## Title

Modeling Variability and Uncertainty of Photovoltaic Generation: A Hidden State Spatial Statistical Approach



## Authors

Tabone, Michaelangelo D
Callaway, Duncan S

## Publication Date

2015

Peer reviewed

# Modeling Variability and Uncertainty of Photovoltaic Generation: A Hidden State Spatial Statistical Approach

Michaelangelo D. Tabone, *Member, IEEE*, and Duncan S. Callaway, *Member, IEEE*

***Abstract*—In this paper, we construct, fit, and validate a hidden Markov model for predicting variability and uncertainty in generation from distributed (PV) systems. The model is unique in that it: 1) predicts metrics that are directly related to operational reserves, 2) accounts for the effects of stochastic volatility and geographic autocorrelation, and 3) conditions on latent variables referred to as "volatility states." We fit and validate the model using 1-min resolution generation data from approximately 100 PV systems in the California Central Valley or the Los Angeles coastal area, and condition the volatility state of each system at each time on 15-min resolution generation data from nearby PV systems (which are available from over 6000 PV systems in our data set). We find that PV variability distributions are roughly Gaussian after conditioning on hidden states. We also propose a method for simulating hidden states that results in a very good upper bound for the probability of extreme events. Therefore, the model can be used as a tool for planning additional reserve capacity requirements to balance solar variability over large and small spatial areas.**

***Index Terms*—Power system planning, solar energy, statistics.**

## I. Introduction

In 2013, U.S. grid-connected solar photovoltaic (PV) capacity increased by almost 40% (4.7 GW), and solar generation accounted for 29% of all newly installed electricity generation by nameplate capacity. This growth is taking place in a diverse setting of locations and sectors [1]. Of 2000 MW installed under California's Solar Initiative, over 99% of systems (82% of nameplate capacity) are less than 1 MW in size [2].

PV generation (along with all solar and wind generation) is different than traditional generation in two important ways: is it *variable*, meaning that it varies uncontrollably as the sun rises and sets, and as clouds pass over PV systems, and it is *uncertain*, meaning that it cannot be perfectly predicted in advance. These properties make PV generation more like electricity demand, which has always been variable and uncertain. Power systems maintain consistent balance of supply and demand as short time-scales by employing reserves, which are readily controllable generators (or loads) placed on stand-by to quickly increase or decrease generation.

These reserves will have to manage increasing amounts of variability and uncertainty as more solar and wind generators are connected. Predicting the amount of variability and uncertainty from PV generation within a balancing area (or an interconnection) is important for predicting future needs for reserves [3]–[6]. Because of the geographic auto-correlation of meteorological phenomena, the locational arrangement of PV panels (i.e., centralized or distributed) will have an effect on the amount of variability and uncertainty exhibited [7], [8].

### A. Renewable Generation and Operational Reserves

In this paper, we define variability and uncertainty to relate directly to the reserve needs of power systems. We examine two classes of operational reserves as they are defined in [9]: "load following reserves" account for the difference between a long time-scale market (typically 1 or 2-h intervals) and a faster market (anywhere between 30-min and 5-min intervals); "regulation reserves" account for the difference between the scheduled generation in the faster market and actual net load.

Most renewable integration studies use the "n-sigma" method to quantify the required amount of reserves following increases in wind and solar generation—see those cited within [9], [10]. The n-sigma method plans for variability or uncertainty that is "n" standard deviations away from a mean. This method implicitly assumes the net load variability and uncertainty are Gaussian, although the true distributions often have heavier tails [3], [8]–[11]. To account for this error many renewable integration studies use an artificially large "n" to compute confidence intervals [10]. A second existing approach, known as the "convolution method," computes the distribution of the sum of two random variables with any distribution shapes. However, historical data are needed to compute the original distribution shapes, making this method obsolete for studies that attempt to predict the effects of variable generators that have yet to be built [10].

### B. Statistical Models of Variability in PV

A number of studies have shown that PV production is geographically autocorrelated [8], [12], which is an important factor for predicting reserve requirements. Murata *et al.* [7] demonstrate that a geographic autocorrelation function relating the

Manuscript received February 03, 2014; revised June 15, 2014 and September 19, 2014; accepted November 03, 2014. This work was supported in part by an NSF Graduate Research Fellowship, the California Public Utilities Commission, and NSF grant CNS-1239467. Paper no. TPWRS-00162-2014.

The authors are with the Energy and Resources Group, University of California at Berkeley, Berkeley, CA 94720-3050 USA (e-mail: m.tabone@berkeley.edu; dcal@berkeley.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRS.2014.2372751

variability of PV generation from each pair of panels in a network[1] can be used to find the standard deviation of the aggregate generation from all panels in the network. Generation from arrangements of PV systems that are closely located are more correlated and the sum of generation from all of these systems exhibits a greater standard deviation than if the systems were more dispersed. Murata *et al.* predict the mean and standard deviation of variability or uncertainty, but these parameters are insufficient on their own to describe the full complexity of distributions of PV variability.

Distributions of PV production at individual sites have high kurtosis [8]. Cloud regimes are one explanation for this distribution shape, i.e., PV generation signals will have high standard deviation in partly cloudy times with fast wind speeds (resulting in fat distribution tails) and a much lower standard deviation during sunny times or fully overcast times (resulting in a high peak).

Recent studies have attempted to identify different cloud regimes and fit separate models for each of them. Lave and Kleissl condition their geographic auto-correlation parameters on cloud size and speed predictions from numerical atmospheric models [12], [13]. Perez *et al.* condition their variability and geographic auto-correlation parameters on the spatial variability of satellite predicted solar insolation [14]. Hummon *et al.* use "variability classes" to simulate the effects of different cloud regimes on PV generation for simulation purposes [11]. Reno and Stein use a "variability index" that classifies days using the standard deviation of the cloud cover ratio [15]. Wegener *et al.* use a hidden Markov model on wavelet coefficients to predict the standard deviation of 1-s variability from observations of 5-min variability at a single system [16]. However there is little to no evidence that conditioning the standard deviation and auto-correlation of PV variability on observations leads to an accurate distribution shape.

### C. Contributions of This Work

In this paper, we present, fit, and validate a model that predicts probability distributions of variability or uncertainty in distributed PV generation from networks of systems with any spatial arrangement. Our objective is to bridge a gap between statistical analysis of variability in PV generation, and power system planning models by

- **Using high temporal and spatial resolution data** from a set of closely located distributed PV systems.
- **Predicting distribution shapes and geographic autocorrelation** of variability and uncertainty between PV systems which allow us to estimate extreme events.
- Defining variability and uncertainty in a PV generation signal to be **directly related to operational reserves**.

