Physics-Informed Sparse Gaussian Process for Probabilistic Stability Analysis of Large-Scale Power System With Dynamic PVs and Loads

Ketian Ye, *Student Member, IEEE*, Junbo Zhao, *Senior Member, IEEE*, Nan Duan, *Senior Member, IEEE*, and Yingchen Zhang, *Senior Member, IEEE*

***Abstract*—This paper proposes a physics-informed sparse Gaussian process (SGP) for probabilistic stability assessment of large-scale power systems in the presence of uncertain dynamic PVs and loads. The differential and algebraic equations considering uncertainties from dynamic PVs and loads are reformulated to a nonlinear mapping relationship that allows the application of SGP. Thanks to the nonparametric characteristic of Gaussian process, the proposed framework does not require distributions of uncertain inputs and this distinguishes it from existing approaches. As the original Gaussian process is not scalable to large-scale systems with high dimensional uncertain inputs, this paper develops the SGP with a stochastic variational inference technique. It leads to approximately two orders of complex reduction. A data preprocessing step is also introduced to tackle the coexistence of stable and unstable cases by sample clustering and constructing separate SGPs. The probabilistic transient stability index is analyzed to assess system stability under different uncertain dynamics loads and PVs. Comparisons are performed with the sampling-based, the polynomial chaos expansion-based, and traditional Gaussian process-based methods on the modified IEEE 118-bus and Texas 2000-bus systems under various scenarios, including different levels of uncertainties and the existence of nonlinear correlations among dynamic PVs. The impacts of data quality and quantity issues are also investigated. It is shown that the proposed SGP achieves significantly improved computational efficiency while maintaining high accuracy with a limited number of data.**

***Index Terms*—Dynamic PVs, nonlinear correlations, physics-informed, power system dynamics, probabilistic transient stability, sparse Gaussian process, uncertainty quantification.**

I. Introduction

THE impacts of increased uncertainties from inverter-based resources (IBRs), such as PVs and the variability of highly flexible demands have challenged the power system secure operation [1]. The traditional power system transient stability analysis assesses the system vulnerabilities under different contingencies without considering these uncertainties. For example, some deterministic approaches considering the worst-case scenarios are developed but this may lead to overestimated or underestimated solutions [2]. To address that, probabilistic methods are proposed and they can obtain detailed statistical information from the dynamic responses, e.g., probability density function (PDF), and further reveal the system vulnerabilities by assessing the responses across the full range of possible values of uncertain parameters [2]. This significantly improves the operator's situational awareness.

The widely used approach to probabilistically quantify the impacts of uncertain parameters on dynamic response is Monte Carlo (MC) simulations. MC-based approach repeatedly generates a large number of samples from the search space to obtain system response statistical information, which is very time-consuming for large-scale systems with a high dimension of uncertainties [2]. Note that the uncertainties from dynamic PVs and loads further increase the computational burden. Some improved MC-based approaches, such as the sequential Monte Carlo [3], quasi-Monte Carlo [4], Markov chain Monte Carlo [5] and enhanced sampling methods like Latin Hypercube sampling (LHS) are also applicable for probabilistic analysis. However, they need accurate PDFs of uncertain resources that are difficult to obtain in practice. Meanwhile, these approaches are not scalable to large-scale systems. To further improve the computational efficiency of MC approach, the surrogate modeling methods are applied [6]–[8]. The key idea is constructing a reduced order model to approximate the original complex power system model, thus significantly simplifying the problem. Other than the sampling-based methods, another approach is to focus on the statistics of model output. A typical example is moments-based method. In [9], the moments of the model output are calculated as a function of uncertain inputs, based on which the distribution of output can be subsequently established using expansion techniques. Point estimate method requires less calculations and avoids building surrogate models. But it only produces raw moments and relies on the accurate PDF of inputs [2]. An alternative to the output statistics is the cumulant-based method, which decomposes the system output into the summands of the cumulants of the independent input

Manuscript received 25 January 2022; revised 21 May 2022; accepted 30 June 2022. Date of publication 4 July 2022; date of current version 24 April 2023. This work was supported in part by National Science Foundation under Grant ECCS 1917308, and in part by the Department of Energy Advanced Grid Modeling Program. Paper no. TPWRS-00137-2022. *(Corresponding author: Junbo Zhao.)*

Ketian Ye and Junbo Zhao are with the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269 USA (e-mail: ketian.ye@uconn.edu; junbo@uconn.edu).

Nan Duan is with the Lawrence Livermore National Laboratory, Livermore, CA 94550 USA (e-mail: duan4@llnl.gov).

Yingchen Zhang is with the National Renewable Energy Laboratory, Golden, CO 80401 USA (e-mail: yingchen.zhang@nrel.gov).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPWRS.2022.3188182.

Digital Object Identifier 10.1109/TPWRS.2022.3188182

uncertainties [9]. Although the cumulant-based method can obtain central moments of output, the accurate derivations of the input-output sensitivity are needed and the method is also subject to the curse of dimensionality issues. Another widely applied method for the dynamic system is the trajectory sensitivity analysis approach based on the partial derivatives and series expansion [7]. Although it may provide insights into system behaviors, the trajectory sensitivity is not applicable for strong nonlinear systems. The interval analysis (IA) method is also an option for uncertainty quantification. Power flow can be solved as a nonlinear interval optimization problem with interval mathematics [10]. Since no probability structure is assumed for an interval, IA can only provide the upper and lower bounds for variables instead of comprehensive probability distributions. To achieve accurate statistical information efficiently, the polynomial chaos expansion (PCE) has been introduced to investigate the uncertainty propagation. PCE approximates the original complicated model using orthogonal polynomials and the corresponding coefficients. The orthogonality of the basis ensures that the first two moments of model output are encoded in the coefficients [6]. PCE-based methods assume the input variables to be independent [6], [11] but this is difficult to hold as correlations exist among PVs that are at geographically close locations. Copula statistics are embedded in PCE to tackle nonlinear correlation [12] but the method is also subject to scalability issues.

The aforementioned approaches either suffer from scalability issues or require accurate PDFs of uncertain resources. This paper develops a computationally efficient probabilistic stability assessment framework without the PDF assumption of uncertain resources. It has the following contributions:

- The proposed framework does not require the uncertain input distributions and thus distinguishes it from existing MC-based and PCE-based approaches. This is achieved via Gaussian process (GP)-based learning approaches. It is worth noting that GP approaches have been applied for distribution grid analysis [13], [14], transmission system power flow, and optimal power flow [15]. However, their extensions to power system dynamics are not trivial.
- It is found that GP is not scalable to large-scale system with high dimensional uncertain inputs as its complexity reaches $\mathcal{O}(n^3)$. To this end, a physics-informed sparse GP (SGP) with variational techniques is developed, where the inducing points are introduced to summarize training data while the variational inference [16] is used to approximate the true posterior. This significantly improves the computational efficiency and makes the approach applicable to large-scale systems, such as the Texas 2000-bus system.
- The proposed framework integrates the SGP and physical information from DAEs. The physic-guided simulation samples allow the SGP to achieve a good prior knowledge and thus reduce the parameter learning difficulty [17]. This leads to accuracy improvement as compared to the data-driven approaches.
- It can handle the coexistence of stable and unstable cases with a data pre-processing procedure, where simulations are clustered into stable and unstable cases. This allows us to analyze the probability of stable and unstable cases, a new phoneme not reported in the existing literature.
- Existing approaches have not considered detailed dynamic PVs and loads and their impacts from uncertainties while the proposed framework is general to deal with that.

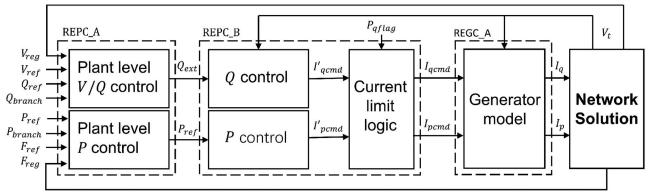

Fig. 1. Model structure for PV plant dynamic model [18].

