Joint Estimation of Behind-the-Meter Solar Generation in a Community

Farzana Kabir, *Student Member, IEEE,* Nanpeng Yu, *Senior Member, IEEE,* Weixin Yao, *Member, IEEE,* Rui Yang, *Member, IEEE,* and Yingchen Zhang, *Senior Member, IEEE*

***Abstract*—Distribution grid planning, control, and optimization require accurate estimation of solar photovoltaic (PV) generation and electric load in the system. Most of the small residential solar PV systems are installed behind-the-meter making only the net load readings available to the utilities. This paper presents an unsupervised framework for joint disaggregation of the net load readings of a group of customers into the solar PV generation and electric load. Our algorithm synergistically combines a physical PV system performance model for individual solar PV generation estimation with a statistical model for joint load estimation. The electric load for a group of customers are estimated jointly by a mixed hidden Markov model (MHMM) which enables modeling the general load consumption behavior present in all customers while acknowledging the individual differences. At the same time, the new model can capture the change in load patterns over a time period by the hidden Markov states. The proposed algorithm is also capable of estimating the key technical parameters of the solar PV systems. Our proposed method is evaluated using net load, electric load, and solar PV generation data gathered from residential customers located in Austin, Texas. Testing results show that our proposed method reduces the mean squared error of state-of-the-art net-load disaggregation algorithms by 67%.**

***Index Terms*—Behind-the-meter solar generation, net load disaggregation, mixed hidden Markov model.**

NOMENCLATURE

Functions

g	Solar PV system performance model

Parameters

$\boldsymbol{\delta}$	Initial state distribution of the Markov chain
δ_k	Initial state probability at state k
η	Inverter efficiency
η_{nom}	Nominal inverter efficiency
$\boldsymbol{\Gamma}$	Tensor of transition probabilities of customers
$\boldsymbol{\Gamma}$	Matrix of transition probabilities of customer n
γ_{njk}	Transition probability from state j to state k of customer n
λ_{nt}^2	Variance of the error terms of customer n at time t
μ	Weight of errors in load estimates
ω	Weight of errors in solar PV generation estimates
$\boldsymbol{\Phi}$	Tensor of technical parameters of solar panels
$\boldsymbol{\Phi}_{max}$	Upper limit of technical parameters of solar panels
$\boldsymbol{\Phi}_{min}$	Lower limit of technical parameters of solar panels
$\boldsymbol{\Phi}_{mn}$	Technical parameters of m-th solar panel of n-th customer
$\boldsymbol{\Phi}_n$	Vector of technical parameters of n-th customer
$\hat{\boldsymbol{\Phi}}$	Tensor of technical parameter estimates of solar panels
$\hat{\boldsymbol{\Phi}}_{HMM}$	Tensor of PV system parameter estimates obtained from the HMM regression
$\boldsymbol{\Psi}_0$	Initial estimates of HMM regression parameters
σ^2	Variance of the customer specific random effect
$\boldsymbol{\Theta}$	MHMM parameters
$\boldsymbol{\Theta}_p$	MHMM parameter update at iteration p
$\boldsymbol{\Theta}_0$	Initial MHMM parameter estimates
θ_t	Solar PV array tilt angle
θ_{az}	Solar PV array azimuth angle
a_k	Common state-specific intercept of all customers for state k
B	Number of random samples
$\mathbb{B}$	Support of the distribution of random effects
c_{nk}	Vector of regression coefficients of explanatory variables for customer n at state k
l	Loss of solar PV array
M	Number of solar panels of each customer
N	Number of customers
P_{ac0}	AC nameplate rating of solar PV array in KW
$P_{dc0,inv}$	Inverter DC rating in KW
P_{dc0}	DC rating of solar PV array in KW
R	Dimension of a single solar panel's parameters
T	Time series length

Variables

α	Forward variable
β	Backward variable
ϵ_{nt}	Error term of customer n at time t
τ	Temperature in ^oC
τ_{wmv}	Weighted moving average of temperature of last 24 hours in ^oC
ε_{Load}	Errors in load estimates
ε_{PV}	Errors in solar PV generation estimates
ζ	Temperature coefficient
$\boldsymbol{b}$	Vector of random effects of all customers
b_n	Random effect corresponding to customer n
d	Day of the year
E	Mean MSE of net load of all customers
E_0	Reference irradiance in W/m^2
E_n	MSE of net load of customer n
E_{tr}	Transmitted irradiance in W/m^2
h	Hour of the day
$\boldsymbol{L}$	Matrix of electric load of customers
$\boldsymbol{L}_n$	Vector of electric loads of customer n

F. Kabir and N. Yu are with the Department of Electrical Engineering, University of California, Riverside, Riverside, CA, 92521-0429 USA e-mail: (fkabi001@ucr.edu, nyu@ece.ucr.edu).

W. Yao is with the Department of Statistics, University of California, Riverside, Riverside, CA, 92521-0429 USA e-mail: (weixin.yao@ucr.edu)

R. Yang and Y. Zhang are with the National Renewable Energy Laboratory, Golden, CO 80401 USA (e-mail: rui.yang@nrel.gov, Yingchen.Zhang@nrel.gov).

L_{nt}	Electric load of customer n at time t in KW
$\hat{\boldsymbol{L}}$	Matrix of electric load estimates of all customers
$\hat{\boldsymbol{L}}_{HMM}$	Matrix of electric load estimates obtained from the HMM regression
$\hat{\boldsymbol{L}}_n$	Vector of electric load estimates of customer n
$\hat{L}_{nt}$	Electric load estimate of customer n at time t in KW
$\boldsymbol{NL}$	Matrix of net load of customers
$\boldsymbol{NL}_n$	Vector of net load of customer n
NL_{nt}	Net load of customer n at time t in KW
$\hat{\boldsymbol{NL}}$	Matrix of net load estimates of all customers
P_{ac}	AC power output of solar PV array in KW
P_{dc}	DC power output of PV array in KW
$\boldsymbol{S}$	Matrix of solar PV generation of customers in KW
$\boldsymbol{S}_n$	Vector of solar PV generation of customer n
S_{nt}	Solar PV generation of customer n at time t in KW
$\hat{\boldsymbol{S}}$	Matrix of solar PV generation estimates of customers
$\hat{\boldsymbol{S}}_n$	Vector of solar PV generation estimates of customer n
$\hat{S}_{nt}$	Solar PV generation estimate of customer n at time t in KW
T_0	Reference cell temperature in ^oC
T_c	Cell temperature in ^oC
x	Matrix of explanatory variables
$\boldsymbol{z}$	Matrix of hidden states of all customers
$\boldsymbol{z}_n$	Vector of hidden states of customer n
z_{nt}	Hidden state of customer n at time t

I. Introduction

Solar PV generation is the fastest growing source of new energy. The introduction of net metering policy enables customers to sell excess electricity to the utility at the retail rate and receive credit on their electricity bill. As a result, small scale residential solar PV generation constituted 33% of total solar PV generation in the United States in 2019 [1]. Moreover, 61% of total solar PV systems in the United States were connected to the distribution system in 2014 [2]. Such high penetration of solar PV poses several challenges in the distribution system operation and planning processes [3]. For example, increasing solar PV generation can cause feeder over-voltage, voltage fluctuations, reverse power flow, protection system malfunction, and can exacerbate cold load pickup problem.

To mitigate these problems, it is imperative to design the system based on the amount of solar PV generation and native load in the distribution network. Thus, it is critical to develop a framework to disaggregate the net load measurements into solar PV generation and electric load. Furthermore, the technical parameters of solar PV systems need to be estimated for planning activities such as solar PV hosting capacity analysis.

The existing net load disaggregation algorithms can be classified into two groups: data-driven methods and model-based methods. The solar PV technical parameters are generally not available to the electric utilities. Detailed physical models such as PVWatts [4] and PV performance modeling collaborative [5] from Sandia National Laboratory are capable of estimating solar generation with information of solar irradiation, solar PV location, time, solar PV size, inverter efficiency, solar PV system loss, module tilt, and module orientation. Such physics-based behind-the-meter solar generation estimations are often inaccurate due to unreliable solar PV geometry data and degradation of PV arrays. The data-driven methods do not employ parametric physical models to estimate solar PV generation. Instead, they rely solely on smart meter data, supervisory control and data acquisition (SCADA), solar irradiance, and weather-related data.The data-driven methods can be further classified into two groups: unsupervised methods and methods that need supervision such as supervised/semi-supervised methods and contextually supervised source separation methods [6].

Supervised net load disaggregation methods need historical solar PV generation and load data of all customers whereas semi-supervised methods need the solar PV generation data for a small number of customers. The contextually supervised source separation method lies between supervised and unsupervised methods. This method also needs the solar PV generation data for a small number of representative customers as solar proxy. The studies by [7]–[11] leverage semi-supervised methods or contextually supervised source separation methods to disaggregate net-loads. The supervised data-driven approach is used in [12] to forecast net load.

The net load disaggregation problem is formulated as an optimization and a signal separation problem in [7]. The net load of a customer is modeled as a composite of representative electric load and solar generation patterns. The study by [11] estimates the power generation of behind the meter solar photovoltaic sites using a small set of selected representative sites while providing information on the uncertainty associated with the estimated solar PV generation volumes. The studies by [8] and [9] adapt a contextually supervised source separation model to disaggregate the net load signals of individual homes located on the same distribution feeder while enforcing various constraints. The contextually supervised source separation model is used to disaggregate the net load signals of feeder-level measurements in [10]. A supervised machine learning algorithm is utilized in [13] to solve the solar PV generation capacity estimation problem as a part of the net load disaggregation method under the assumption that actual measured solar PV generation and capacity data are available for a small number of representative solar PV sites.