We note that this paper has a methodological focus. We delay application of the method to future efforts; our immediate objective is to apply the method for long term forecasts of power system reserve requirements for future high renewable penetration scenarios. However, as we will discuss later, the method could also be used for short time scale (e.g., day ahead) forecasts for reserve requirements.

[1]We define a "network" as a spatial arrangement of PV systems, but not the electrical network connecting them.

The model presented here resembles hidden Markov models (HMM) for stochastic volatility, which have been used in the financial literature for some time [17], and have already been used to downscale 15-min resolution PV generation data to 1-s resolution data at a single system [16]. Hidden Markov models have also been used in the prior literature to forecast mean clearness index of PV insolation [18]–[20].

For this work, we focused exclusively on uncertainty and variability attributable to PV generation. We note that load following and regulation reserve requirements are ultimately a function of *net load*, which also includes wind and load. Existing work on reserve requirements assumes these time series are uncorrelated at the time scales on which reserves are deployed, and can be aggregated post-simulation [10]. We will discuss strategies for dealing with any known correlation between PV generation and wind or load in the conclusions.

### D. Model Overview and Structure of the Paper

We define data inputs and key metrics of interest in Section II, describe the model itself in Section III, and fit and validate the model in Section IV. Though the description of the data and model requires an extended discussion the model itself is relatively straightforward.

The model predicts metrics of variability or uncertainty, which are denoted $y(t)$ and are explained in Section II-C. Our primary innovation is that we condition the standard deviation of $y(t)$ on an endogenously estimated latent state, referred to as a "volatility state," $v(t)$, defined in (4)–(6) and fully explained in Section III. The volatility state for each PV system allows for rapid transitions between periods of high standard deviation and low standard deviation. The transition probabilities of $v(t)$ depend on a set of input variables $x(t)$, which can be any set of discrete observations from the geographic areas and time periods modeled. In this paper, we derive $x(t)$ from widely available 15-min resolution PV generation data, fully explained in Section II-B.

Model parameters are fit using observations of both $y(t)$ and $x(t)$; values of $v(t)$ are latent, meaning that they are never observed and instead are endogenously estimated during the fitting process. For simulation purposes, only observations of $x(t)$ and the parameters are required; distributions of $v(t)$ and $y(t)$ are produced by the simulation; as explained in Section IV-B. The use of latent states is shown to greatly benefit the prediction of extreme events, shown in Section IV-C.

## II. Data and Processing

The data we used for this study comprised instantaneous voltage and current measurements taken from residential and commercial PV installations provided to us by by photovoltaic integrator SolarCity. SolarCity provided 15-min resolution data for over 6000 systems from January 2011 to late September 2012. These data also included metadata on geometry and capacity for each PV system. To study variability at faster timescales SolarCity increased the sampling rate to once per minute at a small subset of systems.

Fig. 1 shows locations for sources of 1-min resolution generation data in the final dataset. We chose these systems to be in one of two 256 $\mathrm{km}^2$ areas, each representative of different types

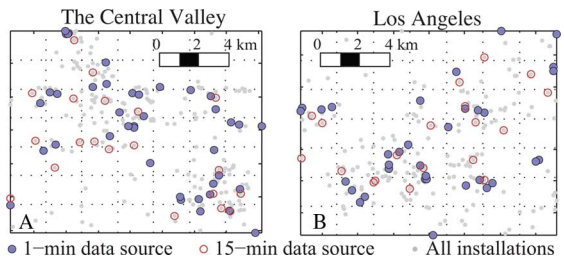


Fig. 1. Locations of systems with 15-min and 1-min resolution data.

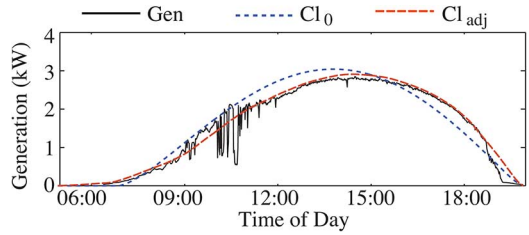


Fig. 2. 1-min resolution generation and clear sky signals calculated for one day in Los Angeles. $Cl_0$ is found using solar-earth geometry and system geometry from metadata, $Cl_{adj}$ is the empirically corrected clear sky signal.

of weather in California: the central valley (CV), and the Los Angeles coast (LA). We sampled 100 systems in each area from about 500 available using an algorithm that combined 1) quota sampling for distances between pairs of locations and 2) geographically random sampling of site pairs. Systems were monitored from mid-June to the end of August 2012. We used only systems for which there we no gaps in data over a period of 30 days, leaving us with data from 39 system inverters in LA and 55 in CV. The data in our sample cover a number of extreme events; half of the systems experience a 1-min duration ramp rate of more than 58% of their nameplate capacity, and half of the systems experience a 5-min average ramp of more than 34% of their nameplate capacity.

The remainder of this section describes the data we used to fit and validate the model. Sections II-A and II-B describe two variables on which we condition the model. Sections II-C describes our metrics for variability and uncertainty.

### A. Empirical Correction of Clear Sky Signal

The statistical model relies on a "clear sky signal," which represents generation that would have occurred in the absence of cloud cover. Solar-earth geometry—which is predictable given time of year, and location, tilt, azimuth, and effective capacity of a PV system—determines the clear sky signal. However, using the solar-earth model described in [21] with system metadata and a derate factor of 0.77, we found that the modeled clear sky signals were poor estimates of production on sunny days. This is likely due to errors in the system metadata and periodic shading from buildings and trees.

Fig. 2 shows 1-min resolution generation for one day along with a clear sky signal based on metadata along with a *corrected* clear sky signal. To implement the correction we first found the difference between actual 15-min production and the clear sky production predicted using only solar-earth geometry. Second, we identified a "clear sky deviation" as the 95th percentile those differences for each observed time of day, during a centered four-week moving window. Using this percentile excluded many low observations (which removed the effects of cloud-cover) as well as a small number of high observations (which removed the effects of occasional cloud reflection). Third, we smoothed the "clear sky deviation" signal using a 2-h moving average. After linearly interpolating between the 15-min intervals, we finally added the deviations back onto the clear sky signal predicted by solar-earth geometry.

### B. Model Inputs From 15-Min Resolution Data

As we explain in Section III, we condition the model's volatility states on discrete input data that are specific to the modeled times and locations. This facilitates simulation for locations and times external to this study. Fig. 1 shows locations for our sources of input data: PV systems that continuously recorded generation at 15-min intervals. To choose systems, we subdivide the study regions into 2-km grids. For each grid cell, we chose the system that was closest to the cell's centroid and not part of the 1-min dataset.