The paper is organized as follows: Section III presents the problem statement, where the dynamic PVs and loads have been introduced; Section III shows the problem reformulation, Gaussian process, and then its enhanced sparse GP for better scalability; Section IV shows and analyzes the numerical results and finally Section V finalizes the paper.

II. Problem Statement

For power systems with uncertainties, their dynamics can be described by the following differential and algebraic equations:

$$\begin{cases} \dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\xi}) \\ \mathbf{0} = \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}) \end{cases} \tag{1}$$

where $\boldsymbol{f}(\cdot)$ and $\boldsymbol{g}(\cdot)$ are nonlinear functions related to the dynamics of synchronous machines, dynamic PVs, and loads as well as the power flow equations, respectively; in this paper, all synchronous generators are modeled by the two-axis model with IEEE-DC1A exciter and TGOV1 turbine-governor; $\boldsymbol{x}$ and $\boldsymbol{y}$ consist of the dynamic state variables of synchronous generators, dynamic PVs and loads and the algebraic state variables, respectively; $\boldsymbol{u}$ is the system input; $\boldsymbol{\xi}$ denotes a random vector that contains all uncertain variables from dynamic loads and dynamic PVs in this paper, such as uncertain parameters, power injections, etc.

The dynamic PVs are modeled using the WECC generic models [18] while the dynamic behavior of loads is modeled by the widely-used composite ZIP-motor model [19]. The dynamic PV model is shown in Fig. 1, where three sub-models plant controller (REPC_A), electrical control module of inverter (RECC_B), converter interface with grid (REGC_A) compose large-scale PV model [18]. Subscript 'ref' indicates the reference values; 'reg' refers to the regulated values; 'cmd' means current command; 'flag' is used to enable frequency control mode; 'ext' denotes external model [20]. Please refer to [18] and [20] for more details about the dynamic PV model and parameters. The ZIP model is described as

$$\begin{cases} P_{ZIP} = P_0 \left[a_0 + a_1 \left(\frac{V}{V_0} \right) + a_2 \left(\frac{V}{V_0} \right)^2 \right] \\ Q_{ZIP} = Q_0 \left[b_0 + b_1 \left(\frac{V}{V_0} \right) + b_2 \left(\frac{V}{V_0} \right)^2 \right] \end{cases} \tag{2}$$

where P_{ZIP} and Q_{ZIP} are the real-time real and reactive power injections from ZIP load; V_0 and V are the nominal and real-time voltage magnitudes; P_0, Q_0, V_0 are nominal values and their coefficients a_0, a_1, a_2 and b_0, b_1, b_2 satisfy

$$a_0 + a_1 + a_2 = 1 - \frac{P_m}{P_0}; \quad b_0 + b_1 + b_2 = 1 - \frac{Q_m}{Q_0} \tag{3}$$

The dynamic part is modeled as an equivalent third-order induction motor [19]:

$$\begin{cases} \frac{\mathrm{d}E'_d}{\mathrm{d}t} = -\frac{1}{T'}\left[E'_d + (X - X')I_q\right] - (\omega - 1)E'_q \\ \frac{\mathrm{d}E'_q}{\mathrm{d}t} = -\frac{1}{T'}\left[E'_q + (X - X')I_d\right] - (\omega - 1)E'_d \\ \frac{\mathrm{d}\omega}{\mathrm{d}t} = -\frac{1}{2H}\left[T_0 - \left(E'_d I_d - E'_q I_q\right)\right] \end{cases} \tag{4}$$

$$\begin{cases} I_d = \frac{1}{R_s^2 + X'^2}\left[R_s\left(V_d - E'_d\right) + X'\left(V_q - E'_q\right)\right. \\ I_q = \frac{1}{R_s^2 + X'^2}\left[R_s\left(V_q - E'_q\right) + X'\left(V_d - E'_d\right)\right. \end{cases} \tag{5}$$

in which the motor parameters $T' = \frac{X_r + X_m}{R_r}$; $X = X_s + X_m$; $X' = X_s + \frac{X_m X_r}{X_m + X_r}$. V_d, V_q, I_d and I_q are the d-axis and q-axis voltages and currents; respectively; E'_d and E'_q are d-axis and q-axis transient voltages; ω is the motor rotor speed; T_0 is the rotor mechanical power;the detailed explanations for motor parameters can be found in [19] and the real and reactive power of the induction motor, i.e., P_m and Q_m are:

$$P_m = V_d I_d + V_q I_q; \quad Q_m = V_q I_d - V_d I_q \tag{6}$$

Since the dynamic load is the combination of static and dynamic parts, the total power of the composite model is:

$$P_{total} = P_{ZIP} + P_m; \quad Q_{total} = Q_{ZIP} + Q_m \tag{7}$$

The main purpose of this paper is to investigate the impacts of uncertainties from dynamic loads and PVs, i.e., $\boldsymbol{\xi}$ on power system dynamic simulations and transient stability. In the literature, the MC-based sampling method is the most widely-used probabilistic approach. By sampling from PDF of $\boldsymbol{\xi}$ (the uncertain static loads or renewable energy instead of the dynamic PVs and loads in this paper), extensive MC simulations are performed to collect all possible outcomes to construct the distribution of dynamic response, i.e., $\boldsymbol{x}$. Power system dynamic simulation is performed for each sample to obtain the trajectory of the model response. However, the high computational cost makes MC inadequate for high-dimensional uncertain inputs. The consideration of dynamic loads and PVs further complicates the system, leading to increased computational cost. Furthermore, PVs that are geographically close are correlated and even a nonlinear dependence structure exists [21]. The PCE-based approaches shown in [6], [11] do not address this type of correlation. They also need accurate PDF information of the uncertain inputs, which is difficult to achieve in practice. As a data-driven method, GP mitigates the inferring error of uncertain input distribution. However, the original GP has a time complexity of $\mathcal{O}(n^3)$, where n is the number of training examples that depends on the dimension of uncertain inputs and system size. This paper develops a physics-informed SGP approach to bridge these gaps.

III. Proposed Physics-Informed SGP for Probabilistic Stability Assessment

In this section, the problem is first reformulated to be compatible with the GP learning framework. Then, the GP-based method is briefly introduced and its enhanced SGP is shown. The probabilistic stability index is also introduced.

A. Problem Reformulation

The differential and algebraic equations shown in (1) makes it challenging to apply the GP-based approaches. This is because GP is usually used for regression analysis and could not be directly applied for dynamic equations. To this end, we rely on the structural-preserve approach and the Kron reduction [22] to represent the algebraic variables $\boldsymbol{y} = \boldsymbol{g}^{-1}(\boldsymbol{x}, \boldsymbol{\xi})$, where $\boldsymbol{g}^{-1}$ is the inverse function of $\boldsymbol{g}$. Then, we obtain the following regression model by substituting it into the differential equation:

$$\dot{\boldsymbol{x}} = \boldsymbol{f}\left(\boldsymbol{x}, \boldsymbol{g}^{-1}(\boldsymbol{x}, \boldsymbol{\xi}), \boldsymbol{u}, \boldsymbol{\xi}\right) \tag{8}$$

which can be rewritten as follows after linearization:

$$\boldsymbol{x}(t) = \boldsymbol{x}(t-1) + \Delta t \times \boldsymbol{f}\left(\boldsymbol{x}(t-1), \boldsymbol{g}^{-1}(\boldsymbol{x}(t-1), \boldsymbol{\xi}), \boldsymbol{u}, \boldsymbol{\xi}\right) \tag{9}$$

where Δt is the time step in the numerical integration process. This can also be rewritten into the following compact form:

$$\boldsymbol{x}(t) = \mathcal{M}(\boldsymbol{x}(t-1), \boldsymbol{\xi}, \boldsymbol{u}, t) \tag{10}$$

where $\boldsymbol{\xi}$ denotes uncertainties from dynamic loads and PVs; $\boldsymbol{x}$ is the dynamic response, such as generator rotor speeds and angles that can be used for transient stability assessment. This reformulation focuses on quantifying the uncertainty propagation from $\boldsymbol{\xi}$ to $\boldsymbol{x}$ through the model $\mathcal{M}$ derived from the differential and algebraic equations. This allows assessing how the variability of uncertain sources affects the system dynamics and stability. Note that the transmission models are generally of good quality and therefore can provide accurate physical information for the framework. The key idea of the proposed method is to use the physic-guided simulation samples to enable a good prior knowledge of the SGP and thus reduce the parameter learning difficulty. This leads to accuracy improvement as compared to the data-driven approaches.