Although supervised and semi-supervised net load disaggregation methods show great promise, they rely on solar PV generation data, which are typically not accessible for behind-the-meter systems. The advanced metering infrastructure (AMI) measurements only provide net load data which equals the load consumption minus solar PV generation. Thus, the historical solar PV generation, load data, and the solar PV technical parameters are not available to the electric utilities. Therefore, it is critical to develop an unsupervised framework to disaggregate the net load measurements into solar PV generation and electric load.

The studies by [14], [15], and [16] all leverage unsupervised net load disaggregation methods. The net load disaggregation problem is formulated as an optimization and a signal separation problem in [14]. In this study, the electric load of a customer is modeled as a composite of representative electric

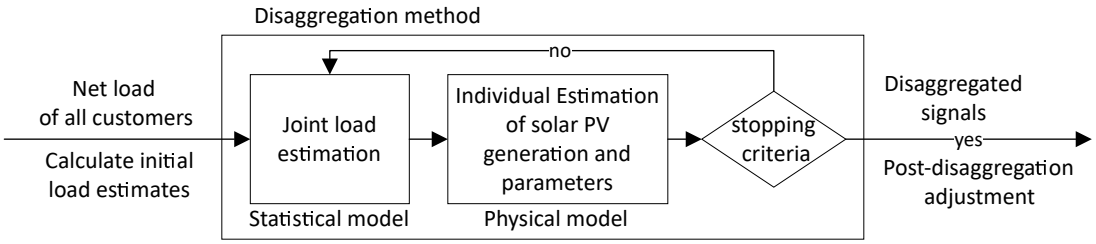

Fig. 1: The overall framework for joint net load disaggregation for a group of residential customers with behind-the-meter solar PV systems.

patterns. An unsupervised algorithm was developed in [15] to disaggregate the net load of a group of customers who have a common point of coupling. The algorithm proposed by [16] estimates electric load by comparing periods before PV installation with similar periods after PV installation that have common weather and activity characteristics and thereby perform net load disaggregation.

Although the pure data-driven methods have achieved some success, they are incapable of estimating the technical parameters of solar PV systems such as the tilt and DC size of the solar panel. These technical parameters of the solar PV systems are extremely useful for both short-term operation and long-term planning activities for electric utilities. Furthermore, the data-driven methods [9], [14], [15] often use a highly simplified linear model, which is incapable of capturing the nonlinear relationship among the solar irradiance, solar PV system geometry, and solar PV generation. In many cases, the pure data-driven methods [7]–[9], [11], [13] require historical solar PV generation data of a subset of customers, which can be difficult for electric utilities to obtain. Some data-driven methods [8], [9] could suffer from transposition errors if solar PV systems of different geometry are not available to serve as solar proxies. Moreover, these two algorithms also require joint estimation of a large number of hyperparameters, which makes the algorithm impractical and brittle. Some data-driven methods, such as [10], [11] only provide estimates of aggregate solar PV generation instead of the solar PV generation estimates for individual sites. The net load disaggregation algorithm proposed by [16] is built under the assumption that energy consumption habits do not significantly change once PV is installed which may not always hold. Moreover, changes in appliance mix or ownership of the house may also impact load patterns. Most net load forecasting algorithms only focus on the net load forecast problem and do not disaggregate the net load into electric load and solar PV generation. In addition, some of the net load forecasting algorithms only provide aggregated net load forecast [12].

Only a few model-based algorithms are developed to estimate the behind-the-meter solar PV generation. The 'SunDance' algorithm not only disaggregates the net-load data but also estimates the solar PV system geometry [17]. It has two key modules, a clear sky solar generation module and a module to map weather variables to solar PV output. However, this algorithm relies heavily on the availability of net-load data of a house when it is unoccupied on a sunny day. The aggregate capacity of all solar PV installations in a specific region is estimated in the algorithm proposed by [18] using correlation analysis, a grid search method, and a physics-based solar PV generation model. The estimated PV capacity is used to decompose the net load and ultimately forecast the net load. However, this method does not provide net load disaggregation of individual solar PV installations. We previously developed an iterative net load disaggregation algorithm for individual customers by seamlessly integrating a physical PV system performance model with a statistical load estimation model [19]. The PV system performance model can capture the complex relationship among the solar PV geometry, weather data, and solar PV generation. The hidden Markov model regression for the electric load is able to capture different customer consumption patterns over time. This algorithm not only provides estimates of technical parameters of the solar PV system but also reduces mean squared error by 44% compared to the state-of-the-art net load disaggregation algorithm by [14].

This paper extends our previous work to estimate behind-the-meter solar generation for a community of customers. This paper proposes estimating the electric load of a community of customers simultaneously with a mixed hidden Markov model (MHMM). The MHMM allows the sharing of information across individual customers, which leads to more accurate load and solar PV generation estimates. Specifically, the MHMM captures both the population-level effects and the individual differences in the power consumption patterns among the community of customers. Furthermore, the physical PV system performance model is extended to account for the case where a customer has multiple strings of solar panels facing different directions. At last, the performance of our proposed method is compared with the state-of-the-art net load disaggregation algorithms using the data from residential customers in Austin, Texas [20].

The unique contributions of this paper are as follows:

1) An MHMM is developed to jointly estimate the electric load of a community of customers, which captures both the population and the individual effects.
2) The proposed net load disaggregation algorithm seamlessly integrates a statistical MHMM with a physical PV system performance model, which accounts for solar panels facing different directions.
3) The proposed behind-the-meter solar generation estimation algorithm yields significantly higher accuracy over state-of-the-art net load disaggregation algorithms including our previous work [19].

The remainder of the paper is organized as follows: Section II provides the overall framework of the proposed algorithm. Section III presents the technical methods, which include the PV system performance model, MHMM, and the net load disaggregation algorithm. Section IV shows the numerical study results. Section V states the conclusions.

II. Overall Framework

The net load measurement of a residential customer equals the electrical load of the customer minus the solar PV generation. Let L_{nt} be the electrical load and S_{nt} be the solar generation of a customer n at time t. Then the net load readings of the customer NL_{nt} for customers $n = 1, \ldots, N$ at time intervals $t \in \{1, 2, ..., T\}$ can be written as follows:

$$NL_{nt} = L_{nt} - S_{nt}; \quad S_{nt} \geq 0 \quad \forall t, n \tag{1}$$

The aim of the net load disaggregation algorithm is to decompose the net load readings NL_{nt} of a group of N residential customers with solar PV systems into the corresponding solar PV generation S_{nt} and electric load L_{nt} at each time interval t. The exact location of the customers, historical PV generation or consumption, solar panel configuration, or other solar PV system parameters are generally not available. Our proposed algorithm does not require this information.

The overall framework of the proposed net load disaggregation algorithm is shown in Fig. 1. A statistical MHMM is first fit to jointly estimate the electric load of all customers with the initial estimates of the load model parameters while keeping the solar generation estimates fixed. The parameter estimation of the mixed hidden Markov model (MHMM) can be computationally intensive. Therefore, a good initial estimate of the load is needed as the starting point of the iterative net load disaggregation algorithm. The electric load estimates obtained from the iterative algorithm with HMM regression from [19] is used as the initial load estimates for the MHMM. Solar PV system parameters and solar PV generation of individual customers are then estimated with a physical model while keeping the load estimates fixed. The iterative estimation procedure continues until the stopping criteria are met. At last, a post-disaggregation adjustment is performed on the disaggregated signals to ensure that the equality constraint (1) relating native electric load, solar PV generation, and net load is satisfied. The joint modeling of load with MHMM and the parameter estimation procedure are described in Section III-A and III-B. The physical solar PV system performance model and the estimation of the technical parameters of solar PV systems are presented in Section III-C. The net load disaggregation algorithm will be discussed in detail in Section III-D.

III. Technical Methods

A. Mixed Hidden Markov Model

Many regression models are used to model load consumption of a customer to incorporate the effect of weather and time. However, the user consumption pattern is expected to be quite different depending on whether a customer is at home or not. For example, when a customer is at home, the load may consist of heating, ventilation, air-conditioning and other appliance usage. On the other hand, when the customer is not at home, the load can be very low and may include power usage from the refrigerator and other appliances, such as water heaters, and TVs in standby mode. In order to model such heterogeneous user consumption patterns, a hidden Markov model (HMM) regression [21] is used to model individual load time series of customers in [19]. However, in [19], the HMM needs to be fitted separately for each individual customer and thus the model is incapable of leveraging the community information to improve the load modeling.

To improve over the individual HMM regression models in [19], a *mixed hidden Markov model* (MHMM) [22] is proposed to provide a joint load estimation of all customers by modeling both the population and the individual effects. The individual heterogeneity can be captured by the individual-specific random effects in the MHMM representing individual deviations from the population averages.