To compute conditioning inputs, we used a heuristic to estimate slow time scale volatility. We first calculated a moving standard deviation as follows:

$$\sigma_i(t) = \frac{1}{m+1}\left[\sum_{j=\left(t-\frac{m}{2}\right)}^{t+\frac{m}{2}}\left(\frac{S_i(j)}{CL_i(j)} - \mu(j)\right)^2\right]^{\frac{1}{2}} \quad (1)$$

where $S_i(t)$ is the solar generation from system $i$ at time $t$, $CL_i(t)$ is the clear-sky signal for system $i$ at time $t$, and $m$ is the number of intervals for the moving window ($m = 4$ to account for four 15-min intervals in an hour). We placed each standard deviation reading into one of 5 bins, resulting in the vector of data inputs for the model:

$$x(t) \in \{1, 2, \ldots 5\}^{N_g} \quad (2)$$

which contains one element for each of $N_g$ grid cells. Binning the data was necessary because the model is conditioned on discrete (not continuous) inputs, as explained in Section III. We defined the bin edges using equally spaced exponential intervals: 0, $e^{-3.5}$, $e^{-2.83}$, $e^{-2.16}$, and $e^{-1.5}$.

Panel A of Fig. 3 shows one day of generation at 15-min resolution along with the moving standard deviation of this signal; panel B shows and the resulting volatility heuristics (conditioning inputs) for this day.

### C. Variability and Uncertainty of PV Generation

Operational reserves are used to manage both uncertainty and variability in net load. Uncertainty arises from forecast error on the time scale of dispatch. For example, if hour-ahead markets dispatch generators in one hour blocks, there will be error between forecasted hourly average demand and actual hourly average demand. Variability arises because dispatch instructions for an interval must be further adjusted within the interval to

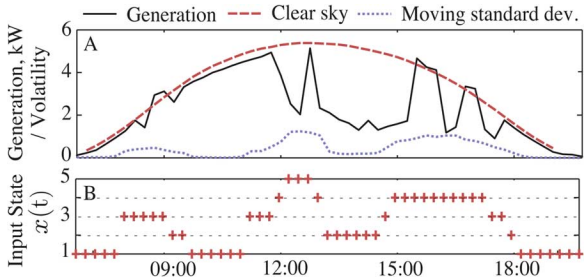


Fig. 3. A: 15-min generation, clear sky signal, and moving standard deviation for one system. B: Volatility heuristic based on the moving standard deviation.

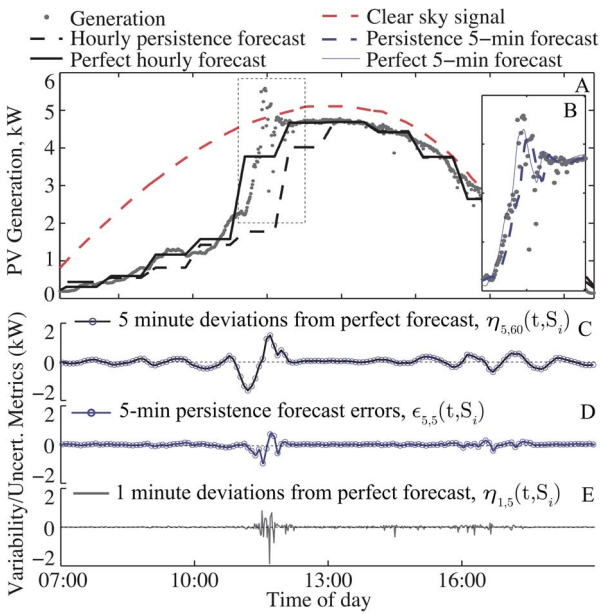


Fig. 4. Decomposition of 1 min PV generation. Panel A: Generation, hourly persistence forecast and hourly perfect forecast. Panel B (inset): 5-min persistence forecast and 5-min perfect forecast between 11 am and 12:30 pm (boxed). Panels C through E show each of the metrics we use to build models.

maintain supply-demand balance. We calculated metrics representing forecast errors (uncertainty) and deviations from perfect forecasts (variability) corresponding to the markets used by the CAISO [22], defined as:

- *Hourly forecast errors*: the difference between an hour-ahead forecast of hourly average demand and a perfect forecast of hourly average of demand.
- *5 min deviations*, $\eta_{5,60}(t, S_i)$: the difference between a 5-min resolution forecast (i.e., the forecast used for the load following market) and the perfect forecast of hourly average of demand. See Fig. 4(c).
- *5 min forecast errors*, $\epsilon_{5,5}(t, S_i)$: errors in a 5-min ahead persistence forecast of 5-min intervals. See Fig. 4(d).
- *1 min deviations*, $\eta_{1,5}(t, S_i)$: deviation of observed generation from a perfect 5-min forecast. See Fig. 4(e).

The total generation required from load following reserves is the sum of coincident hourly forecast errors and 5-min deviations; the generation required from regulation reserves is the coincident 5-min forecast error and 1-min deviations. Because the model we present in this paper is designed to describe sub-hourly variability and uncertainty, and because numerical weather methods would likely provide a much better hour-ahead forecast than the persistence method, we will not model hourly forecast errors.

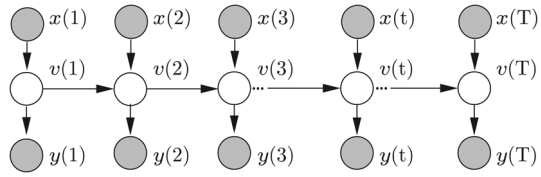


Fig. 5. Hidden Markov model represented as a directed acyclic graph.

Fig. 4 shows the decomposition process for each metric using one day of PV generation from one system. To forecast PV generation on short time scales, we used the persistence of a clearness index as suggested in [23]. Panel A of Fig. 4 shows the hour-ahead forecast and a "perfect" hourly forecast. The hourly profiles are as described in [22] where generators provide contracted energy during the middle 40 min of an hour and ramp to the next hour's contract during the following 20 min. Panel B of Fig. 4 shows the perfect and persistence forecast at 5-min intervals, where generators may use the entire 5-min interval to ramp.

## III. Volatility State Model

The model we develop in this section is an adaptation of a hidden Markov model (HMM) for stochastic volatility [17]. HMMs for stochastic volatility endogenously estimate the occurrence of sharp changes in the standard deviation of a signal such as those in Figs. 2 and 4. For each system and each time, the HMM classifies a reading as being in one of M possible latent states (referred to as "volatility states"), where the latent state defines the standard deviation of the signal.