B. GP for Uncertainty Quantification

It is worth noting that the derivation of $\mathcal{M}$ can be a complicated process with detailed dynamic models and the assessment of (10) via MC-based approaches is still time-consuming while requiring the accurate PDF of $\boldsymbol{\xi}$. We advocate the use of GP for such uncertainty quantification as it has rigorous theoretical justifications and proofs, the ability to construct a data-driven model without the distributions of uncertain inputs in the presence of a small sample size [23]. The key idea is to learn the mapping function $\mathcal{M}$ from a limited number of samples of $\boldsymbol{\xi}$ and the system response.

Since the model responses for dynamic systems are time-dependent and as a result, the statistics of model responses evolve as a function of time. Consequently, the parameters of $\mathcal{M}$ are time-varying. Without loss of generalizability, we take the time

step t as an example to elaborate on GP for uncertainty quantification. Given a set of observations $\boldsymbol{\Xi}$ and system responses $\boldsymbol{X}$, we want to identify a reduced-order model of the original $\mathcal{M}$, i.e., $f:\boldsymbol{\Xi}\rightarrow\boldsymbol{X}$:

$$\boldsymbol{X} = f(\boldsymbol{\Xi}) + \boldsymbol{\epsilon} \tag{11}$$

where $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma_\epsilon^2\boldsymbol{I})$ represents the errors from observations as well as the mapping function; σ_ϵ is the standard deviation; f is a GP defined by the mean function and the covariance function as follows:

$$f \sim \mathcal{GP}\left(m(\boldsymbol{\xi},t), k\left(\boldsymbol{\xi},\boldsymbol{\xi}';\boldsymbol{\theta}\right)\right) \tag{12}$$

where the trend $m(\boldsymbol{\xi},t)$ is usually a polynomial of $\boldsymbol{\xi}$ and $k(\boldsymbol{\xi},\boldsymbol{\xi}';\boldsymbol{\theta})$ is a kernel function with parameters $\boldsymbol{\theta}=\boldsymbol{\theta}(t)$; $\boldsymbol{\xi}'$ is the data excluding the training points $\boldsymbol{\xi}$. The joint prior distribution of the training outputs $\boldsymbol{f}_{tr}=f(\boldsymbol{\Xi})$ and test outputs $\boldsymbol{f}_*=f(\boldsymbol{\Xi}_*)$ is:

$$\begin{bmatrix}\boldsymbol{f}_{tr}\\ \mathbf{f}_*\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{m}\\ \mathbf{m}_*\end{bmatrix}, \begin{bmatrix}k(\boldsymbol{\Xi},\boldsymbol{\Xi}) & k(\boldsymbol{\Xi},\boldsymbol{\Xi}_*)\\ k(\boldsymbol{\Xi}_*,\boldsymbol{\Xi}) & k(\boldsymbol{\Xi}_*,\boldsymbol{\Xi}_*)\end{bmatrix}\right) \tag{13}$$

The likelihood follows a Gaussian distribution related to $\boldsymbol{\Xi}$:

$$\boldsymbol{X}|\boldsymbol{\Xi},\boldsymbol{f}_{tr} \sim \mathcal{N}\left(\boldsymbol{f}_{tr},\sigma_\epsilon^2\boldsymbol{I}\right) \tag{14}$$

By applying Gaussian marginalization rule, the marginal likelihood is derived: $\boldsymbol{X}\sim\mathcal{N}(\mathbf{m},\boldsymbol{K})$, where $\boldsymbol{K}$ is the Gram matrix with terms $\boldsymbol{K}_{ij}=k(\boldsymbol{\xi}_{(i)},\boldsymbol{\xi}_{(j)})$. With the prior and the likelihood, the posterior distribution over the same set of random variables introduced in the prior follows a GP with the posterior mean $\hat{\mathbf{m}}$ and posterior covariance matrix $\hat{\boldsymbol{K}}$ as:

$$\begin{cases}\hat{\mathbf{m}} = \mathbf{m}_* + k(\boldsymbol{\Xi}_*,\boldsymbol{\Xi})\left(\boldsymbol{K}+\sigma_\epsilon^2\boldsymbol{I}\right)^{-1}(\boldsymbol{X}-\mathbf{m})\\ \hat{\boldsymbol{K}} = \boldsymbol{K}_* - k(\boldsymbol{\Xi}_*,\boldsymbol{\Xi})\left(\boldsymbol{K}+\sigma_\epsilon^2\boldsymbol{I}\right)^{-1}k(\boldsymbol{\Xi},\boldsymbol{\Xi}_*)\end{cases} \tag{15}$$

It is observed from the above equation that the system response is approximated as a GP, where the mean and variance corresponding to each time instant are analytically calculated. That is, the Bayes rule models the propagation of the uncertainties from the prior and the likelihood. Note that the common trends are typically modeled as linear while the kernel functions are represented as exponential kernel, or Gaussian kernel, or Matérn kernel. Their hyperparameters are optimized via a maximum likelihood estimator.

C. Scalable SGP for Large System Uncertainty Quantification

A major drawback of GP is the curse of dimensionality issue as its model complexity reaches $\mathcal{O}(n^3)$, where n is the number of uncertain inputs. Commonly-used techniques to reduce computational cost include subset-of-data, sparse kernel, and low-rank approximation [24]. They introduce inducing variables to approximate involved distributions for either the prior or the posterior. By introducing a new set of l inducing points $\boldsymbol{Z}=\{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_l\}$ to summarize the training data, the function values at these inducing points are $\mathbf{r}=f(\boldsymbol{Z})$.

According to the Gaussian marginalization rule, the joint distribution $p(\mathbf{f},\mathbf{f}_*)$ can be obtained by marginalizing out $\mathbf{r}$ from the joint prior $p(\mathbf{f},\mathbf{f}_*,\mathbf{r})$:

$$p(\mathbf{f},\mathbf{f}_*) = \int p(\mathbf{f},\mathbf{f}_*|\mathbf{r})p(\mathbf{r})\,\mathrm{d}\mathbf{r} \tag{16}$$

where $p(\mathbf{r})=\mathcal{N}(\mathbf{r}|\mathbf{m}_{\boldsymbol{Z}},\boldsymbol{K}_{\boldsymbol{ZZ}})$ and $\mathbf{m}_{\boldsymbol{Z}}$ and $\boldsymbol{K}_{\boldsymbol{ZZ}}$ are the mean and covariance matrix for $\boldsymbol{Z}$. By design, $\mathbf{f}$ and $\mathbf{f}_*$ are conditionally independent given inducing points $\mathbf{r}$, i.e., $\mathbf{f}\perp\mathbf{f}_*|\mathbf{r}$, (16) can be approximated by:

$$p(\mathbf{f},\mathbf{f}_*) \simeq q(\mathbf{f},\mathbf{f}_*) = \int q(\mathbf{f}|\mathbf{r})q(\mathbf{f}_*|\mathbf{r})\,\mathrm{d}\mathbf{r} \tag{17}$$

where $q(\cdot)$ represents the variational distribution. Subsequently, the training and test conditionals are

$$\begin{cases}p(\mathbf{f}|\mathbf{r}) = \mathcal{N}\left(\boldsymbol{K}_{\mathbf{f},\mathbf{r}}\boldsymbol{K}_{\mathbf{r},\mathbf{r}}^{-1}\boldsymbol{r}, \boldsymbol{K}_{\mathbf{f},\mathbf{f}} - \boldsymbol{K}_{\mathbf{f},\mathbf{r}}\boldsymbol{K}_{\mathbf{f},\mathbf{f}}\boldsymbol{K}_{\mathbf{r},\mathbf{f}}\right)\\ p(\mathbf{f}_*|\mathbf{r}) = \mathcal{N}\left(\boldsymbol{K}_{\mathbf{f}_*,\mathbf{r}}\boldsymbol{K}_{\mathbf{r},\mathbf{r}}^{-1}\boldsymbol{r}, \boldsymbol{K}_{\mathbf{f}_*,\mathbf{f}_*} - \boldsymbol{K}_{\mathbf{f}_*,\mathbf{r}}\boldsymbol{K}_{\mathbf{f}_*,\mathbf{f}_*}\boldsymbol{K}_{\mathbf{r},\mathbf{f}_*}\right)\end{cases} \tag{18}$$

where the calculations for covariance matrices $\boldsymbol{K}$ follow the expression shown below the (14). By introducing inducing points to characterize the dependencies between training and test outputs, the model complexity is reduced to $\mathcal{O}(nl^2)$, where l is much less than n.