Let L_{nt} be the load and z_{nt} be the hidden state associated with the customer n at time t, $n = 1, \ldots, N$, $t = 1, \ldots, T$. Let $z_{nt} = 1$ if the customer is at home and $z_{nt} = 2$, if not home, making the number of total states $K = 2$. Let $\boldsymbol{L}_n$ denote the T-dimensional vector of observations, i.e., load of customer n across T time points and $\boldsymbol{L}$ denote the $T \times N$-dimensional matrix of load of all customers. The vectors of hidden states, $\boldsymbol{z}_n$ and $\boldsymbol{z}$, are defined analogously. Let $\boldsymbol{x}$ be the $T \times Q$-dimensional matrix of explanatory variables or fixed effects. The explanatory variables include temperature (τ), exponential moving average of the temperature of last 24 hours (τ_{wmv}), hour of the day (h), and the interaction of temperature and hour of the day ($\tau \times h$). To capture the non-linear relationship between temperature and load, a 3rd-degree polynomial of temperature is used, denoted by $\boldsymbol{\tau}, \boldsymbol{\tau^2}$, and $\boldsymbol{\tau^3}$ following the proposal of Hagan [23]. Based on some empirical analysis, a 3rd-degree polynomial of the *hour* of the day is also used. The explanatory variable matrix $\boldsymbol{x}$ is denoted by $\boldsymbol{x} = \left[\boldsymbol{\tau}, \boldsymbol{\tau^2}, \boldsymbol{\tau^3}, \boldsymbol{\tau}_{wmv}, \boldsymbol{d}, \boldsymbol{h}, \boldsymbol{h^2}, \boldsymbol{h^3}, \boldsymbol{\tau} \times \boldsymbol{h}\right]$.

A hidden Markov model (HMM) is defined as a pair of stochastic processes $\{z_{nt}, y_{nt}\}$, where z_{nt} is an unobserved finite state Markov chain and the output process y_{nt} is related to the latent state process z_{nt}. An MHMM extends HMMs to a regression setting in a generalized linear mixed model framework. MHMM combines HMMs with a linear mixed effect regression model in a longitudinal setting and enables the incorporation of covariates and random effects in both the conditional and/or transition model. A random intercept model is assumed for the conditional model to allow the customers to borrow information from each other and to simultaneously incorporate the heterogeneity across different customers.

Several assumptions are made for the MHMM. First, the random effects are assumed to follow a normal distribution and are independent of the hidden states. Second, given the random effects, the dependence structure of the latent time series $\{z_{nt}\}_{t=1}^{T}$ can be modeled by an underlying Markov chain. The transition probability from state j to state k for customer n is denoted by $\gamma_{njk} = P\left(z_{nt} = k | z_{n(t-1)} = j, z_{n(t-2)} = l, \ldots\right) = P\left(z_{nt} = k | z_{n(t-1)} = j\right)$ where $j, k = 1, 2$ and γ_{njk}

satisfies $\sum_{k=1}^{2} \gamma_{njk} = 1$ for each j and n. The initial state distribution of the Markov chain is denoted by $\boldsymbol{\delta}$ and the transition matrix of all of the customers is denoted by $\boldsymbol{\Gamma}$. Third, conditional on the random effects, the n-th process, $\{L_{nt}\}_{t=1}^{T}$, is a HMM, and observations on different processes from different customers are independent. Therefore, given state z_{nt}, an MHMM with customer-specific random intercepts in the conditional model can be written as follows:

$$L_{nt} = a_{z_{nt}} + b_n + \boldsymbol{c}_{n,z_{nt}}\boldsymbol{x}_t + \epsilon_{nt}, \quad n = 1 \dots N, \quad t = 1 \dots T, \quad b_n \sim N\left(0, \sigma^2\right), \quad \epsilon_{nt} \sim N\left(0, \lambda^2_{z_{nt}}\right) \tag{2}$$

Here, $a_{z_{nt}}$ is the common state-specific intercept of all customers and $\boldsymbol{c}_{n,z_{nt}}$ is the Q-dimensional vector of customer and state-specific regression coefficients of explanatory variables. Both $\boldsymbol{a}$ and $\boldsymbol{c}$ are fixed effect coefficients. Here, b_n is the customer-specific random effect common to all states and follows a normal distribution with variance σ^2. The individual error term ϵ_{nt} follows a normal distribution with state-specific variance $\lambda^2_{z_{nt}}$. Therefore, conditional on the state z_{nt} and random effect, the distribution of L_{nt} is

$$f\left(L_{nt}|z_{nt}, b_n\right) \sim N\left(a_{z_{nt}} + \boldsymbol{c}_{n,z_{nt}}\boldsymbol{x}_t + b_n, \lambda^2_{z_{nt}}\right), \quad b_n \sim N\left(0, \sigma^2\right) \tag{3}$$

There are a few advantages of using MHMM instead of HMM to jointly model electric load of a community of customers. First, the random effects enter additively in the linear predictor and thus may represent the influence of omitted covariates or individual heterogeneity not captured by the observed covariates. Second, traditional HMM assumes that the observations are independent given the hidden states. To meet this assumption, an extremely large number of latent states is often required. However, in this case the HMM becomes uninterpretable. MHMM allows for the dependence between the longitudinal observations of the same customers by means of the customer-specific random effect and hence provides more efficient estimates of fixed model parameters. As the number of latent states in MHMM is not required to be large, MHMM is relatively easy to interpret. Third, MHMM assumes that the random effects follow a common distribution which makes the estimates of the random effect shrunk towards their mean (i.e., a weighted average between the overall mean effect and the individual effect). Thus, the estimation of individual effects also borrows information from each other.

Borrowing information across customers as an advantage of MHMM is further elaborated below. An advantage of joint load estimation over individual load estimation is its ability to borrow strength across customers by obtaining estimates of parameters common to all customers known as the fixed effect or population level effect. In addition, MHMM is also able to capture the individual heterogeneity of customers through the individual-specific random effects while retaining the strength of joint load estimation. Thus, MHMM effectively treats the customers as distinct entities but from the same general population. Additionally, MHMM provides an estimate of the variance of the random effect distribution. MHMMs have been successfully applied in various scientific fields, notably for modeling animal movement and behavior [24], lesion count in multiple sclerosis patients [22], forest tree growth [25], and teenage driving behavior [26].

B. Estimation of MHMM by MCEM Algorithm

Traditionally, the expectation–maximization (EM) algorithm has been used to estimate the parameters of a HMM. The EM algorithm is an iterative method for performing maximum likelihood estimation when some of the data are missing. Unfortunately, the basic EM algorithm can not be applied to MHMM directly due to the existence of random effects and the complex numerical integration. The Monte Carlo expectation maximization (MCEM) algorithm is used [27] to estimate the unknown parameters $\boldsymbol{\Theta} = \left[\{a_k\}_{k=1}^{K}, \{c_{nk}\}_{n=1,k=1}^{n=N,k=K}, \{\lambda_k\}_{k=1}^{K}, \boldsymbol{\delta}, \boldsymbol{\Gamma}, \sigma\right]$ of the MHMM. MCEM is a stochastic approximation method that is especially useful for cases where numerical integration and maximization are not advised, e.g., when there is a large number of random effects or a large number of parameters. Let $F_{nt} = f\left(L_{nt}|z_{nt}, b_n, \boldsymbol{\Theta}\right)$ and $\mathbb{B}$ be the support of the distribution of random effects $\boldsymbol{b}$. The complete data likelihood (CDLL) for N customers can be written as:

$$\begin{aligned} &\mathcal{L}_c\left(\boldsymbol{\Theta}; \boldsymbol{L}, \boldsymbol{z}, \boldsymbol{b}\right) \\ &= \int_{\mathbb{B}} f(\mathbf{L}|\mathbf{z}, \mathbf{b}, \boldsymbol{\Theta}) f\left(\mathbf{z}|\mathbf{b}, \boldsymbol{\Theta}\right) f\left(\mathbf{b}; \boldsymbol{\Theta}\right) \mathbf{db} \\ &= \int_{\mathbb{B}} \prod_{n=1}^{N} \left\{ \sum_{z_n} \delta_{z_{n1}} F_{n1} \prod_{t=2}^{T} \gamma_{z_{n(t-1)}, z_{nt}} F_{nt} \right\} f\left(b_n; \boldsymbol{\Theta}\right) db_n \end{aligned} \tag{4}$$

For notational convenience, the following indicator variables are defined for $t = 2, 3, \dots T$, $u_{ntj} = 1$ if and only if $z_{nt} = j$ and $v_{ntjk} = 1$ if and only if $z_{n(t-1)} = j$ and $z_{it} = k$

Then the complete log likelihood can be written as

$$\begin{aligned} &\log \mathcal{L}_c\left(\boldsymbol{\Theta}; \mathbf{L}, \mathbf{z}, \mathbf{b}\right) \\ &= \sum_{n=1}^{N} \left\{ \sum_{j=1}^{K} u_{n1j} \log \delta_j + \sum_{t=2}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} v_{ntjk} \log \gamma_{njk} \right. \\ &\quad + \int_{\mathbb{B}} \sum_{t=1}^{T} \sum_{j=1}^{K} u_{ntj} \log F_{nt} db_n \\ &\quad \left. + \int_{\mathbb{B}} f\left(b_n|\boldsymbol{L}_n\right) \log f\left(b_n\right) db_n \right\} \end{aligned} \tag{5}$$

The MCEM is an iterative algorithm requiring two steps at each iteration: computation of a particular conditional expectation of the log-likelihood (E-step) and the maximization of this expectation over the relevant parameters (M-step).