Fig. 5 depicts the model as a directed acyclic graph, where shaded nodes represent observed variables and unshaded nodes represent unobserved, latent variables. We will estimate a separate model for each variability or uncertainty metric. $y(t) \in \mathbb{R}^{N_s}$, represents normalized variability or uncertainty, $\eta_{1,5}$, $\epsilon_{5,5}$, or $\eta_{5,60}$, it is defined over $N_s$ PV systems. $x(t) \in \{1, 2, \ldots 5\}^{N_g}$ represents the inputs to the model, which in our case are volatility heuristics for each of $N_g$ grid cells, discussed in Section II-B. $v(t) \in \{1, 2, \ldots M\}^{N_s}$ is a vector of unobserved volatility states, i.e., each system at each time is in one of $M$ volatility states. In what follows we describe these variables in further detail.

Define $y_i(t)$ as the normalization of a given variability or uncertainty metric for the $i$th PV system. For example, normalized regulation variability is

$$y_i(t) = \frac{\eta_{1,5}(t, S_i) - \eta_{1,5}(t, CL_i)}{\max_{j \in \text{hour}(t)} CL_i(j)} \tag{3}$$

where $S_i$ is the original PV generation time series and $CL_i$ is the clear sky time series. Subtracting $\eta_{1,5}(t, CL_i)$ and dividing by the clear sky trend $CL_i(t)$ removes non-stationary variability resulting from the solar diurnal cycle. Equations for $\epsilon_{5,5}$, or $\eta_{5,60}$ follow the same form. We assume each vector of metrics is mean zero multivariate Gaussian:

$$y(t) \sim MVG\left(0, \mathbf{\Sigma}\left(v(t); \phi\right)\right) \tag{4}$$

where the covariance matrix $\mathbf{\Sigma}$ is dependent on the volatility states, $v(t)$, and on an exponential geographic autocorrelation function, as defined in (5) and (6):

$$\mathbf{\Sigma}_{ij}\left(v(t); \sigma^2, \phi\right) = \begin{cases} \sigma^2_{v_{i(t)}} & i = j \\ \sigma_{v_{i(t)}} \sigma_{v_{j(t)}} \rho\left(v_i(t), v_j(t); d_{i,j}, \phi\right) & i \neq j \end{cases} \tag{5}$$

$$\rho(m, n; d_{i,j}, \phi) = a_{m,n} \cdot \exp\{-d_{i,j}/\tau_{m,n}\}. \tag{6}$$

Diagonal elements of the covariance matrix, $\Sigma$, contain the variances of each individual system such that if the $i$th system is in the $m$th volatility state $\Sigma_{i,i} = \sigma^2_m$. The off diagonals of $\Sigma$ represent covariance between systems, defined by the exponential geographic autocorrelation functions defined in (6), where $m$ and $n$ are the volatility states of systems $i$ and $j$, respectively, $d_{i,j}$ is the distance between systems $i$ and $j$, and $\phi$ is a set of parameters $\{\mathbf{a}, \boldsymbol{\tau}\}$. $\tau_{m,n}$ is a range parameter representing the distance over which correlation decreases by 63%. $a_{m,n}$ represents the correlation when $d_{i,j} = 0$. Due to heterogeneous cloud cover for adjacent systems, $a_{m,n}$ can be less than one.

We assume that the probability of being in a volatility state is conditionally dependent on the volatility state at the previous time step and on the input heuristic from the grid cell containing the system, $x_g(t)$ where $g$ indexes the grid cell. Equation (7) shows a set of Markov chain transition matrices that govern the progression of the volatility state for each system; $\mathbf{A}^{(k)} \in \mathbb{R}^{M \times M}$, $k$ indexes the input heuristic. Equation (8) describes each matrix element:

$$\mathbb{A} = \{\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots \mathbf{A}^{(K)}\} \tag{7}$$

$$\mathbf{A}^{(k)}_{m,n} = P\left(v_i(t) = n | v_i(t-1) = m, x_g(t) = k\right). \tag{8}$$

### A. Estimating the Model

We tested the performance of model fits with $M$ = 3,4,5,6,7 or 8 total volatility states. We cross validated the model fits by withholding 25% of the data during fitting (referred to as the "test data") and using it for validation. Data used for fitting are referred to as the "model data." We estimated parameters in two stages: the first stage estimates the entire model assuming no geographic autocorrelation, the second stage estimates autocorrelation parameters for each pair of volatility states given the output from stage 1.

*Stage 1*: $v(t)$, $\sigma^2$, and $\mathbb{A}$ are estimated via expectation-maximization (EM): First, EM chooses parameters that maximize likelihood given an expected value of the volatility states. Second, it recalculates the expected value of the volatility states given the updated model parameters. This gradient ascent process repeats to convergence; it is not guaranteed to find the global maximum but will reach a local maximum.

The "expectation" step of EM provides expected values of volatility states given the model parameters, defined in (9), where $\gamma_{i,m}(t)$ is the expected value of an indicator for whether system $i$ is in volatility state $m$ at time $t$, i.e., it is the probability of finding a given system in a particular state:

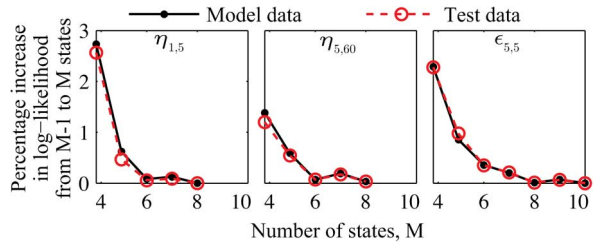


Fig. 6. Percent log-likelihood increase from increasing the number of states.

$$\gamma_{i,m}(t) = E\left[\mathbb{I}\left\{v_i(t) = m\right\}\right]. \tag{9}$$

*Stage 2*: Equation (10) describes a weighted correlation coefficient $\rho_{i,j,m,n}$ for each variability or uncertainty metric when system $i$ is in state $m$ and system $j$ is in state $n$:

$$\rho_{i,j,m,n} = \frac{\frac{\sum_{t=1}^{T} y_i(t) y_j(t) \gamma_{i,m}(t) \gamma_{j,n}(t)}{\sum_{t=1}^{T} \gamma_{i,m}(t) \gamma_{j,n}(t)}}{\sqrt{\frac{\sum_{t=1}^{T} \gamma_{i,m}(t) \gamma_{j,n}(t) y_i(t)^2}{\sum_{t=1}^{T} \gamma_{i,m}(t) \gamma_{j,n}(t)}} \sqrt{\frac{\sum_{t=1}^{T} \gamma_{i,m}(t) \gamma_{j,n}(t) y_j(t)^2}{\sum_{t=1}^{T} \gamma_{i,m}(t) \gamma_{j,n}(t)}}}. \tag{10}$$

The autocorrelation parameters, $a$ and $\tau$, are fit using weighted least squares from the correlation for each pair of sites. Equation (11) shows the weighted objective function for fitting the autocorrelation parameters:

$$\{a_{m,n}, \tau_{m,n}\} = \arg\min_{a,\tau} \sum_{i,j=1}^{N} \left(a \cdot e^{\frac{-d_{ij}}{\tau}} - \rho_{i,j,m,n}\right)^2 \times \sum_{t=1}^{T} \gamma_{i,m}(t) \gamma_{j,n}(t). \tag{11}$$

## IV. Results

### A. Parameter Estimation

Fig. 6 displays the log-likelihood of the model data (data used to fit the model) and the test data (reserved data for testing) fit to each metric. When $M > 5$ for $\eta_{1,5}$ and $\eta_{5,60}$ and $M > 7$ for $\epsilon_{5,5}$, improvements in log likelihood are small and added states are encountered less than 0.5% of the time. Therefore for the remainder of the analysis we use models with 5 states for $\eta_{1,5}$ and $\eta_{5,60}$, and 7 states for $\epsilon_{5,5}$.