In [25], a variational approach is proposed for sparse approximation by minimizing the Kullback-Leibler (KL) divergence between a variational GP and the true posterior GP, i.e., $\mathrm{KL}[q(\mathbf{f},\mathbf{r})||p(\mathbf{f},\mathbf{r}|\boldsymbol{X})]$. In this paper, we advocate the use stochastic variational inference (SVI) [16]. SVI allows variational inference for very large data that is needed for large-scale power system applications.

To integrate SVI with GP, the minimization of the KL divergence between the variational posterior q and the true posterior p is transformed to maximizing the lower bound of the true log marginal likelihood:

$$\mathcal{L} = \int p(\mathbf{f}|\mathbf{r})p(\mathbf{r}|\boldsymbol{X})\log\frac{p(\boldsymbol{X}|\mathbf{f})p(\mathbf{r})}{p(\mathbf{r}|\boldsymbol{X})}\,\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{r} \tag{19}$$

Then, the variational posterior to be approximated is:

$$q(\mathbf{f},\mathbf{r}) = p(\mathbf{f}|\mathbf{r})q(\mathbf{r}) \tag{20}$$

where $q(\mathbf{r})=\mathcal{N}(\mathbf{r}|\mathbf{m}_{\mathbf{q}},\mathbf{S})$ and $\mathbf{m}_q$ and $\mathbf{S}$ are their corresponding mean and covariance matrix derived from the PDF. By marginalizing $\mathbf{r}$, (20) is

$$q(\mathbf{f}|\mathbf{m}_q,\mathbf{S}) = \int p(\mathbf{f}|\mathbf{r})q(\mathbf{r})\,\mathrm{d}\mathbf{r} = \mathcal{N}\left(\mathbf{f}|\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Xi}}\right) \tag{21}$$

where the mean and covariance of SGP are as follows:

$$\begin{cases}\hat{\boldsymbol{\mu}} = \mathbf{m}_q + \boldsymbol{P}\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{K}_{\boldsymbol{Z\Xi}}(\mathbf{m}_q - \mathbf{m}_{\boldsymbol{Z}})\\ \hat{\boldsymbol{\Xi}} = \boldsymbol{K} - \boldsymbol{P}^{\mathrm{T}}\left(\boldsymbol{K}_{\boldsymbol{ZZ}} - \mathbf{S}\right)\boldsymbol{P}\end{cases} \tag{22}$$

and $\boldsymbol{P}=\boldsymbol{K}_{\boldsymbol{\Xi Z}}\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}$. The variational parameters $(\boldsymbol{Z},\mathbf{m}_q,\mathbf{S})$ can be estimated analytically using the approach in [26].

Remark: the physics-informed SGP means that the samples from uncertain inputs must follow the differential and algebraic equations but the construction of the SGP model only relies on the sampled data as well as those propagated via the differential and algebraic equations. One key advantage is that SGP only needs a few hundred of samples, distinguishing it from other sample-extensive or data-hungry deep learning approaches.

Based on these properties, SGP is suitable for application in large systems with high-dimensional data and unknown input distributions. The GP-based method can handle noise to some extent but may have biased data in the presence of high-noise data (not very common in practice). In this paper, data quality and quantity issues are investigated, see Section IV. In future work, we will develop a robust SGP approach to deal with noisily data and even bad data.

D. Probabilistic Transient Stability Assessment

To evaluate the impacts of uncertain resources on transient stability, the probabilistic transient stability assessment (TSA) is carried out. The probabilistic TSA framework consists of sampling and simulations, transient stability index (TSI) calculation, and statistical analysis. Traditionally, MC simulations are performed to calculate TSI shown below [2]:

$$\mathrm{TSI} = 100 \times \frac{360 - \delta_{\max}}{360 + \delta_{\max}} \tag{23}$$

where $\delta_{\max}$ is the maximum rotor angle difference between any two generators. The system is transient stable when $\mathrm{TSI} > 0$, and vice versa. Large TSI indicates the system has a large stability margin. For each dynamic simulation result, its TSI is calculated and the shape of the TSI distribution allows for assessing the possibility of losing stability. The latter can provide operator system probabilistic stability by quantifying and characterizing the risk, based on which the trade-off between risk and cost can be achieved via preventive control. To perform probabilistic analysis, kernel density estimation (KDE) [27] is performed on all TSI values to infer the PDF, i.e.,

$$\widehat{f}\,(\mathrm{TSI}) = \frac{1}{\alpha h} \sum_{i=1}^{\alpha} Ker\left(\frac{\mathrm{TSI} - \mathrm{TSI}_i}{h}\right) \tag{24}$$

where α is the number of samples; h denotes the bandwidth parameter; Ker is standard Gaussian kernel in this paper.

E. Handling the Co-Existence of Stable and Unstable Scenarios

When the system is operated under stressed conditions or subject to large disturbances, the perturbations on differential and algebraic equations state initialization caused by uncertainties from dynamic loads and PVs can lead to stable, unstable, and co-existence of stable and unstable scenarios with different probabilities. Since the patterns for absolute stable and unstable scenarios are quite different, it is not feasible to use one SGP model to capture the co-existence of stable and unstable scenarios, which will be shown in the numerical results section.

To address that, a data processing strategy is proposed. In particular, a number of samples obtained by Latin hypercube sampling is first generated. Then, these samples are propagated via the power system dynamic simulations to get corresponding responses. If the TSI is beyond a preset value, the corresponding sample will be claimed as unstable. This means that a simple classification of the samples with boundaries via k-means algorithm or support vector machine can be used to categorize the samples into stable and unstable cases. Then, two SGP models are constructed separately using stable and unstable samples for uncertainty quantification. Note that when the system is absolutely stable or unstable, only one SGP model is needed.

F. Algorithm Implementation

The proposed framework consists of clustering for stable and unstable cases, SGP model construction, uncertainty quantification, stability index calculation, and probabilistic assessment. The implementation has the following steps:

- *Step 1:* Generate training samples from historical data using LHS and collect corresponding simulation results based on dynamic simulations on (1).
- *Step 2:* Cluster all samples into two groups to obtain stable and unstable datasets at a given criterion. In this paper, TSI is used.
- *Step 3:* Construct a SGP model for each group and estimate parameters. The uncertainty associated with the system response can be calculated via (22).
- *Step 4:* TSI is calculated by (23) and its probability density function estimated by (24) is utilized for probabilistic transient stability assessment.

IV. Numerical Results

Simulations are carried out on the modified IEEE 118-bus and large-scale Texas 2000-bus systems [28] to demonstrate the effectiveness and advantages of the proposed method. The synchronous generators are assumed to be the two-axis model with their parameters following the benchmark [29]. For the 118-bus system with 54 generators, a three-phase fault is applied at bus 15 at 0.01 s and is cleared after 5 cycles by opening line 15–33. For the 2000-bus system, the fault happens at bus 793 at 0.01 s and the duration is 5 cycles until line 793–823 is opened. The uncertain resources introduced include dynamic loads and PVs. The uncertain loads for initial power flow calculation are assumed to follow Gaussian distributions with means being the true means and standard derivations being a proportion (uncertainty level) of the means. Similarly, PVs are assumed to follow Beta distributions, where the upper bounds are determined by load consumption. Dynamic loads and PVs are modeled as the ZIP-motor load model and WECC generic model [18], respectively. The rotor angles and speeds are chosen to be desired outputs for transient stability analysis. Different scenarios are designed to investigate the impacts of uncertain resources on dynamic responses.