In the E-step, the expectation of the complete data log likelihood (CDLL) conditional on the observed data and parameter estimates at iteration p are calculated. To write the CDLL, the hidden states and the random effects are treated as missing data. We replace the indicator variables by their conditional expectations given the observations $\boldsymbol{L}$ and the current parameter estimates Θ^p. The computation of

CDLL is not easy due to the high-dimensional integration. The MCEM approximates the CDLL by a Monte Carlo method. Let B random samples $b_n^1, \ldots, b_n^B$ be generated from the distribution $f(\mathbf{b_n}; \mathbf{\Theta}^p)$. Defining $F_{nt}^l = f\left(L_{nt}|z_{nt}, b_n^l, \mathbf{\Theta}^p\right)$, the following approximation can be obtained to the conditional expectation of the CDLL.

$$
\begin{aligned}
&E\left[\log \mathcal{L}_c\left(\mathbf{\Theta}; \mathbf{L}, \mathbf{z}, \mathbf{b}|\mathbf{L}, \mathbf{\Theta}^p\right)\right] \qquad (6)\\
&\approx \sum_{n=1}^{N}\left\{\sum_{j=1}^{K}\hat{u}_{n1j}\log\delta_j + \sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K}\hat{v}_{ntjk}\log\gamma_{njk} + \right.\\
&\left.\sum_{l=1}^{B}\sum_{t=1}^{T}\sum_{j=1}^{K}\hat{u}_{ntj}h(b_n^l|\boldsymbol{L}_n)\log F_{nt}^l + \sum_{l=1}^{B}h\left(b_n^l|\boldsymbol{L}_n\right)\log f\left(b_n^l\right)\right\}\\
&\text{where}\quad h\left(b_n^l|\boldsymbol{L}_n\right) = \frac{f\left(\boldsymbol{L}_n|b_n^l, \mathbf{\Theta}^p\right)}{\sum_{l=1}^{B} f\left(\boldsymbol{L}_n|b_n^l, \mathbf{\Theta}^p\right)}
\end{aligned}
$$

For ease of implementation of the MCEM algorithm, forward variable is defined as follows:

$$
\begin{aligned}
\alpha_{nt}\left(j, b_n^l\right) &= f\left(L_{n1,\ldots,}, L_{nt}, z_{nt} = j|b_n^l\right) \qquad (7)\\
&= \boldsymbol{\delta} f\left(L_{n1}\right)\left(\prod_{t=2}^{t}\mathbf{\Gamma}_n f\left(L_{nt}\right)\right)
\end{aligned}
$$

The forward variable can be computed recursively by

$$
\alpha_{n1}\left(j, b_n^l\right) = \delta_j f\left(L_{n1}|z_{n1} = j, b_n^l\right) \qquad (8)
$$

$$
\begin{aligned}
\alpha_{n(t+1)}\left(k, b_n^l\right) = \sum_{j=1}^{m}\{&\alpha_{nt}\left(j, b_n^l\right)\gamma_{njk} \qquad (9)\\
&f\left(L_{n(t+1)}|z_{n(t+1)} = k, b_n^l\right)\}
\end{aligned}
$$

Similarly, backward variable is defined as follows:

$$
\begin{aligned}
\beta_{nt}\left(j, b_n^l\right) &= f\left(L_{n(t+1)}, \ldots, L_{nT}|z_{nt} = j, b_n^l\right) \qquad (10)\\
&= \left(\prod_{t=t+1}^{T}\mathbf{\Gamma}_n f\left(L_{nt}\right)\right)1'
\end{aligned}
$$

Backward variable can be calculated recursively by

$$
\beta_{nT}\left(j, b_n^l\right) = 1 \qquad (11)
$$

$$
\begin{aligned}
\beta_{nt}\left(j, b_n^l\right) = \sum_{k=1}^{K}\{&\gamma_{njk}\beta_{n(t+1)}\left(k, b_n\right) \qquad (12)\\
&f\left(L_{n(t+1)}|z_{n(t+1)} = k, b_n^l\right)\}
\end{aligned}
$$

Let $F_{ntk} = f\left(L_{nt}|z_{nt} = k, b_n^l\right)$. The conditional expectation of the indicatior variables $\hat{u}_{ntj}$ and $\hat{u}_{ntj}$ can then be defined as follows:

$$
\begin{aligned}
\hat{u}_{ntj} &= f\left(z_{nt} = j|\boldsymbol{L}_n, \boldsymbol{\theta}_p\right) \qquad (13)\\
&= \frac{\sum_{l=1}^{B}\alpha_{it}\left(j, b_n^l\right)\beta_{nt}\left(j, b_n^l\right)h(b_n^l|\boldsymbol{L}_n)}{\sum_{j=1}^{M}\sum_{l=1}^{B}\alpha_{nt}\left(j, b_n^l\right)\beta_{nt}\left(j, b_n^l\right)h\left(b_n^l|\boldsymbol{L}_n\right)}
\end{aligned}
$$

and

$$
\begin{aligned}
\hat{v}_{ntjk} &= f\left(z_{n(t-1)} = j, z_{nt} = k|\mathbf{y}, \boldsymbol{\theta}_p\right) = \qquad (14)\\
&\frac{\sum_{l=1}^{B}\alpha_{n(t-1)}\left(j, b_n^l\right)\gamma_{njk}F_{ntk}\beta_{nt}\left(k, b_n^l\right)h(b_n^l|y_n)}{\sum_{j,k=1}^{M}\sum_{l=1}^{B}\alpha_{n(t-1)}\left(j, b_n^l\right)\gamma_{njk}F_{ntk}\beta_{nt}\left(k, b_n^l\right)h(b_n^l|\boldsymbol{L}_n)}
\end{aligned}
$$

To avoid numerical underflow when α and β are very small, $\hat{u}_{ntjk}$ and $\hat{v}_{ntjk}$ can be calculated using logarithms, the approximation of $\log(p+q)$ by [28] and the log-sum-exp trick. In the M-step, the parameters $\mathbf{\Theta}$ are updated by maximizing the expected CDLL in (6) with respect to $\mathbf{\Theta}$. The first, second, and fourth term of (6) are maximized with respect to $\boldsymbol{\delta}$, $\mathbf{\Gamma}$ and σ^2, respectively. The third term of (6) is maximized with respect to $\boldsymbol{c}, \boldsymbol{a}$, and $\boldsymbol{\lambda}$. Since the conditional distribution and the random effects follow a normal distribution, closed form solutions of $\mathbf{\Theta}$ are available.

$$
a_j = \frac{\sum_{n=1}^{N}\sum_{l=1}^{B}\sum_{t=1}^{T}\left(L_{nt} - c_{nj}\boldsymbol{x}_t - b_n^l\right)\hat{u}_{ntj}h\left(b_n^l|\boldsymbol{L}_n\right)}{\sum_{n=1}^{N}\sum_{l=1}^{B}\sum_{t=1}^{T}\hat{u}_{ntj}h\left(b_n^l|\boldsymbol{L}_n\right)} \qquad (15)
$$

$$
\boldsymbol{c}_{nj} = \frac{\sum_{l=1}^{B}\sum_{t=1}^{T}\left(L_{nt} - a_j - b_n^l\right)\hat{u}_{ntj}h\left(b_n^l|\boldsymbol{L}_n\right)\boldsymbol{x}_t}{\sum_{l=1}^{B}\sum_{t=1}^{T}\hat{u}_{ntj}h\left(b_n^l|\boldsymbol{L}_n\right)\boldsymbol{x}_t^2} \qquad (16)
$$

$$
\lambda_j^2 = \frac{\sum_{n=1}^{N}\sum_{l=1}^{B}\sum_{t=1}^{T}\left(L_{nt} - a_j - \boldsymbol{c}_{nj}\boldsymbol{x}_t - b_n^l\right)^2\hat{u}_{ntj}h\left(b_n^l|\boldsymbol{L}_n\right)}{\sum_{n=1}^{N}\sum_{l=1}^{B}\sum_{t=1}^{T}\hat{u}_{ntj}h\left(b_n^l|\boldsymbol{L}_n\right)} \qquad (17)
$$

$$
\delta_j = \frac{\sum_{n=1}^{N}\hat{u}_{n1j}}{N},\ \gamma_{njk} = \frac{\sum_{t=1}^{T}\hat{v}_{ntjk}}{\sum_{t=1}^{T}\sum_{k=1}^{M}\hat{v}_{ntjk}} \qquad (18)
$$

Once the MCEM algorithm converges and the parameter estimates are available, the random effect estimates can be calculated by the expectation of b_n for each customer,

$$
\bar{b}_n = \sum_{l=1}^{B}h(b_n^l|L_n)b_n^l \qquad (19)
$$

Then the expected load can be estimated using (2). The state probabilities at each time step can be estimated by calculating the filtered probabilities of regimes for each customer.

C. PV System Performance Model and Parameter Estimation

In this subsection, the physical PV system performance model is presented first. Then, the estimation method of the technical parameters of a solar PV system is described.

A PV system performance model, g calculates the AC output of a solar PV system with the relevant weather data and the solar PV system specifications. The model used in our

study is based on the PV performance modeling collaborative [5] from Sandia National Laboratory and PVWatts from NREL [29], [4]. The inputs to the model include the solar PV system specifications (system nameplate DC rating in kW P_{dc0}, tilt angle θ_t and azimuth angle θ_{az} of the PV array, nominal efficiency of the inverter η_{nom}, and loss of the PV system l), weather-related data (temperature and wind speed), and solar irradiance data (direct normal irradiance, diffuse horizontal irradiance, and global horizontal irradiance). The solar irradiance data incorporates The solar PV performance model has four main submodels, the radiation submodel, the thermal submodel, module submodel, and the inverter submodel.