Tables I and II show estimates of the autocorrelation parameters, $a_{m,n}$ and $\tau_{m,n}$, for 1-min and 5-min deviations, where higher volatility state index corresponds to higher standard deviation. For 5-min deviations, $a$ (correlation at a distance of 0 m) generally increases with volatility state standard deviation. Trends in $\tau$ suggest a non-monotonic relationship with volatility state, where the decay range is short for high and low variance states, and long for moderate variance.

### B. Validation by Simulation

To validate the model we compare observations to simulated distributions, which require estimates of the volatility states. We simulate volatility states with the following methods:

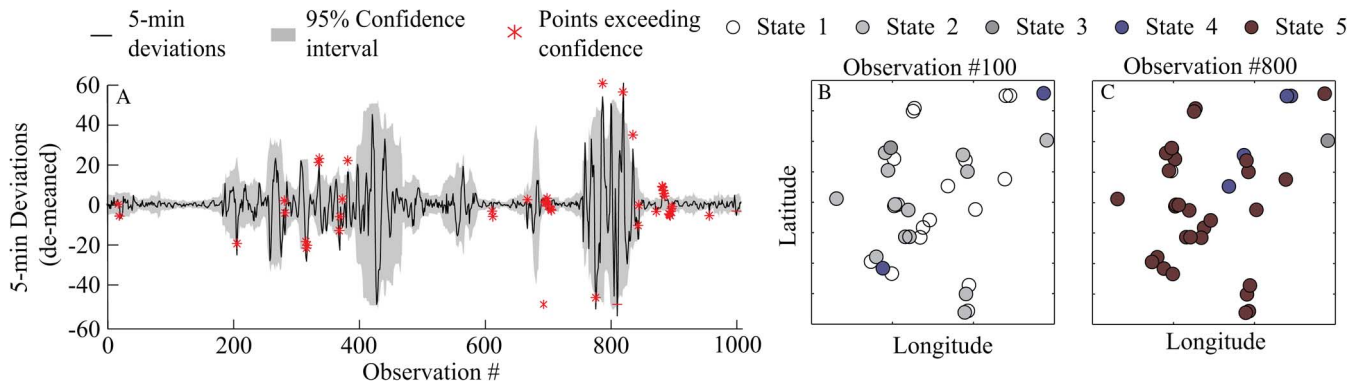


Fig. 7. Panel A: Observed, demeaned, aggregate 5-min variability ($\eta_{5,60}$) from the entire network of observed systems. Light grey boundaries represent the 95% confidence interval from method 1; points exceeding this interval are highlighted by red stars. Panels B and C: Example volatility states of systems.

TABLE I
PARAMETERS FOR AUTOCORRELATION FUNCTIONS FIT TO $\eta_{1,5}(t)$

| | a | | | | | $\tau$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **S#** | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| 1 | 0.41 | 0.18 | 0.05 | 0.02 | 0.00 | 0.21 | 0.40 | 0.53 | 0.21 | 0.10 |
| 2 | - | 0.21 | 0.14 | 0.04 | 0.01 | - | 1.54 | 2.05 | 1.76 | 0.10 |
| 3 | - | - | 0.53 | 0.35 | 0.10 | - | - | 1.25 | 1.18 | 0.56 |
| 4 | - | - | - | 0.66 | 0.39 | - | - | - | 0.97 | 0.68 |
| 5 | - | - | - | - | 0.61 | - | - | - | - | 0.38 |

TABLE II
PARAMETERS FOR AUTOCORRELATION FUNCTIONS FIT TO $\eta_{5,60}(t)$

| | a | | | | | $\tau$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **S#** | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| 1 | 0.20 | 0.17 | 0.07 | 0.06 | 0.00 | 12.76 | 12.94 | 6.94 | 0.61 | 0.10 |
| 2 | - | 0.23 | 0.17 | 0.06 | 0.00 | - | 8.45 | 5.98 | 2.46 | 0.10 |
| 3 | - | - | 0.49 | 0.38 | 0.18 | - | - | 6.51 | 8.73 | 7.58 |
| 4 | - | - | - | 0.61 | 0.51 | - | - | - | 7.84 | 5.88 |
| 5 | - | - | - | - | 0.74 | - | - | - | - | 5.42 |

- $v_i(t)$ *Simulation Method 1*: Set $v_i(t)$ equal to the most likely state as estimated during model fitting. This requires that we restrict simulations to times and locations for which we have one min data. Because we ultimately want to use the model to estimate reserves in the absence of 1-min data, this method is only a baseline.
- $v_i(t)$ *Simulation Method 2*: For each system at each time, simulate $N_p$ samples of volatility states using the stationary probabilities of the transition matrices in $\mathbb{A}$. In this paper, we use $N_p = 40$ samples. We model the distribution for each system at each time as a Gaussian mixture with $N_p$ equally likely components, one for each sampled covariance matrix. This method neglects correlation between volatility states.
- $v_i(t)$ *Simulation Method 3*: Simulate $N_p$ volatility states per site as in method 2, then independently sort the $N_p$ volatility states for each system at each time from highest to lowest variance, such that $v_i^{(1)}(t)$ and $v_i^{(N_p)}(t)$ contain the highest and lowest standard deviations, respectively; this maximizes correlation of volatility states.
- *No latent states*: We also construct, fit and test a separate benchmark model without latent states by conditioning standard deviation and the geographic autocorrelation parameters directly on the 15-min volatility heuristics, instead of on a latent state.

While it may be possible to estimate correlation between discrete volatility states, the problem is non-trivial; most geographic autocorrelation models use continuous distributions.

We use the simulated input states to calculate the standard deviation of aggregate variability or uncertainty from the entire network of monitored systems at each time. First, we calculate the normalized covariance matrix with (5). Second, we transform the covariance matrix to represent the de-normalization of the variability or uncertainty, i.e., the inverse of (3). For this transformation we multiply each element of the normalized covariance matrix by the hourly maximum of the clear sky signal for each system. Finally, we sum all elements of the covariance matrix to represent the summation of variability or uncertainty from all systems.