Four methods are compared and discussed, which are LHS, PCE, and the proposed GP and SGP. Results obtained using MC simulations (MCS) with 10,000 samples are considered as the benchmark. For LHS, 3000 samples are assumed. All simulations are conducted on MATLAB with 2.60 GHz Intel Core i7-6700HQ. The mean absolute percentage error (MAPE) is used to measure the overall performance of each method, i.e.,

$$\mathrm{MAPE} = \frac{1}{T} \sum \left| \frac{x^* - \hat{x}}{x^*} \right| \times 100\% \tag{25}$$

where T is simulation time, x^* and $\hat{x}$ represent the true and estimated dynamic responses, respectively. Since this paper

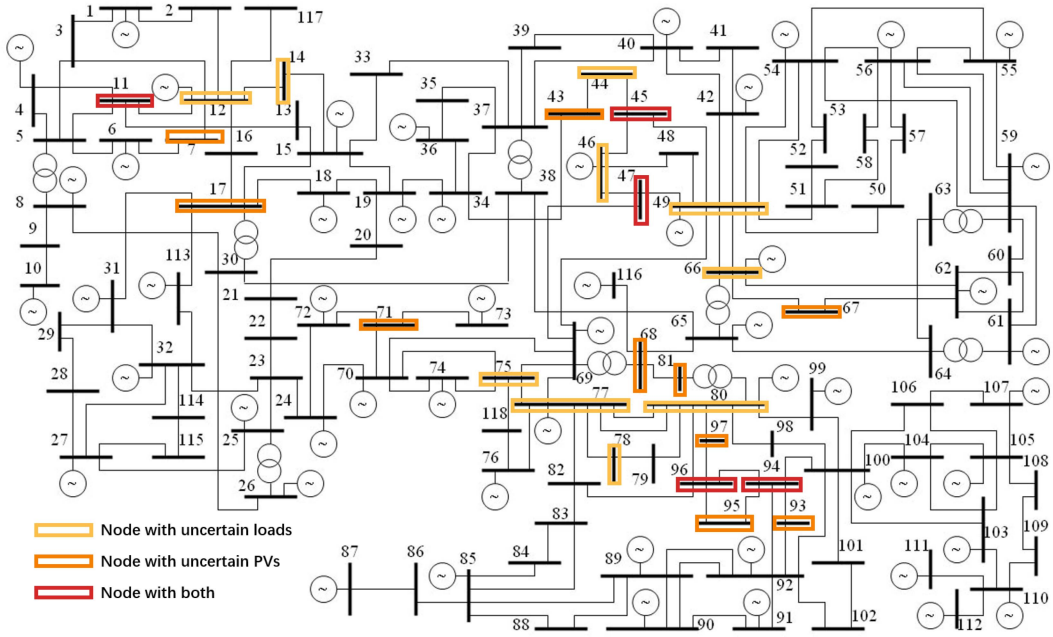

Fig. 2. Single-line diagram of the modified IEEE 118-bus system with uncertain dynamic loads and PVs.

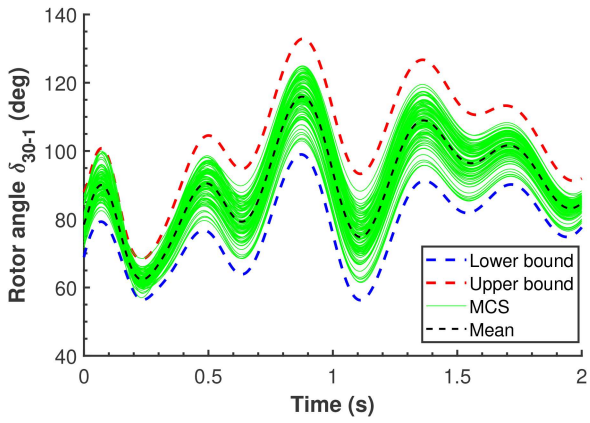

Fig. 3. The upper and lower bounds of rotor angle δ_{30-1} in 118-bus system Scenario 1.

focuses on transient stability, the statistics of generator rotor angles will be used as critical information. More specifically, the average MAPE of the mean and variance of the model response over entire simulation time will be used as error indices, denoted as e_μ and e_{σ^2}. Note that MAPE is a widely used metric to evaluate the performance of many different approaches in the literature and it is thus adopted in this paper.

A. Validation on the 118-Bus System

In this section, all four methods are examined on the 118-bus system, where 15 uncertain loads and 15 uncertain PVs are located at buses (11, 12, 14, 44–47, 49, 66, 75, 77, 78, 80, 94, 96) and (7, 11, 17, 43, 45, 47, 67, 68, 71, 81, 93–97), respectively, see Fig. 2. The angle of generator 30 at bus 69 (δ_{30-1}) with respect to the reference generator 1 is used for illustration. Fig. 3 shows

TABLE I
PDFs OF LOADS AND PVs FOR 118-BUS AND 2000-BUS SYSTEMS

Scenarios	$PV(kW)$	$PL(kW)$
Scenario 1 (118-bus)	$PV_i \times Beta(2.06, 2.2)$	$\mathcal{N}(PL_i, (0.1PL_i)^2)$
Scenario 3 (2000-bus)	$\mathcal{N}(PV_i, (0.1PV_i)^2)$	$\mathcal{N}(PL_i, (0.1PL_i)^2)$

the uncertainty of rotor angle response δ_{30-1} obtained from MC simulations with dynamic PVs and loads, where the upper and lower bounds are calculated using 3σ rule. It can be observed that the trajectories follow certain distributions and the uncertainties from model inputs will lead to different dynamic responses. For the PCE methods, the degree of PCE is set to be $n = 2$ and the coefficients are computed by the regression method with 200 samples of rotor angle trajectories. For GP, a total number of 200 samples are provided for inference while 100 samples along with 50 inducing points are used for SGP. Gaussian kernel and gradient descent algorithm are utilized for model construction and parameter estimation.

1) Validation of the Proposed Method: The distributions of loads and PVs are listed in Table I under the Scenario 1. The comparison results of four methods are shown in Fig. 4 and Table III. The numerical results in Table III show that both GP and SGP methods outperform LHS in terms of both accuracy and efficiency. The performance of GP is better than PCE with the same number of samples. SGP yields the best performance with the lowest computational cost thanks to the variational inference for reducing problem complex without losing of accuracy.

2) Robustness to Nonlinear Correlations Among Uncertain Inputs: The impacts of uncertain inputs with nonlinear correlations are also tested and analyzed. It is assumed that the neighboring PVs at buses 93–97 have a dependence structure described by C-Vine copulas [30] and their types and parameters

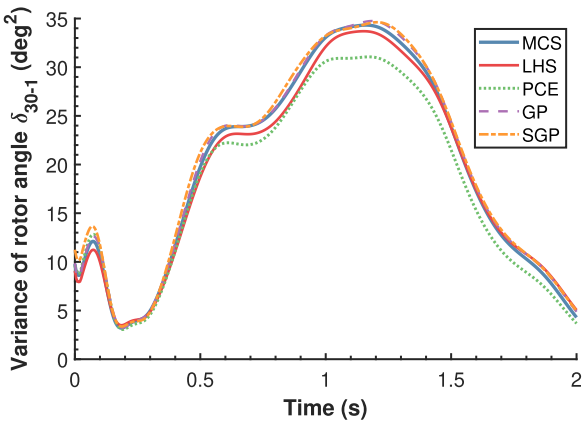

Fig. 4. True and estimated variances of δ_{30-1} from different methods for 118-bus system under the Scenario 1.