The radiation submodel translates the solar irradiation data into the energy incident on the PV module cover. First, solar position algorithms [30] can be used to calculate the sun position from the date, time, and geographic position data. The irradiance incident on the plane of the array (E_{poa}) is defined as follows.

$$E_{POA} = E_b + E_g + E_d \tag{20}$$

where E_b is the POA beam component, E_g is the POA ground-reflected component, and E_d is the POA sky-diffuse component. The sun position data, albedo, PV array orientation, solar irradiance data, and array tracking mode are used to calculate, E_b, E_g, and E_d and hence plane of array irradiance, E_{poa}. The solar irradiance data is collected from National Solar Radiation Database (NSRDB). The physical solar model (PSM) employed by NSRDB utilizes the cloud physical and optical properties to produce cloudy-sky solar radiation [31]. For a fixed system, the angle of incidence is calculated following the standard geometrical calculation. Next, to account for reflection losses on the module cover, a correction is applied for incidence angles greater than 50^o using the polynomial correction from [5] and the transmitted irradiance, E_{tr} is calculated.

The thermal submodel calculates the operating cell temperature, T_{cell} using the total incident POA irradiance E_{poa}, wind speed, and dry bulb temperature following the Sandia cell temperature model.

The module submodel computes the DC output power P_{dc} by using the DC nameplate rating P_{dc0}, cell temperature T_{cell}, transmitted POA irradiance E_{tr}, and loss of the PV array l. The loss is modeled as a percentage of DC energy. It includes the impacts of soiling, shading, mismatch, wiring, system age, etc. The reference cell temperature E_0 is 25^oC, temperature coefficient ζ is $-0.5\%/^oC$, and reference irradiance T_0 is $1000\ W/m^2$.

$$P_{dc} = (1-l) \times \frac{E_{tr}}{E_0} P_{dc0} \left[1 + \zeta \left(T_c - T_0\right)\right] \tag{21}$$

The inverter submodel calculates the AC power output of the PV system P_{ac} using P_{dc}. The AC nameplate rating of the inverter (P_{ac0}) is calculated by $P_{ac0} = \frac{P_{dc0}}{\text{DC-to-AC ratio}}$. The nominal efficiency of the inverter is defined as the ratio of the AC nameplate rating of the inverter P_{ac0} and the inverter DC rating $P_{dc0,inv}$. Then, the inverter efficiency η can be calculated following [4] and P_{ac} can be calculated as follows:

$$P_{ac} = \begin{cases} P_{ac0} & \text{if} P_{dc} \geq P_{dc0,inv} \\ \eta P_{dc} & \text{if} P_{dc} < P_{dc0,inv} \end{cases} \tag{22}$$

Next, the description of how to estimate the technical parameters of a solar PV system with multiple strings of solar panels facing different directions is provided. Although most residential houses have a single south-facing solar panel to maximize solar energy production over the year, many houses have multiple strings of solar panels often facing south and west. The west-facing solar installations may receive additional local government incentives because they produce more energy during the peak demand hours in the late afternoon. Our proposed solar PV system technical parameter estimation algorithm accounts for the cases of both single and multiple strings of solar panels.

Let $\mathbf{\Phi}$ denote a tensor of order three with dimensions $N \times M \times R$ representing the technical parameters of M panels of N customers. Let R denote the dimension of a single solar panel's parameters and g_t denote the solar PV system's generation at time t. The technical parameters of the m-th solar panel of the customer n is denoted by $\mathbf{\Phi}_{mn} = [P_{dc0}, \theta_t, \theta_{az}, \eta_{nom}, l]$, which includes the DC rating, array tilt angle, array azimuth angle, nominal inverter efficiency, and loss of the PV array, respectively. The inputs to the solar PV system parameters estimator are the estimated solar PV generation S_{nt} of a customer n for time $t = 1, 2, ..., T$. Solar PV system parameters are estimated by minimizing the sum of squared error between the estimated solar PV generation S_{nt} and the calculated solar generation $g\left(\mathbf{\Phi}_{mn}\right)$ from M strings of solar panels of customer n.

$$\begin{aligned} &\arg\min_{\mathbf{\Phi}_{mn}} \sum_{t=1}^{T} \left(S_{nt} - \sum_{m=1}^{M} g_t\left(\mathbf{\Phi}_{mn}\right) \right)^2 \\ &\text{subject to } \mathbf{\Phi}_{min} \leq \mathbf{\Phi}_{mn} \leq \mathbf{\Phi}_{max} \end{aligned} \tag{23}$$

where T is the time series length. $\mathbf{\Phi}_{min}$ and $\mathbf{\Phi}_{max}$ denote the lower and upper limits of the solar PV system technical parameters. The highly nonlinear nature of the solar PV system performance model makes Equation (23) a nonlinear optimization problem, which can be solved by an interior-point algorithm.

D. Summary of Net Load Disaggregation Algorithm

It is proposed to disaggregate net load measurements $\boldsymbol{NL}$ into electric load $\hat{\mathbf{L}}$ and solar PV generation $\hat{\mathbf{S}}$ for a group of residential customers by integrating the physical solar PV system performance model introduced in Section III-C and the statistical MHMM introduced in Section III-A and III-B. The detailed process for joint net load disaggregation of a community of customers is shown in Algorithm . The MCEM estimation of the MHMM parameters can be computationally intensive. Therefore, it is needed to have a good initial estimate of load $\hat{\mathbf{L}}^{(0)}$ as the starting point of the iterative algorithm. The initial estimates for electric load $\hat{\boldsymbol{L}}_{HMM}$ and PV system parameters $\hat{\mathbf{\Phi}}_{HMM}$ of all customers are set to be the estimates based on the iterative algorithm with HMM regression [19].

For each iteration i, an MHMM is fitted to the estimated load $\hat{\mathbf{L}}^{(i-1)}$. There are J_i sets of initial MHMM parameter estimates $\mathbf{\Theta}_0^{(i,j)}$ with $J_i = 2$ for the first iteration and $J_i = 3$ for the subsequent iterations. The first and second sets of initial MHMM parameter estimates $\mathbf{\Theta}_0^{(i,1)}$ and $\mathbf{\Theta}_0^{(i,2)}$ are obtained from fitting HMM regression to the estimated electric load $\hat{\mathbf{L}}_n^{(i-1)}$ of each customer n where $\mathbf{\Psi}_0^{(i,j)}$ are the initial estimates for the HMM regression parameters. Then, $\mathbf{\Psi}_0^{(i,1)}$ is set to be equal to be the multiple linear regression model parameters on $\boldsymbol{L}_n^{(i-1)}$ for N customers and their negatives for states $k = 1, 2$. To set $\mathbf{\Psi}_0^{(i,2)}$, the HMM regression is run with ten sets of initial values obtained by adding random noise to $\mathbf{\Psi}_0^{(i,1)}$ and then choose the initial value set that yields the maximum log likelihood for the HMM regression. The third set of initial MHMM parameter estimates $\mathbf{\Theta}_0^{(i,3)}$ is equal to the MHMM parameters estimated in the previous iteration $\mathbf{\Theta}^{(i-1)}$. Now, the MHMM parameter estimates $\mathbf{\Theta}^{(i,j)}$ and the updated load estimates $\hat{\boldsymbol{L}}^{(i,j)}$ for $j = 1 \dots J_i$ are obtained.

For the jth set of initial value, the estimated solar PV generation $\hat{\boldsymbol{S}}_n^{(i,j)} = \hat{\boldsymbol{L}}_n^{(i,j)} - \boldsymbol{NL}_n$ is calculated for each customer n and estimate the technical parameters of M solar PV panels $\mathbf{\Phi}_n^{(i,j)}$ by solving a constrained optimization following Equation (23). The inputs to the optimization problem include the estimated solar PV generation $\hat{\boldsymbol{S}}_n^{(i,j)}$, the solar PV system performance model g, and the initial solar PV system parameter estimates $\mathbf{\Phi}_n^{(i-1)}$. The solar PV generation $\hat{\boldsymbol{S}}_n^{(i,j)}$ for each customer n can then be updated by feeding the estimated solar PV parameters $\mathbf{\Phi}_n^{(i,j)}$ into the PV system performance model g. With the updated estimates for the load and solar generation, the net load $\hat{\mathbf{NL}}_n^{(i,j)}$ and the average MSE of the customers' net load $E^{(i,j)}$ can be calculated for the jth set of initial MHMM parameter estimates.

At the end of the i-th iteration, among the J_i sets of outputs, the one that corresponds to the lowest average MSE of the net load $E^{(i,j)}$ is calculated. The corresponding index of the initial MHMM parameter estimates is denoted as j^*. In other words, at the end of iteration i, the following variables are updated: $\hat{\boldsymbol{S}}^{(i)} = \hat{\boldsymbol{S}}^{(i,j^*)}$, $\mathbf{\Phi}^{(i)} = \hat{\mathbf{\Phi}}^{(i,j^*)}$ and $\mathbf{\Theta}^{(i)} = \hat{\mathbf{\Theta}}^{(i,j^*)}$, and $E^{(i)} = E^{(i,j^*)}$. The load estimate is then updated as $\hat{\boldsymbol{L}}^{(i)} = \boldsymbol{NL} + \hat{\mathbf{S}}^{(i)}$. The iterative algorithm continues until the average MSE of net load, $E^{(i)}$ converges or the maximum number of iterations is reached. Finally, the solution that yields the lowest average MSE for the customers' net loads is selected.

