "Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation," *Monthly Weather Review*, vol. 133, no. 5, pp. 1098–1118, 2005.
[20] S. Baran and S. Lerch, "Mixture EMOS model for calibrating ensemble forecasts of wind speed," *Environmetrics*, vol. 27, no. 2, pp. 116–130, 2016.
[21] S. Chai, Z. Xu, and Y. Jia, "Conditional density forecast of electricity price based on ensemble ELM and logistic EMOS," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3031–3043, May 2019.
[22] A. W. Aryaputera, H. Verbois, and W. M. Walsh, "Probabilistic accumulated irradiance forecast for Singapore using ensemble techniques," in *Proceedings of the IEEE 43rd Photovoltaic Specialists Conference (PVSC)*, June 2016, pp. 1113–1118.
[23] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski, "Using Bayesian model averaging to calibrate forecast ensembles," *Monthly Weather Review*, vol. 133, no. 5, pp. 1155–1174, 2005.
[24] J. M. Sloughter, A. E. Raftery, T. Gneiting, and C. Fraley, "Probabilistic quantitative precipitation forecasting using Bayesian model averaging," *Monthly Weather Review*, vol. 135, no. 9, pp. 3209–3220, 2007.
[25] J. M. Sloughter, T. Gneiting, and A. E. Raftery, "Probabilistic wind speed forecasting using ensembles and Bayesian model averaging," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 25–35, 2010.
[26] R. M. Chmielecki and A. E. Raftery, "Probabilistic visibility forecasting using Bayesian model averaging," *Monthly Weather Review*, vol. 139, no. 5, pp. 1626–1636, 2011.
[27] R. J. Bessa, V. Miranda, A. Botterud, Z. Zhou, and J. Wang, "Time-adaptive quantile-copula for wind power probabilistic forecasting," *Renewable Energy*, vol. 40, no. 1, pp. 29–39, 2012.
[28] Y. Zhang and J. Wang, "K-nearest neighbors and a kernel density estimator for GEFCom2014 probabilistic wind power forecasting," *International Journal of Forecasting*, vol. 32, no. 3, pp. 1074–1080, 2016.
[29] M. Lotfi, M. Javadi, G. J. Osório, C. Monteiro, and J. P. S. Catalão, "A novel ensemble algorithm for solar power forecasting based on kernel density estimation," *Energies*, vol. 13, no. 1, p. 216, 2020.
[30] N.-B. Heidenreich, A. Schindler, and S. Sperlich, "Bandwidth selection for kernel density estimation: a review of fully automatic selectors," *AStA Advances in Statistical Analysis*, vol. 97, pp. 403–433, 2013.
[31] A. Bracale and P. De Falco, "An advanced Bayesian method for short-term probabilistic forecasting of the generation of wind power," *Energies*, vol. 8, no. 9, pp. 10 293–10 314, 2015.
[32] A. Bracale, G. Carpinelli, and P. De Falco, "A probabilistic competitive ensemble method for short-term photovoltaic power forecasting," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 2, pp. 551–560, 2017.
[33] P. Grana, "Push it to the limit: Rethinking inverter clipping," *Solar Power World*, September 8, 2017, https://www.solarpowerworldonline.com/2017/09/folsom-rethinking-inverter-clipping/.
[34] Y. Zhang, M. Beaudin, R. Taheri, H. Zareipour, and D. Wood, "Day-ahead power output forecasting for small-scale solar photovoltaic electricity generators," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2253–2262, Sept. 2015.
[35] S. Baran, "Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components," *Computational Statistics & Data Analysis*, vol. 75, pp. 227–238, 2014.
[36] P. W. Mielke, "Convenient beta distribution likelihood techniques for describing and comparing meteorological data," *Journal of Applied Meteorology*, vol. 14, no. 6, pp. 985–990, 1975.
[37] G. Heinze and M. Schemper, "A solution to the problem of separation in logistic regression," *Statistics in Medicine*, vol. 21, no. 16, pp. 2409–2419, 2002.
[38] C. Liu and D. B. Rubin, "The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence," *Biometrika*, vol. 81, no. 4, pp. 633–648, Dec. 1994.
[39] P. Lauret, M. David, and P. Pinson, "Verification of solar irradiance probabilistic forecasts," *Solar Energy*, vol. 194, pp. 254–271, 2019.
[40] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 2, pp. 243–268, 2007.
[41] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.
[42] T. Gneiting and R. Ranjan, "Comparing density forecasts using threshold- and quantile-weighted scoring rules," *Journal of Business & Economic Statistics*, vol. 29, no. 3, pp. 411–422, 2011.
[43] NOAA National Weather Service, "Environmental Modeling Center," https://www.emc.ncep.noaa.gov/.
[44] European Centre for Medium-Range Weather Forecasts, "Set I - Atmospheric Model high resolution 10-day forecast (HRES)," https://www.ecmwf.int/en/forecasts/datasets/set-i.
[45] R. N. Hoffman and E. Kalnay, "Lagged average forecasting, an alternative to Monte Carlo forecasting," *Tellus A: Dynamic Meteorology and Oceanography*, vol. 35, no. 2, pp. 100–118, 1983.
[46] S. D. Jascourt, D. Kirk-Davidoff, and C. Cassidy, "Forecasting solar power and irradiance – lessons from real-world experiences," in *Proceedings of American Solar Energy Society National Solar Conference 2016*, R. Perez and D. Renne, Eds., 2016, pp. 112–120.
[47] L. Ayompe, A. Duffy, S. McCormack, and M. Conlon, "Validated real-time energy models for small-scale grid-connected PV-systems," *Energy*, vol. 35, no. 10, pp. 4086–4091, 2010.
[48] "System Advisor Model," https://sam.nrel.gov/content/downloads, National Renewable Energy Laboratory (NREL), Golden, CO, USA, 2019.

Kate Doubleday (S'13) is a Ph.D. student in the Department of Electrical, Computer, and Energy Engineering and the Renewable and Sustainable Energy Institute at the University of Colorado Boulder. She is currently a Visiting Scholar in the Power Systems Engineering Center at the National Renewable Energy Laboratory.

Stephen Jascourt is a Senior Scientist at Maxar, where he created and continually works on advancing Maxar's solar irradiance and power forecasting capabilities and a variety of applied meteorology topics serving customers in the energy and agriculture industries. He previously was a scientific lead on numerical weather prediction professional development training for UCARs acclaimed COMET Program. He has a Ph.D. in Atmospheric Science from the University of Wisconsin-Madison.

William Kleiber is an Associate Professor of Applied Mathematics at the University of Colorado Boulder. He received his Ph.D. in Statistics from the University of Washington, and was a postdoctoral scholar at the National Center for Atmospheric Research. He received the Young Investigator Award from the American Statistical Association's Section on Statistics and the Environment, and was elected the junior Lebesgue Chair at the University of Rennes, France.

Bri-Mathias Hodge (SM' 17) is an Associate Professor in the Department of Electrical, Computer and Energy Engineering and a Fellow of the Renewable and Sustainable Energy Institute at the University of Colorado Boulder. He is also Chief Scientist in the Power Systems Engineering Center at the National Renewable Energy Laboratory.