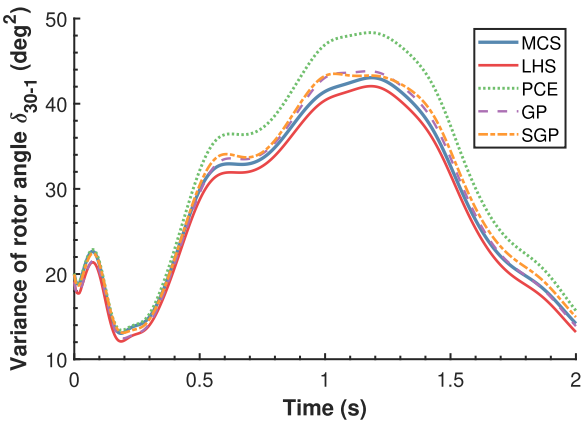

Fig. 5. True and estimated variances of δ_{30-1} from different methods for 118-bus system under the Scenario 2.

TABLE II
NONLINEAR CORRELATIONS OF PVS FOR 118- AND 2000-BUS SYSTEM

Scenarios	Pair Copula type	Parameter
Scenario 2 (118-bus)	$[Clayton, Clayton, Gumbel, Frank, t, [Clayton, Clayton, Gumbel, Frank, t]$	$[2.5, 2, 1.5\ 0.5, [0.5, 1.5], [2.5, 2, 1.5\ 0.5, [0.5, 1.5]]$
Scenario 4 (2000-bus)	$[Clayton, Gumbel, Frank, Clayton, t, [Clayton, Gumbel, Frank, Clayton, t]$	$[2.5, 1.5, 1.5\ 1.5, [0.4, 2], [2, 1.5, 2\ 1.5, [0.5, 1.2]]$

TABLE III
COMPARISON RESULTS AMONG DIFFERENT METHODS FOR 118-BUS SYSTEM UNDER SCENARIOS 1-2

Scenario	Method	MAPE		CPU time (s)
		$e_\mu(\times 10^{-2}\%)$	$e_{\sigma^2}(\%)$	
Scenario 1 (Independent)	MCS	–	–	3158.43
	LHS	4.27	2.31	831.60
	PCE	13.70	10.89	112.91
	GP	3.18	1.67	92.84
	SGP	3.42	1.74	70.96
Scenario 2 (Correlated)	MCS	–	–	3227.05
	LHS	4.52	2.32	857.14
	PCE	15.26	20.85	120.06
	GP	3.25	1.89	97.93
	SGP	3.14	1.92	73.18

are shown in Table II under the Scenario 2. Fig. 5 displays the results for four methods. One of the main impacts of correlation is the uncertainty increase of system dynamic responses. The PCE is adjusted to handle dependent inputs using the approach shown in [30]. The shapes of rotor angle responses are similar since the system is stable in these scenarios. As shown in Table III, the GP-based methods are able to achieve the best performances as compared to others. By comparing Scenarios 1 and 2, it can be observed that the CPU time is slightly increased since the problem complexity has increased due to the existence of nonlinear correlations among uncertain inputs. One of the disadvantages of PCE is that it requires the distribution information of inputs, which is difficult to obtain accurately in practice. Hence, the complexity of the dependence structure of input brings additional difficulties and computational burdens for PCE. By contrast, GP is a non-parametric model and is only slightly affected by the complicated correlation relationships among uncertain inputs.

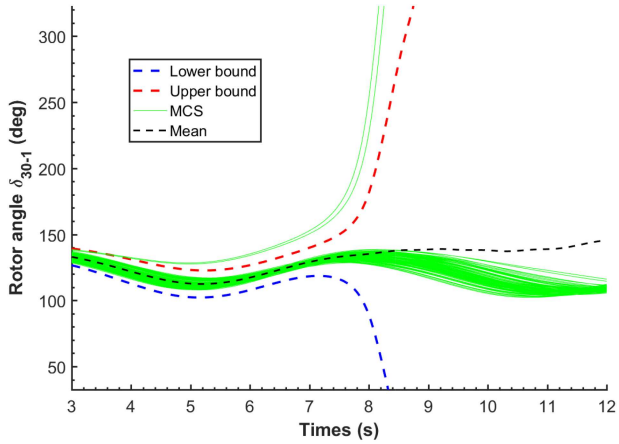

Fig. 6. MCS results of δ_{30-1} where both stable and unstable cases coexist.

3) Application for Probabilistic Transient Stability: In this subsection, we investigate the scenario, where the unstable and stable cases coexist due to the existence of uncertain inputs. Specifically, in Scenario 5, the settings and configurations of the 118-bus system are unchanged except that the power ratings of generations are reduced, the fault duration is extended to 15 cycles, and the uncertainty level is increased. The results are shown in Fig. 6 and it can be observed that few unstable cases occur while most uncertain inputs still lead to stable cases. We believe these unstable cases may be yielded from extreme initial conditions due to the uncertainty of inputs. According to our previous discussions, one SGP model is inadequate to model both stable and unstable dynamic responses. The pre-processing strategy is then used to cluster two types of trajectories and each model is built for both of them. This is achievable since both PCE and GP contain uncertainty quantification of model output. Parameters and configurations are kept unchanged except the

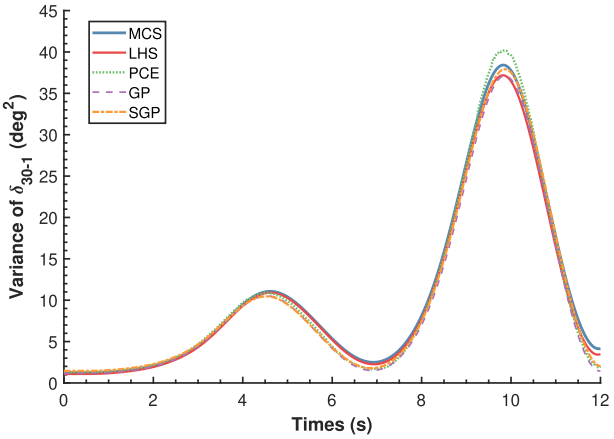

Fig. 7. True and estimated variance of δ_{30-1} in the 118-bus system for stable cases.

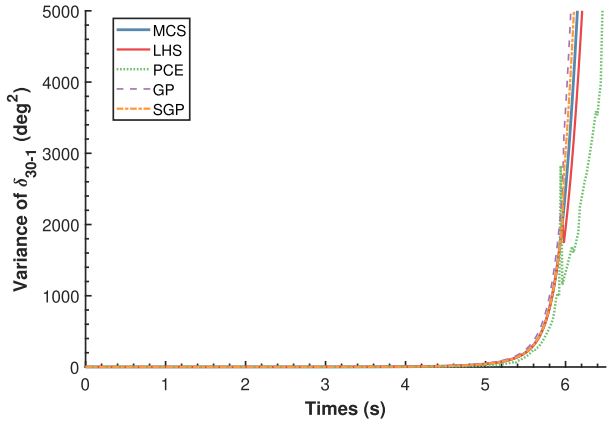

Fig. 8. True and estimated variance of δ_{30-1} in the 118-bus system for unstable cases.

TABLE IV
COMPARISON RESULTS AMONG DIFFERENT METHODS FOR 118-BUS SYSTEM SCENARIOS 5

Scenario	Method	MAPE		CPU time (s)
		$e_\mu(\times 10^{-2}\%)$	$e_{\sigma^2}(\%)$	
Scenario 5 (stable)	MCS	–	–	678.45
	LHS	4.46	2.66	199.98
	PCE	7.32	4.09	172.04
	GP	3.70	1.21	164.45
	SGP	3.57	1.04	90.77
Scenario 5 (unstable)	MCS	–	–	686.35
	LHS	7.72	3.48	204.93
	PCE	49.78	22.016	209.30
	GP	6.32	3.34	176.10
	SGP	5.69	2.68	92.53

number of samples used for PCE and GP is increased to 300. Figs. 7 and 8 and Table IV demonstrate that the proposed framework is able to approximate both stable and unstable cases with accurate statistic information. The PCE with 300 samples is insufficient for unstable cases. The number of training sample is not further increased because of the rapid growth of computational cost.