Post-disaggregation Adjustment: To further improve the net load disaggregation algorithm, the post-disaggregation adjustment is performed by enforcing the constraint that the electric load minus solar generation must equal to the net load measurement. The following optimization problem inspired from [8] is solved for each customer n using the disaggregated signals $\hat{\boldsymbol{L}}_n$ and $\hat{\boldsymbol{S}}_n$.

$$\begin{aligned} &\underset{L_{nt}, S_{nt}}{\arg\min} \sum_{t=0}^{T} \mu\Big(L_{nt} - \hat{L}_{nt}\Big)^2 + \omega\Big(S_{nt} - \hat{S}_{nt}\Big)^2 \\ &\text{subject to } S_{nt} \geq 0, \quad L_{nt} - S_{nt} = NL_{nt} \end{aligned} \tag{24}$$

Here, μ and ω are parameters that denote the weights for the errors in the load and solar generation estimates. μ and ω can be calculated as the inverse of the variance of the errors of the load and solar generation estimates.

$$\mu = 1/Var\left(\varepsilon_{Load}\right), \quad \omega = 1/Var\left(\varepsilon_{PV}\right) \tag{25}$$

Since the load and solar PV generation data are not available, the variance of the errors is estimated by the load and solar PV generation from steps 4 and 7 of the net load disaggregation algorithm from [19] at step 1 of Algorithm 1.

Algorithm Algorithm for joint net load disaggregation of N customers and estimation of their solar PV parameters

Input: A matrix of net load of customers, $\mathbf{NL}$
Output: Matrices of estimates for load $\hat{\mathbf{L}}$ and solar generation $\hat{\mathbf{S}}$, and a tensor of PV system parameters $\mathbf{\Phi}$

1: Initialize the matrix of load $\hat{\mathbf{L}}^{(0)} = \hat{\mathbf{L}}_{\mathbf{HMM}}$ with the load estimates from [19]. Initialize the tensor of PV parameters, $\mathbf{\Phi}^{(0)} = \hat{\mathbf{\Phi}}_{HMM}$ with estimates from [19].
2: **for** i=1 to maxiter **do**
3: Determine J_i sets of initial MHMM parameter estimates $\mathbf{\Theta}_0^{(i,j)}$. Set them equal to HMM regression model parameters $\mathbf{\Theta}_{HMM}^{(i,j)}$ based on $\hat{\mathbf{L}}^{(i-1)}$ with initial HMM regression parameters $\mathbf{\Psi}_0^{(i,j)}$ for $j = 1 \dots J_{i-1}$. Set $\mathbf{\Theta}_0^{(i,J_i)} = \mathbf{\Theta}_{HMM}^{(i,J_i)}$ if $i = 1$, $\mathbf{\Theta}_0^{(i,J_i)} = \mathbf{\Theta}^{(i-1)}$ for $i > 1$.
4: **for** j=1 to J_i **do**
5: Fit MHMM, $f(\boldsymbol{x}, \mathbf{\Theta})$, to $\hat{\mathbf{L}}^{(i-1)}$ with initial parameter estimates $\mathbf{\Theta}_0^{(i,j)}$ and calculate $\mathbf{\Theta}^{(i,j)}$
6: Update load estimates, $\hat{\mathbf{L}}^{(i,j)} = f\left(\boldsymbol{x}, \mathbf{\Theta}^{(i,j)}\right)$
7: Update solar generation, $\hat{\boldsymbol{S}}^{(i,j)} = \hat{\mathbf{L}}^{(i,j)} - \mathbf{NL}$
8: **for** customers n=1 to N **do**
9: Determine ${\mathbf{\Phi}_n}^{(i,j)}$ from Equation (23) using ${\mathbf{\Phi}_n}^{(i-1)}$ as initial value
10: Update solar generation, $\hat{\boldsymbol{S}}_n^{(i,j)} = g\left({\mathbf{\Phi}_n}^{(i,j)}\right)$
11: Estimate net load, $\hat{\mathbf{NL}}_n^{(i,j)} = \hat{\mathbf{L}}_n^{(i,j)} - \hat{\mathbf{S}}_n^{(i,j)}$ and MSE of the estimated net load, $E_n^{(i,j)}$
12: **end for**
13: Calculate $E^{(i,j)} = \frac{1}{N} \sum_{n=1}^{N} E_n^{(i,j)}$
14: **end for**
15: Determine $j^* = \underset{j}{argmin}\, E^{(i,j)}$
16: Update $\hat{\mathbf{S}}^{(i)} = \hat{\mathbf{S}}^{(i,j^*)}$, $\mathbf{\Phi}^{(i)} = \mathbf{\Phi}^{(i,j^*)}$, $\mathbf{\Theta}^{(i)} = \mathbf{\Theta}^{(i,j^*)}$, and $E^{(i)} = E^{(i,j^*)}$
17: Update load estimates, $\hat{\mathbf{L}}^{(i)} = \mathbf{NL} + \hat{\mathbf{S}}^{(i)}$
18: **if** $\left|E^{(i)} - E^{(i-1)}\right| \leq \varepsilon$ Break **end if**
19: **end for**
20: Determine $i^* = \underset{i}{argmin}\, E^{(i)}$
21: Calculate $\hat{\mathbf{L}} = \hat{\mathbf{L}}^{(i^*)}$, $\hat{\mathbf{S}} = \hat{\mathbf{S}}^{(i^*)}$, and $\mathbf{\Phi} = \mathbf{\Phi}^{(i^*)}$
22: **return** $\hat{\mathbf{L}}$, $\hat{\mathbf{S}}$, and $\mathbf{\Phi}$

IV. Numerical Study

A. Dataset for Numerical Study

The energy data of 193 residential customers in Austin, Texas gathered by Pecan Street Inc. [20] are used to validate our proposed net load disaggregation algorithm. The dataset

includes 15-minute interval net load, electric load, and solar PV generation data. The tilt and azimuth angle information are not available. However, the solar panel's DC rating data are reported for 90% customers which can be used for validation. The study period is selected as 10/03/2015-10/30/2015 to compare our estimates with [14]. The solar irradiance and weather-related data is collected with 4×4 km spatial and 30-minute temporal resolutions from the National Solar Radiation Database [32]. It is converted into 15-minute interval data by linear interpolation.

The approximate longitude and latitude of Austin, Texas (30.29^oN,-97.69^oE) is used as a common proxy location for all customers as their exact locations are not available. Similarly, the same weather variables are used for net load disaggregation for all of the customers. Since most of the residential rooftop solar PV systems use fixed array, it is assumed that none of the residential solar PV systems in this study have tracking system. Note that if the solar PV system's tracking mode information is available, then the incorporation of either 1-axis or 2-axis tracking in the PV system performance model is straightforward.

B. Experimental Setup

The proposed net load disaggregation method is implemented following Algorithm under two scenarios. In the first scenario, it is assumed that every customer's solar PV system only has one string of solar panels, which means $M = 1$ and the number of solar PV technical parameters $R = 5$. In the second scenario, it is recognized that in the Pecan Street dataset, 71 out of 193 customers have two strings of solar panels facing different directions. Thus, it is assumed that these customers have two solar panels with potentially different DC ratings but the same tilt angle, nominal inverter efficiency, and loss. In this scenario $M = 2$ and $R = 7$. The rest of the customers have a single string of solar panels. Note that the data indicating one or two strings of solar panels may be erroneous. Thus, the estimated total effective DC sizes for the 71 customers [19] is compared for both one and two strings of solar panel installation assumption. If the difference between the outputs under the two assumptions is significant, then it is still assumed that the customer have a single string of solar panels. This is because customers often have larger solar panels installed on the main roof. The secondary roof usually can only support smaller solar panels. Thus, the difference between the total estimated DC ratings of solar PV systems is typically not significant. When a large difference occurs, it might suggest that the proposed iterative algorithm with two strings of solar panels setup has converged to a local optimum. This is possible given that the dimensionality of the search space is much larger for the two-string setup. Therefore, when a large difference in the DC size estimates is encountered, a single string of solar panels is assumed. Finally, 64 out of 71 customers are identified to have two strings of solar panels.

To strike a balance between the computational efficiency and accuracy, the number of random samples B of the MCEM is selected to be 500. The tolerance for the convergence of the MCEM algorithm is set as $\epsilon' = 0.001$. The initial parameter estimates of the MHMM are obtained from fitting the HMM regression [19] to the individual customer's electric load data. Note that, the HMM regression is fitted using the EM algorithm [33] instead of MS regress package [34] to make the parameter estimation procedure comparable to the MCEM algorithm.

The feasible ranges of solar PV parameters $P_{dc}, \theta_T, \theta_{az}, \eta_{nom}, l$ are set as 1-15 kW, $5^o - 50^o$, $0^o - 360^o$, $0.92 - 0.99$ and $9\% - 40\%$, respectively [19], [35], [36]. The DC-to-AC ratio is fixed at 1.1. When testing the benchmark algorithm to perform net load disaggregation for individual customers with HMM regression [19], 8 initial solar PV system technical parameter sets are chosen for the single string of solar panel scenario by gradually increasing P_{dc0} in 7 steps from 1 kW to 8 kW. The other initial parameters $[\theta_T, \theta_{az}, \eta_{nom}, l]$ were set at their most common values $25^o, 180^o, 0.96$, and 14%, respectively. For the scenario with two strings of solar PV panels, 64 initial PV parameter sets are obtained by enumerating the two DC sizes from 1 to 8 kW. The initial estimates for θ_{az} are set at 180^o and 270^o. The tolerance for converge criteria of Algorithm 1 is set as $\varepsilon = 0.001$. The performance of the proposed and benchmark algorithms is evaluated with three commonly used error metrics: mean squared error (MSE), mean absolute squared error (MASE), and coefficient of variation (CV) [19].

C. Result and Analysis

In this section, the performance of our proposed Algorithm is compared with four other state-of-the-art net load disaggregation algorithms including our earlier work [19] that uses HMM regression for individual load estimation. In addition, the benefits of considering multiple strings of solar panels facing different directions are also evaluated in the numerical study.