Fig. 7 shows the output of the method 1 simulation for $\eta_{5,60}(t)$. The 95% confidence interval is defined as twice the simulated standard deviation, and the volatility states are ordered such that state 1 has the lowest standard deviation and state 5 has the greatest. The confidence interval is wider when simulated volatility states have a greater standard deviation.

Fig. 8 shows quantile-quantile (QQ) plots that compare standardized quantiles of the observed data to quantiles of a standard normal. We standardize the quantiles by first computing the position of the observed data within the simulated model's cumulative density function for all observations in the study period, and then taking the standard normal inverse CDF of the result. If the model and its parameterization predict the empirical distribution, the standardized quantiles should be normally distributed, and points in the QQ plot will lie along the $y = x$ line.

Row 1 of Fig. 8 shows results for the no latent state model. For 1-min and 5-min deviations, the tails of the observed data are "heavy" compared to the standard normal, meaning that the simulated distribution will under-predict extreme events. For 5-min forecast errors, the tails of the observed data are "light" compared to the standard normal, meaning that the model over-predicts extreme events. In contrast, Row 2 of Fig. 8 shows that the volatility state model more accurately estimates distributions in the baseline scenario (method 1); distributions tails are only slightly light for deviations and slightly heavy for persistence forecast errors. This result indicates that if the volatility state distribution across sites is well characterized, the model

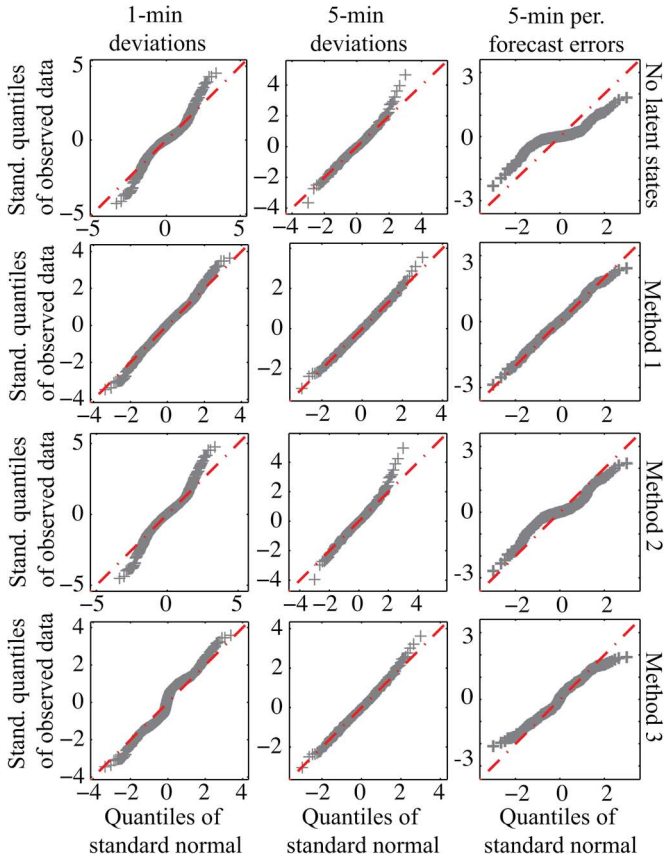


Fig. 8. QQ plots of test data pseudo-residuals using each volatility states simulation method and for each metric.

will work well in times and periods for which one min data are unavailable.

Row 3 of Fig. 8 show that method 2 results are worse than method 1 and comparable to the model without hidden states. Because method 2 does not model volatility state correlation across systems, the probability that multiple systems will be in a high volatility state simultaneously is relatively low, and we expect this simulations to under-predict extreme events—i.e., the observed data tails will be heavy. For 1- and 5-min deviations the observed tails are heavy, as expected. For 5-min persistence forecast errors the tails are light, though trend toward crossing the $y = x$ line back to heavy.

The bottom row of Fig. 8 shows results from method 3, which maximizes volatility state correlation across systems; the effect is evident if one compares the tails of the QQ plots between methods 2 and 3. For 1-min and 5-min deviations, method 3 performs similarly in the tails to method 1, suggesting that volatility states are in fact highly correlated. For 5-min forecast errors the simulated distribution has heavy tails.

### C. Predicting Maximum Events

Power system planners are concerned with high impact, low probability events, i.e., the maximum regulation or load following reserve required within a given time period. Equation (12) shows a method for finding the probability that all observations within a time period fall below some threshold, $x$, assuming independent observations; where $T_{HOD}$ is a specified hour of day. The 95% confidence interval for the maximum requirement corresponds to probabilities of 0.975, and 0.025:

$$p\left\{\eta_{1,5}(t) \leq x : t \in T_{HOD}\right\} = \prod_{t \in T_{HOD}} p\left\{\eta_{1,5}(t) \leq x\right\}. \tag{12}$$

Fig. 9 shows 95% confidence intervals for the maximum reserve requirement estimated for the test data, stratified by each hour of day. The left-hand column shows predictions from the volatility state model and the right-hand column shows predictions using the model without latent states. The dark grey boundaries are those calculated with known volatility states. The light grey (dotted line) boundary is 97.5% bound (i.e., the upper bound of a 95% confidence interval) of the predicted distribution using method 3, the worst-case assumption for geographic autocorrelation of volatility states. In expectation, observed maxima should fall above the 97.5% confidence bound between 0 and 1 times for 27 observations. For method 3 there are 3 observations above the bound, whereas the no latent state model has 7 above the bound. We note that the reserved test data—though randomly chosen—have slightly more extreme events than the data use to estimate the model. For 1-min and 5-min deviations, the method 3 upper bound dips below the upper bound for method 1 for a few hours; this results from small differences between stationary hidden state probabilities from method 3 versus the most likely volatility states predicted in method 1.

## V. Concluding Remarks

In this paper, we presented, fit, and validated a hidden Markov statistical model for variability and uncertainty in PV generation that parametrically estimates both geographic autocorrelation and stochastic volatility. The model differs from others in the literature by conditioning on latent "volatility states," which account for discontinuous changes in the standard deviation of variability or uncertainty from PV generation. We fit the model to metrics of PV generation that are useful for the planning of load following and regulation reserves: 1) 5-min persistence forecast errors made 5 min ahead, 2) 1-min deviations from a perfect 5-min interval forecast, and 3) 5-min deviations from a perfect hour ahead forecast. These metrics relate to the use of load following and regulation reserves by power system operators.