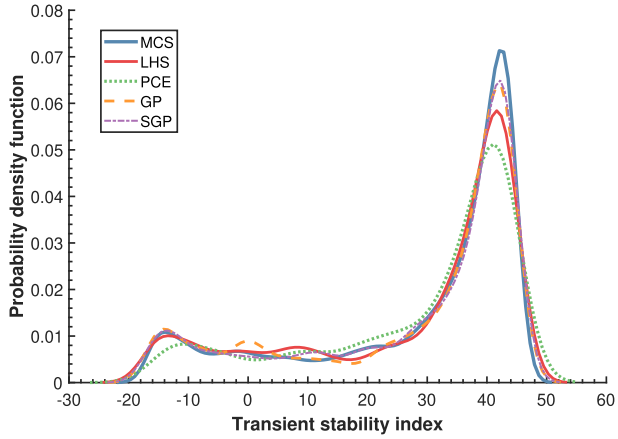

Fig. 9. Estimated PDF of TSI by different methods for 118-bus system under the Scenario 5.

The PDF of TSI shown in Fig. 9 illustrates the probabilistic analysis of system stability. In most cases, TSI is larger than 0 while in some cases TSI is less than 0. The errors accumulate through KDE, mainly at the peak of the PDF of TSI. PCE performs worse than LHS, especially at the peak while two GP-based methods achieve better accuracy. SGP outperforms GP in capturing the tail of TSI distribution, which means SGP can provide more accurate estimates for extreme cases. The overall information of TSI presents a comprehensive depiction of system stability, which can be used to estimate quantiles and the overall probability of being unstable.

B. Scalability to Large-Scale 2000-Bus System

Tests are also carried out on the 2000-bus system to show the effectiveness and applicability of the proposed method to large-scale systems. Similar to the 118-bus system, 100 uncertain dynamic loads and 100 dynamic PVs are designed and placed on the 2000-bus system in Fig. 10, where 100 loads are divided into 10 groups and are controlled by 10 factors to simulate similar load shapes within the same area. On the other hand, dynamic PVs with synthetic output data are assigned near actual PVs locations and are initially viewed as independent uncertain sources. The parameters and hyperparameters selection are the same as those in Section IV-A expect that 1000 samples are used for PCE and GP. Two sets of parameters, 500 samples with 200 inducing points and 400 samples with 100 inducing points (SGP$*$) are tested for SGP.

1) Feasibility Results: The rotor angle response δ_{40} is demonstrated in Fig. 11 and the numerical comparisons are shown in Table V corresponding to the Scenario 3. As the system becomes larger, MC simulations are more expensive for practical use. The accuracy of PCE is insufficient for this large system as compared to other methods. PCE also confronts with the curse of dimensionality issue and GP is approaching the time restriction to complete estimation and analysis. With the help of variational inference, SGP is able to speed up the calculation process while maintaining high accuracy. The comparisons of two sets of parameters for SGP indicate that the computational

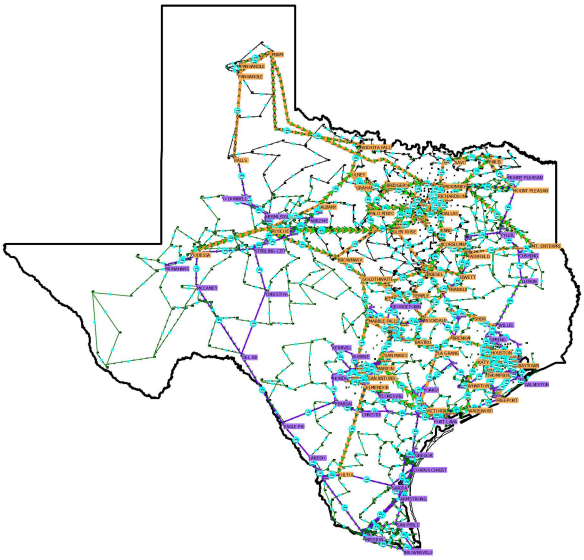

Fig. 10. Diagram of the Texas 2000-bus system [28].

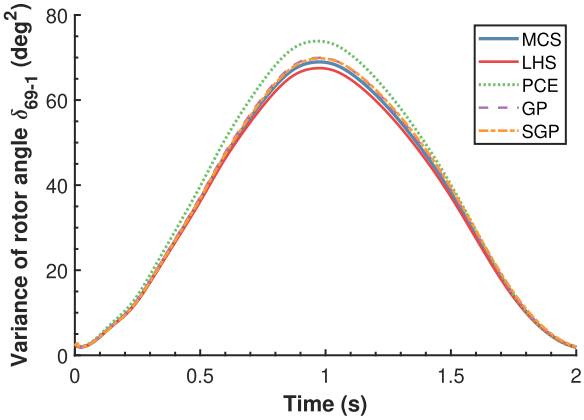

Fig. 11. True and estimated variances of δ_{40-1} from different methods for 2000-bus system under the Scenario 3.

TABLE V
COMPARISON RESULTS AMONG DIFFERENT METHODS FOR 2000-BUS SYSTEM UNDER SCENARIOS 3-4

Scenario	Method	MAPE $e_\mu(\times 10^{-2}\%)$	MAPE $e_{\sigma^2}(\%)$	CPU time (s)
Scenario 3 (Independent)	MCS	–	–	25886.64
	LHS	4.12	2.54	6639.61
	PCE	8.92	6.44	2289.78
	GP	3.86	1.88	1847.13
	SGP	4.15	2.11	1212.70
	SGP*	5.73	2.68	922.48
Scenario 4 (Correlated)	MCS	–	–	26682.41
	LHS	4.58	2.48	6785.44
	PCE	10.72	8.56	2766.76
	GP	3.44	1.76	1879.53
	SGP	4.67	1.92	1267.71
	SGP*	5.50	2.23	938.64

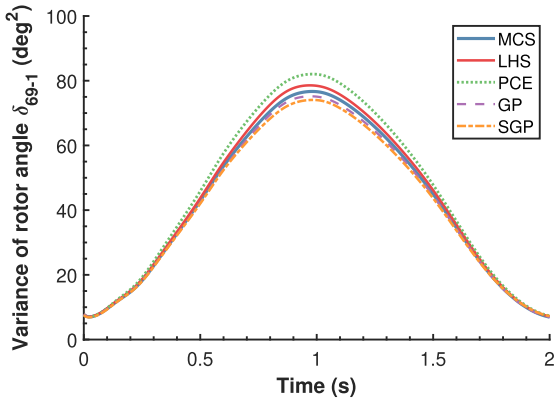

Fig. 12. True and estimated variances of δ_{40-1} from different methods for 2000-bus system under the Scenario 4.

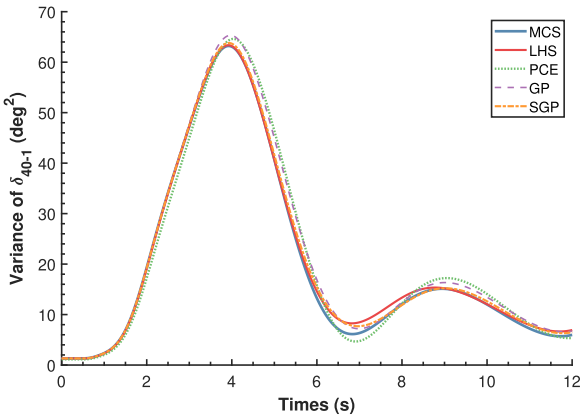

Fig. 13. True and estimated variances of δ_{40-1} in the 2000-bus system for stable cases.

efficiency of the proposed method can be further improved if some increased error is acceptable depending on practical operational limits. Anyway, the proposed approach can be used for large-scale system preventive analysis every 30 minutes as the time spent is only about 15 minutes.