1) Comparison with state-of-the-art net load disaggregation algorithms: The four state-of-the-art benchmark net load disaggregation algorithms are as follows: the unsupervised consumer mixture model [14], the SunDance algorithm [17], the algorithm proposed in [15], and our earlier behind-the-meter solar generation estimation work that uses HMM regression to model individual customers' electric load. The details of the experimental setup of the consumer mixture model and the SunDance model can be found in [19]. Since method C yields the best results among the four methods proposed in [15] for this dataset, it is used as one of the benchmarks. This method assumes electric load to be piecewise constant and models the solar PV generation by a linear combination of the solar irradiance.

The MSE, MASE, and CV for the load and solar generation estimates of the proposed algorithm and the four benchmark algorithms are reported in Table I. As shown in the table, our proposed algorithm which seamlessly integrates the physical solar PV performance model with statistical MHMM significantly outperforms all benchmark algorithms. Our proposed method reduces the MSE of the solar PV generation estimates by 67% and 33% from the consumer mixture model [14] and our earlier work that uses HMM regression [19].

TABLE I: Comparison of various net load disaggregation methods

Error Metric	Variable	MHMM (solar panel scenario 1)	HMM reg. (solar panel scenario 1)	MHMM (solar panel scenario 2)	HMM reg. (solar panel scenario 2)	Consumer Mixture Model	SunDance Model	Algorithm by [15]
MSE	Solar	**0.13**	0.19	**0.12**	0.18	0.37	0.54	0.42
	Load	**0.13**	0.19	**0.12**	0.18	0.37	0.49	0.28
MASE	Solar	**2.13**	2.61	**2.11**	2.58	3.85	3.74	4.44
	Load	**0.43**	0.48	**0.42**	0.48	0.74	0.81	0.56
CV	Solar	**0.47**	0.58	**0.45**	0.57	0.77	0.85	0.78
	Load	**0.29**	0.33	**0.28**	0.32	0.46	0.57	0.43

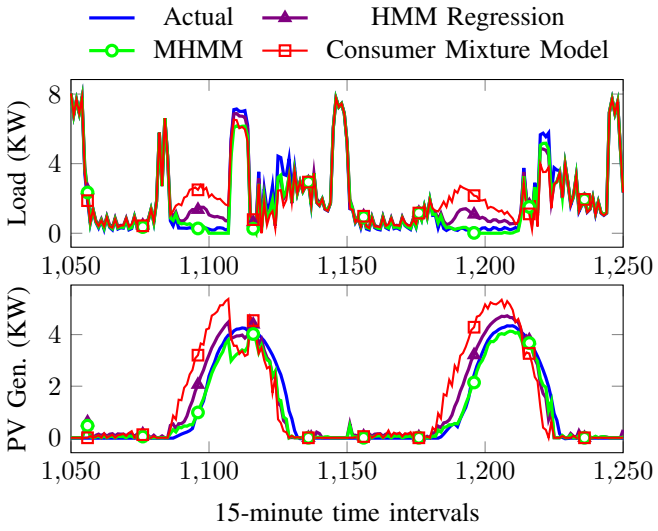

Fig. 2: Comparison of disaggregated load and solar PV generation with actual values for a customer from October 11 to October 12, 2015

There are two reasons why our proposed algorithm yields better results. First, the high fidelity physical PV system performance model incorporated in our proposed algorithm can better capture the nonlinear relationships between the solar PV generation, solar PV system specifications, and weather data. Second, MHMM is better suited to emulate the customers' energy behavior in different regimes. This is especially evident during the low load periods when the customer may be absent as depicted in Fig. 2. As shown in the figure, the MHMM follows the actual load much more closely than the benchmark algorithms during the low load periods, which leads to better solar PV generation estimation. Therefore, the comparative advantage of our proposed model is more pronounced for customers who are absent from home for a long period. In the numerical study, 24 out of 193 customers are suspected to be absent from their residence for an extended period. For these customers, our proposed model with MHMM regression reduces the MSE by 71% compared to the consumer mixture model.

2) Comparison between MHMM and HMM regression: The proposed net load disaggregation algorithm with MHMM outperforms the algorithm with HMM regression by 33% in terms of MSE of load estimates. The MHMM provides a more accurate estimation of load compared to the HMM regression. The proposed net load disaggregation method is an iterative method that estimates the load and PV generation parameters in turn. An improved load estimate at step 6 of the algorithm leads to an improved estimate of solar PV technical parameters, which in turn leads to an improved solar PV generation estimate. Our proposed net load disaggregation algorithm with MHMM outperforms the algorithm with HMM regression by 33% in terms of MSE of solar PV generation estimates.

The improved load estimate by MHMM can be attributed to the following factors. First, MHMM jointly models the electric load of customers in a community while capturing the individual heterogeneity by incorporating the individual-specific random effects. Second, MHMM provides a more efficient estimation of the fixed model parameters. Third, MHMM enables information sharing by the population level effect and the random effects components follow a normal distribution. As a result, it can be observed in this study that the MHMM algorithm yields more pronounced improvement for customers with unreliable intercept estimates in the HMM regression. The MHMM algorithm corrects such problems by moving these outliers toward the mean intercept. As shown in Fig. 3, the histogram of the intercepts from the HMM regression is skewed to the right with 54 customers having a very large intercept (> 1). The intercept estimates of the MHMM for these customers have been shifted toward the mean. The improvement in MSE of solar PV generation estimates is 50% for these customers and only 29% for the rest of the customers. The histogram of the MSE of the solar PV generation estimate for the HMM and MHMM regression algorithms are shown in Figure 4. It can be observed that the percentage of customers with lower MSE of solar PV generation estimates of the MHMM algorithm is much higher than that of the HMM algorithm. For example, the percentage of customers with MSE of solar PV generation estimates smaller than $0.1\ kW$ is only 29% for the HMM regression. By adopting the proposed MHMM regression, this percentage increases to 45%.

The net load disaggregation algorithm with both HMM regression and MHMM provides accurate solar PV generation estimates both in sunny and cloudy days. In this study, October 21 to October 26 are cloudy days with low DNI. The average MSE of solar PV generation estimates is $0.10\ KW$ for the algorithm with MHMM regression. The average MSE of solar PV generation estimates is $0.12\ KW$ for the algorithm with HMM regression. For both algorithms, the MSE for the cloudy days is lower than the overall average MSE. As shown in figure 5, the PV generation estimates of the customer with the median MSE of solar PV generation from HMM regression and MHMM closely follow the actual solar PV generation.

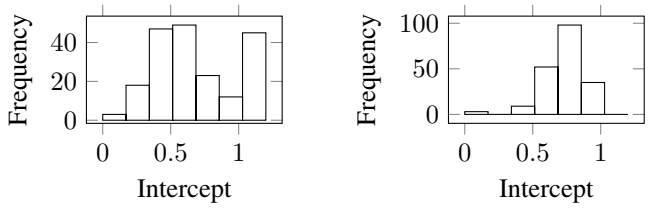

Fig. 3: Histogram of the intercepts from the HMM regression (left, $std = 0.49$) and the MHMM (right, $std = 0.14$)

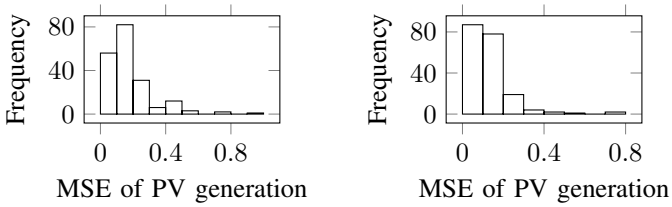

Fig. 4: Histogram of the MSE of solar PV generation estimates for the HMM regression (left) and the MHMM (right)

3) Advantage of modeling multiple strings of solar panels: By considering the possibility of having multiple strings of solar panels facing different directions, our proposed algorithm in scenario 2 further improves the estimation accuracy when compared to scenario 1. This modeling flexibility better captures the physical configurations of real-world solar PV systems. As shown in Table I, a 7% reduction in MSE of the solar generation estimates is achieved in scenario 2 over scenario 1.

4) Accuracy of the PV array technical parameters: The ground truth tilt and azimuth angle of the solar PV installations are not available. The DC rating of solar PV panels is available for 90% of the customers. The performance of the proposed model in estimating the DC size of the solar PV systems is illustrated in Fig. 6. As shown in the figure, the estimated solar DC ratings and the actual are quite similar. The MAPE of the estimated solar DC ratings is 20% for the algorithm with HMM regression and 18% for the algorithm with MHMM.

5) Computation time and scalability: The computation time for the net load disaggregation algorithm with MHMM is 6 minutes per hour of net load data using an Intel core i9 processor. The computation time is measured for the case where the number of initial MHMM parameters J_i is 3. The number of random samples B equals 500. The tolerance for the convergence of the MCEM algorithm ϵ' is set as 0.001 and the number of customers is 193. However, 45% time is spent on solving the HMM regression problem which is used as initial parameters for the MHMM at step 3 of the algorithm. This process can be parallelized to save computation time since the HMM regression is estimated for each customer separately. To perform net load disaggregation for a large number of customers, one could first separate all customers into different communities based on geographical location. Then the net-load disaggregation problem can be solved for different communities in parallel, which makes the proposed algorithm extremely scalable.