Given knowledge of the latent states (only possible for locations or times for which we have 1-min resolution data), the model predicts distributions well, even in the tails. For regions that lack 1-min data we built a method to simulate latent states in a way that maximizes their correlation (see method 3 in Section IV-B). This latent state simulation method produces comparable results to those when latent states are known (though the method is overly conservative for uncertainty). We also presented results for a model similar to existing models in that it conditions on observations instead of latent variables; this model under-predicts extreme events associated with variability and over-predicts extreme events associated with uncertainty relative to the model we developed in this paper. We expect that, by increasing the number of extreme events available to fit the model, additional data would increase

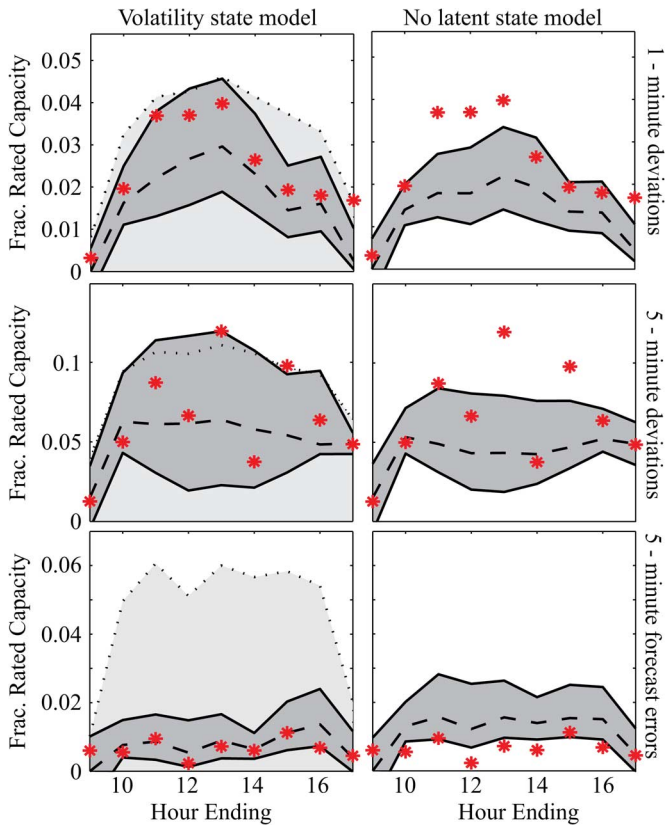


Fig. 9. Predicted distributions of maxima observed during each hour ending in the test data. Rows represent each metric ($\eta_{1,5}(t)$, $\epsilon_{5,5}(t)$, $\eta_{5,60}(t)$). Columns represent predictions by the volatility state model (left), and the no latent state model (right). The dark grey boundaries signify a 95% confidence interval found using the baseline scenario where the most likely volatility states predicted by EM is used, the light grey boundary signifies the 97.5th percentile of the predicted distribution of maxima using the sorting based worst case assumption. Red stars signify observed maxima.

the performance of the latent state model relative to the model without latent states.

Though earlier research [3], [8]–[11] and the results in this paper suggest that unconditioned Gaussian distributions do not characterize PV variability well, it is possible that as larger data sets become available, future research will show that a single Gaussian describes aggregate PV variability well across large spatial scales, for reasons related to the central limit theorem (CLT). We conjecture that unconditioned Gaussian distributions will work well for an aggregation of small systems distributed across hundreds to thousands of kilometers, but not for a small number of very large utility-scale systems, even if they are relatively far apart.

Our intended application of the model is to predict the amount of reserves required to integrate PV generation into power systems. Because the model accounts for spatial auto-correlation and is conditioned on spatial inputs, it is uniquely situated to compare the additional reserve requirements from centralized versus distributed PV systems. Any model that predicts reserve requirements must also account for variability from non-PV sources, namely load variability and wind variability. For future work, we plan to identify increases in operational reserve needs in California as centralized and distributed arrangements of PV are added. To identify these requirements, we will combine this model's simulated distributions of PV variability with similar metrics for wind and load predicted by the CAISO for the same time periods (as simulated for [24]). Combining wind, load, and solar variability is achieved by summing coincident forecast errors or deviations for each; i.e., the same method used in the n-sigma method or the convolution method [10]. We note that if there is known positive or negative correlation between solar and wind or load, the model we have developed can be combined with wind and load variability models in a number of ways that preserve this correlation. For example, one could condition wind and net load models on the same inputs as the PV model, which will induce some correlation due to the common predictors. Alternately, one may include variables representing wind and load directly within the model presented and estimate correlations endogenously—using the same mathematical framework developed in this paper.

One could also use this model's estimates for reserve requirements in unit commitment optimal dispatch models for different renewable energy penetration scenarios. Recent work from the National Renewable Energy Laboratory (NREL) uses a unit commitment optimal dispatch model that accounts for time-varying reserve needs, but in a way that would be improved by this model [5], [6]. This model, or a variation of it, may also be used by system operators to predict the required amount of reserves to procure.

## Acknowledgment

The authors would like to thank E. Carlson and K. Varadarajan of SolarCity who provided the data required for this work. The authors also would like to thank A. Von Meier and four anonymous reviewers, whose comments greatly improved this work.

## References

[1] Solar Energy Industry Association, Solar Market Insight Report 2013 Year in Review, Mar. 2014, Solar Energy Industry Association, Tech. Rep.

[2] The California Solar Initiative, Apr. 2014 [Online]. Available: http://www.gosolarcalifornia.org/about/csi.php

[3] G. E. Energy, Western Wind and Solar Integration Study, National Renewable Energy Laboratory, Golden, CO, USA, Tech. Rep. NREL/SR-550-47434, May 2010.

[4] M. Rothleder, Track I Direct Testimony of Mark Rothleder on Behalf of the California Independent System Operator Corporation, California Public Utilities Commission, 2011.

[5] M. Hummon, P. Denholm, J. Jorgenson, D. Palchak, B. Kirby, and O. Ma, Fundamental Drivers of the Cost and Price of Operating Reserves, National Renewable Energy Laboratory (NREL), Golden, CO, USA, Tech. Rep., 2013 [Online]. Available: http://www.nrel.gov/docs/fy13osti/58491.pdf

[6] E. Ela, V. Diakov, E. Ibanez, and M. Heaney, "Impacts of variability and uncertainty in solar photovoltaic generation at multiple timescales," *Contract* vol. 303, p. 2753000, 2013 [Online]. Available: http://www.nrel.gov/docs/fy13osti/58274.pdf

[7] A. Murata, H. Yamaguchi, and K. Otani, "A method of estimating the output fluctuation of many photovoltaic power generation systems dispersed in a wide area," *Elect. Eng. Jpn.* vol. 166, no. 4, pp. 9–19, Mar. 2009 [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/eej.20723/abstract

[8] A. Mills and R. Wiser, Implications of Wide-Area Geographic Diversity for Short-Term Variability of Solar Power, Lawrence Berkeley National Laboratory, 2010, Tech. Rep.