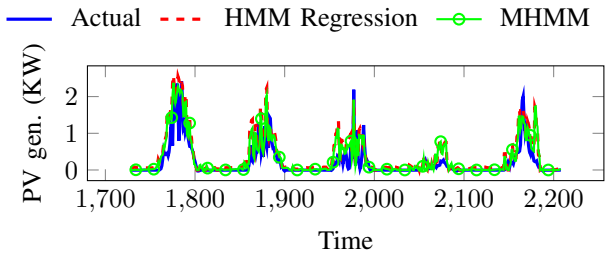

Fig. 5: Comparison of disaggregated solar PV generation with actual values for the customer with median MSE of solar PV generation for the cloudy days from October 18 to October 23

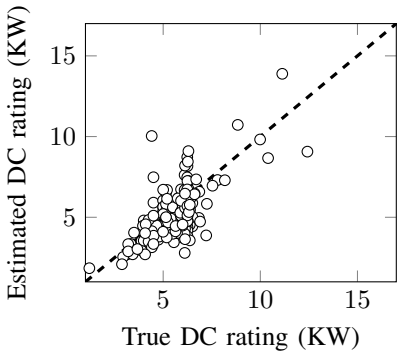

Fig. 6: Comparison of true and estimated DC rating of PV array

V. Conclusion

An unsupervised algorithm is developed to disaggregate the observed net load signals of a group of residential customers with behind-the-meter solar PV systems into unknown solar PV generation and electric load. The iterative algorithm synergistically combines a physical PV system performance model for individual solar PV generation estimation with a statistical mixed hidden Markov model for joint load estimation. The mixed hidden Markov model not only models the general load consumption behavior of the entire community but also captures the individual differences with the random effects. Furthermore, the high fidelity PV system performance model considers real-world configurations with multiple strings of solar panels facing different directions. These technical advancements result in a significant reduction in the estimation error of the solar PV generation from the state-of-the-art net load disaggregation algorithms. Once the estimated solar PV systems' technical parameters are obtained with the proposed algorithm, online estimation of behind-the-meter solar PV generation becomes feasible with real-time solar irradiance data.

Several interesting extensions of our proposed algorithm can be explored in the future. First, semi-parametric mixed hidden Markov model can be developed to further improve the computational efficiency of the proposed algorithm. Second, in the current model, the customer-specific random effect and its variance are assumed to be independent of the hidden states. The mixed hidden Markov model can be improved by assuming that the random effect variance depends on the

hidden states. Third, a robust version of the proposed net-load disaggregation algorithm can be developed to improve estimation accuracy in the presence of outliers. Finally, a mixed effect model to jointly estimate load in several communities with a community-specific random effect along with the customer-specific random effect will be of interest.

REFERENCES

[1] B. Tyra, "Electric Power Monthly," US Energy Information Administration, Tech. Rep., 2020.

[2] M. Bolinger and J. Seel, "Utility-scale solar 2014: An empirical analysis of project cost, performance, and pricing trends in the United States," Lawrence Berkeley National Lab., United States, Tech. Rep., 2015.

[3] R. Seguin, J. Woyak, D. Costyk, J. Hambrick, and B. Mather, "High-penetration pv integration handbook for distribution engineers," National Renewable Energy Lab.(NREL), United States, Tech. Rep., 2016.

[4] A. P. Dobos, "PVWatts version 5 manual," National Renewable Energy Laboratory, Golden, CO, USA, Tech. Rep., 2014.

[5] J. S. Stein, "The photovoltaic performance modeling collaborative (PVPMC)," in *38th IEEE Photovoltaic Specialists Conference*. IEEE, 2012, pp. 003 048–003 052.

[6] M. Wytock and J. Z. Kolter, "Contextually supervised source separation with application to energy disaggregation," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, ser. AAAI'14. AAAI Press, 2014, pp. 486–492.

[7] F. Bu, K. Dehghanpour, Y. Yuan, and Z. Wang, "A data-driven game-theoretic approach for behind-the-meter pv generation disaggregation," *arXiv*, 2019.

[8] E. C. Kara, M. Tabone, C. Roberts, S. Kiliccote, and E. M. Stewart, "Estimating behind-the-meter solar generation with existing measurement infrastructure: Poster abstract," in *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments*. ACM, 2016, pp. 259–260.

[9] M. Tabone, S. Kiliccote, and E. C. Kara, "Disaggregating solar generation behind individual meters in real time," in *Proceedings of the 5th Conference on Systems for Built Environments*. ACM, 2018, pp. 43–52.

[10] E. C. Kara, C. M. Roberts, M. Tabone, L. Alvarez, D. S. Callaway, and E. M. Stewart, "Disaggregating solar generation from feeder-level measurements," *Sustainable Energy, Grids and Networks*, vol. 13, pp. 112–121, Mar. 2018.

[11] H. Shaker, H. Zareipour, and D. Wood, "Estimating power generation of invisible solar sites using publicly available data," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2456–2465, Sep. 2016.

[12] M. Sun, T. Zhang, Y. Wang, G. Strbac, and C. Kang, "Using Bayesian deep learning to capture uncertainty for residential net load forecasting," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 188–201, Jan. 2020, conference Name: IEEE Transactions on Power Systems.

[13] K. Li, F. Wang, Z. Mi, M. Fotuhi-Firuzabad, N. Duić, and T. Wang, "Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation," *Applied Energy*, vol. 253, p. 113595, Nov. 2019.

[14] C. M. Cheung, W. Zhong, C. Xiong, A. Srivastava, R. Kannan, and V. K. Prasanna, "Behind-the-meter solar generation disaggregation using consumer mixture models," in *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, Oct. 2018, pp. 1–6.

[15] F. Sossan, L. Nespoli, V. Medici, and M. Paolone, "Unsupervised disaggregation of photovoltaic production from composite power flow measurements of heterogeneous prosumers," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 9, pp. 3904–3913, Sep. 2018.

[16] W. Stainsby, D. Zimmerle, and G. P. Duggan, "A method to estimate residential PV generation from net-metered load data and system install date," *Applied Energy*, vol. 267, p. 114895, Jun. 2020.

[17] D. Chen and D. Irwin, "SunDance: Black-box behind-the-meter solar disaggregation," in *Proceedings of the Eighth International Conference on Future Energy Systems*. ACM, 2017, pp. 45–55.

[18] Y. Wang, N. Zhang, Q. Chen, D. S. Kirschen, P. Li, and Q. Xia, "Data-driven probabilistic net load forecasting with high penetration of behind-the-meter PV," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3255–3264, May 2018.

[19] F. Kabir, N. Yu, W. Yao, R. Yang, and Y. Zhang, "Estimation of behind-the-meter solar generation by integrating physical with statistical models," in *2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 2019, pp. 1–6.

[20] C. Holcomb, "Pecan street inc.: A test-bed for NILM," in *International Workshop on Non-Intrusive Load Monitoring*, 2012.

[21] M. Fridman, "Hidden Markov model regression," Institute of Mathematics, University of Minnesota, Tech. Rep., 1993.

[22] R. M. Altman, "Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 201–210, 2007.

[23] M. T. Hagan and S. M. Behr, "The time series approach to short term load forecasting," *IEEE Transactions on Power Systems*, vol. 2, no. 3, pp. 785–791, Aug 1987.

[24] S. L. DeRuiter, R. Langrock, T. Skirbutas, J. A. Goldbogen, J. Calambokidis, A. S. Friedlaender, B. L. Southall *et al.*, "A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure," *The Annals of Applied Statistics*, vol. 11, pp. 362–392, 2017.

[25] F. Chaubert-Pereira, Y. Guédon, C. Lavergne, and C. Trottier, "Markov and semi-Markov switching linear mixed models used to identify forest tree growth components," *Biometrics*, vol. 66, no. 3, pp. 753–762, 2010.

[26] J. C. Jackson, P. S. Albert, and Z. Zhang, "A two-state mixed hidden Markov model for risky teenage driving behavior," *The annals of applied statistics*, vol. 9, no. 2, p. 849, 2015.

[27] A. Maruotti, "Mixed hidden Markov models for longitudinal data: An overview," *International Statistical Review*, vol. 79, pp. 427–454, 2011.

[28] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis*. Cambridge university press, 1998.

[29] A. P. Dobos, "PVWatts version 1 technical reference," National Renewable Energy Laboratory, Golden, CO, USA, Tech. Rep., 2013.

[30] I. Reda and A. Andreas, "Solar position algorithm for solar radiation applications," *Solar Energy*, vol. 76, no. 5, pp. 577–589, 2004.

[31] M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, "The National Solar Radiation Data Base (NSRDB)," *Renewable and Sustainable Energy Reviews*, vol. 89, pp. 51–60, Jun. 2018.

[32] ——, "The National Solar Radiation Data Base (NSRDB)," *Renewable and Sustainable Energy Reviews*, vol. 89, pp. 51–60, 2018.

[33] I. L. MacDonald and W. Zucchini, *Hidden Markov and other models for discrete-valued time series*. Chapman & Hall, 1997.

[34] M. Perlin, "MS_regress - the MATLAB package for Markov regime switching models," *Available at SSRN 1714016*, 2015.

[35] C. D. G. Statistics, "The California Solar Initiative - CSI working data set," 2019, https://www.californiadgstats.ca.gov/downloads/.

[36] B. Marion, J. Adelstein, K. e. Boyle, H. Hayden, B. Hammond, T. Fletcher, B. Canada, D. Narang, A. Kimber, L. Mitchell *et al.*, "Performance parameters for grid-connected PV systems," in *Conference Record of the Thirty-first IEEE Photovoltaic Specialists Conference, 2005.* IEEE, 2005, pp. 1601–1606.