[9] M. Milligan, P. Donohoo, D. Lew, E. Ela, B. Kirby, H. Holttinen, E. Lannoye, D. Flynn, M. O'Malley, and N. Miller, "Operating reserves and wind power integration: An international comparison," in *Proc. 9th Int. Workshop Large-Scale Integration of Wind Power Into Power Systems*, 2010, pp. 18–29.

[10] H. Holttinen, M. Milligan, E. Ela, N. Menemenlis, J. Dobschinski, B. Rawn, R. Bessa, D. Flynn, E. Gomez Lazaro, and N. Detlefsen, "Methodologies to determine operating reserves due to increased wind power," in *Proc. 2013 IEEE Power and Energy Society (PES) General Meeting*, Jul. 2013, pp. 1–10.

[11] M. R. Hummon, E. Ibanez, G. Brinkman, and D. Lew, Sub-Hour Solar Data for Power System Modeling From Static Spatial Variability Analysis, National Renewable Energy Laboratory (NREL), Golden, CO, USA, 2012, Tech. Rep. Pending.

[12] M. Lave, J. Kleissl, and J. Stein, "A wavelet-based variability model (WVM) for solar PV power plants," *IEEE Trans. Sustain. Energy*, vol. 4, no. 2, pp. 501–509, 2013.

[13] M. Lave and J. Kleissl, "Cloud speed impact on solar variability scaling—Application to the wavelet variability model," *Solar Energy* vol. 91, pp. 11–21, May 2013 [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X13000406

[14] R. Perez, S. Kivalov, J. Schlemmer, K. Hemker, Jr., and T. E. Hoff, "Short-term irradiance variability: Preliminary estimation of station pair correlation as a function of distance," *Solar Energy* vol. 86, no. 8, pp. 2170–2176, Aug. 2012 [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X12000928

[15] M. J. Reno and J. Stein, Using Cloud Classification to Model Solar Variability, Sandia National Laboratories, Tech. Rep., 2013 [Online]. Available: http://energy.sandia.gov/wp/wp-content/gallery/uploads/SAND-2013-1692C%_ASES-CloudCategoryVariability.pdf

[16] J. Wegener, M. Lave, J. Luoma, and J. Kleissl, Temporal Downscaling of Irradiance Data via Hidden Markov Models on Wavelet Coefficients: Application to California Solar Initiative Data, 2012 [Online]. Available: http://solar.ucsd.edu/datasharing/doc/UCSDReport_1secCSI.pdf

[17] R. Langrock, I. L. MacDonald, and W. Zucchini, "Some nonstandard stochastic volatility models and their estimation using structured hidden Markov models," *J. Empiric. Finance* vol. 19, no. 1, pp. 147–161, Jan. 2012 [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0927539811000661

[18] Y. Li, L. He, and R.-Q. Nie, "Short-term forecast of power generation for grid-connected photovoltaic system based on advanced Grey-Markov chain," in *Proc. Int. Conf. Energy and Environment Technology, 2009 (ICEET '09)*, Oct. 2009, vol. 2, pp. 275–278.

[19] P. Poggi, G. Notton, M. Muselli, and A. Louche, "Stochastic study of hourly total solar radiation in Corsica using a Markov model," *Int. J. Climatol.* vol. 20, no. 14, pp. 1843–1860, Nov. 2000 [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/1097-0088(20001130)20:14<1843::AID-JOC561>3.0.CO;2-O/abstract

[20] A. Mellit, M. Benghanem, A. H. Arab, and A. Guessoum, "A simplified model for generating sequences of global solar radiation data for isolated sites: Using artificial neural network and a library of Markov transition matrices approach," *Solar Energy* vol. 79, no. 5, pp. 469–482, Nov. 2005 [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038092X05000204

[21] G. M. Masters, *Renewable and Efficient Electric Power Systems*, 1st ed. New York, NY, USA: Wiley-IEEE Press, Aug. 2004.

[22] Y. Makarov, C. Loutan, J. Ma, and P. de Mello, "Operational impacts of wind generation on California power systems," *IEEE Trans. Power Syst.* vol. 24, no. 2, pp. 1039–1050, May 2009 [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4808228

[23] E. Ibanez, G. Brinkman, M. Hummon, and D. Lew, "A solar reserve methodology for renewable energy integration studies based on subhourly variability analysis," in *Proc. 2nd Int. Workshop Integration of Solar Power in Power Systems*, Lisbon, Portugal, 2012 [Online]. Available: http://www.nrel.gov/docs/fy12osti/56169.pdf

[24] C. Goebel and D. Callaway, "Using ICT-Controlled plug-in electric vehicles to supply grid regulation in California at different renewable integration levels," *IEEE Trans. Smart Grid*, vol. 4, no. 2, pp. 729–740, Jun. 2013.

**Michaelangelo D. Tabone** (M'13) received the M.S. degree in energy and resources from the University of California at Berkeley, Berkeley, CA, USA, in 2012. He received the B.S. degree in chemical engineering, the B.A. degree in political science, and wrote a B.Phil. degree in life cycle assessment in 2010 from the University of Pittsburgh, Pittsburgh, PA, USA. Currently, he is pursuing the Ph.D. degree in the Energy and Resources Group at the University of California, Berkeley.

He is an affiliate of the Grid Integration Group at the Lawrence Berkeley National Laboratory. His research focuses on data driven methods for energy analysis, including 1) integration of renewable electricity into power systems, 2) enabling demand response, and 3) disaggregating behavioral and infrastructural drivers of energy use.

**Duncan S. Callaway** (M'08) received the B.S. degree in mechanical engineering from the University of Rochester, Rochester, NY, USA, and the Ph.D. degree in theoretical and applied mechanics from Cornell University, Ithaca, NY, USA.

He is currently an Assistant Professor of Energy and Resources at the University of California, Berkeley, Berkeley, CA, USA, and a faculty scientist at Lawrence Berkeley National Laboratory. Prior to joining the University of California, he was first a National Science Foundation (NSF) Postdoctoral Fellow at the Department of Environmental Science and Policy, University of California, Davis, Davis, CA, USA, subsequently worked as a Senior Engineer at Davis Energy Group, Davis, CA, USA, and PowerLight Corporation, Berkeley, CA, USA, and was most recently a Research Scientist at the University of Michigan, Ann Arbor, MI, USA. His current research interests are in the areas of 1) modeling and control of distributed energy resources and 2) using information technology to improve building energy efficiency.