2) Robustness to Nonlinear Correlations Among Uncertain Inputs: In Scenario 4, the impacts of nonlinear correlation for the 2000-bus system are studied following the settings in Table II. The comparison results on rotor angle variance is demonstrated in Fig. 12 and Table V. Similarly, a noticeable increase in variance can be observed and the computational burden of PCE increases more rapidly than GP-based methods, which further highlights the advantage of GP-based methods.

3) Application for Probabilistic Transient Stability: The ability of handling the co-existence of stable and unstable cases in the 2000-bus system is also demonstrated in Scenario 6. The power ratings of generators nearby the connected PVs are decreased, fault duration is extended to 15 cycles, and the levels of uncertainty for loads and PVs are amplified to force the system to be stressed. The number of training samples for PCE and GP is increased to 1500. Figs. 13 and 14 demonstrate the capability of the proposed method for such a large-scale system. Since the number of training sample is increased, the performance of PCE

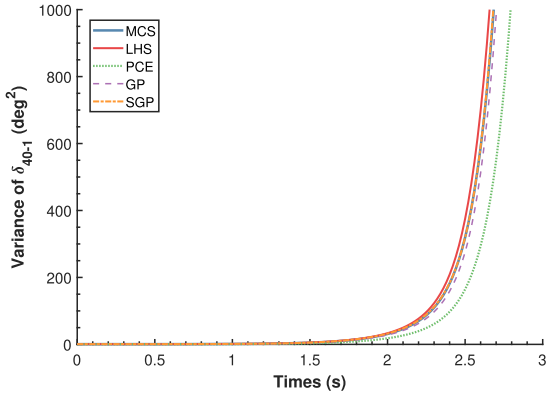

Fig. 14. True and estimated variances of δ_{40-1} in the 2000-bus system for unstable cases.

TABLE VI
COMPARISON RESULTS AMONG DIFFERENT METHODS FOR 2000-BUS SYSTEM SCENARIOS 6

Scenario	Method	MAPE		CPU time (s)
		$e_\mu(\times 10^{-2}\%)$	$e_{\sigma^2}(\%)$	
Scenario 6 (stable)	MCS	–	–	18637.26
	LHS	5.72	4.40	5931.20
	PCE	19.80	11.41	1805.90
	GP	4.93	5.13	1376.73
	SGP	3.55	3.43	937.45
Scenario 6 (unstable)	MCS	–	–	1876.59
	LHS	10.34	6.59	1927.31
	PCE	13.92	8.86	1823.44
	GP	8.34	4.69	1390.74
	SGP	5.63	3.65	942.24

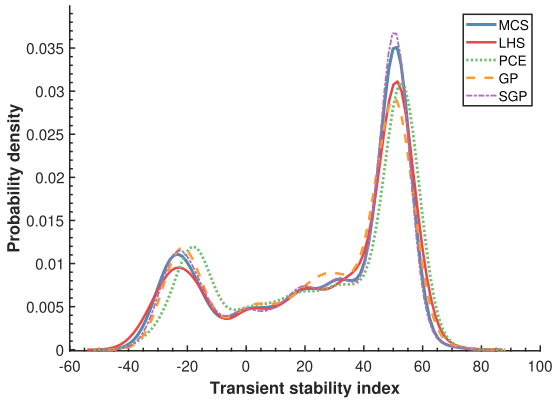

Fig. 15. Estimated PDF of TSI by different methods for 2000-bus system Scenario 6.

is not collapsed as compared to in the 118-bus system. Still, the results presented in Table VI indicate SGP achieves the best computational efficiency while maintaining high accuracy.

The probabilistic transient stability assessment is performed and its results are shown in Fig. 15. The majority of TSIs lie in the range 40–60 while another group occurs around −25. Several samples are located around 0, which implies that only a few cases are near the threshold. For the unstable cases, rotor angles will reach an abnormal value quickly while in stable cases, rotor swings will be damped out. The improvement in the accuracy

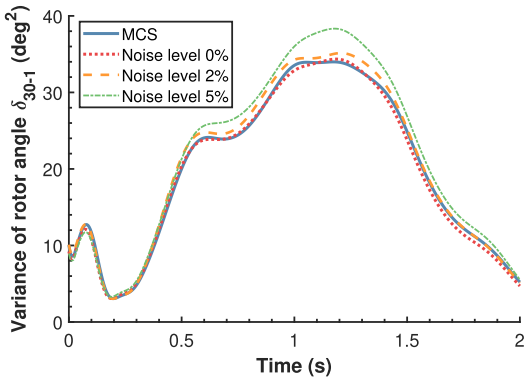

Fig. 16. True and estimated variances of δ_{30-1} from SGP with different noise levels in the 118-bus system.

TABLE VII
ROBUSTNESS OF SGP TO DATA QUALITY AND QUANTITY ISSUES

Scenario	Noise Level/ Number	MAPE		CPU time (s)
		$e_\mu(\times 10^{-2}\%)$	$e_{\sigma^2}(\%)$	
Scenario 7 (noised)	0%	3.42	1.74	70.96
	2%	4.56	2.88	78.57
	5%	11.94	9.49	81.62
Scenario 7 (limited)	500	4.67	1.92	1267.71
	400	5.33	2.08	979.12
	200	16.15	9.66	910.83

of the SGP method over other approaches can be observed in Fig. 15 as it captures much better the peak as well as the upper and lower tails.

C. Robustness to Data Quality and Quantity Issue

In practice, there are always data quality and quantity issues for learning-based approaches. Data quality issue means that the uncertain inputs are corrupted by noise and it has an influence on the training accuracy. Data quantity issue refers to the limited number of samples for algorithm training. They are investigated below.

1) Robustness to Data Quality Issue: Gaussian noise is added to the uncertain inputs and the noise level is increased from 0% to 5%. Simulations are carried out based on the noisy inputs. The results are illustrated in Fig. 16 and Table VII. The increased noise level to uncertain inputs will be propagated to the estimation results. It can be found that the SGP is robust to the noise level within 2% as the estimation error is kept below 3%. But the approximation variance curve deviates a large when the noise level reaches 5% (again not very common in practice). Meanwhile, the simulation time increases slightly as the computational burden arises due to noise level estimation.

2) Robustness to Data Quantity Issue: The number of samples for the proposed approach in the training stage is assumed to be 500, 400, and 200. As shown in Fig. 17, fewer samples result in higher estimation error as expected. Numerical results in Table VII indicate at least 400 samples are needed for SGP to produce good results. The computing time is not reduced significantly with the decreased number of samples. It is worth noting that 400 samples can be easily met in practice. In our future

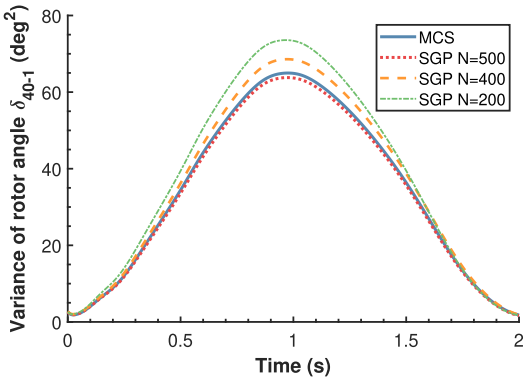

Fig. 17. True and estimated variances of δ_{40-1} from SGP with different sampling numbers in the 2000-bus system.

work, we will investigate other physics-informed approaches to see if a less number of samples can be sufficient.

V. Conclusion

In this paper, a physics-informed SGP approach has been developed for large-scale power system uncertainty quantification and probabilistic stability assessment with high dimensional uncertain inputs from dynamic loads and PVs. The differential and algebraic equations considering uncertainties from dynamic PVs and loads are reformulated to integrate the domain knowledge into GP. It is found that GP is not scalable to large-scale systems with high-dimensional uncertain inputs. To this end, SGP with variational technique is developed, significantly reducing the computational complexity while maintaining high accuracy. A data pre-processing step is also introduced to tackle the coexistence of stable and unstable cases. Extensive simulation results conducted on the modified IEEE 118-bus and the Texas 2000-bus systems show that the proposed GP and SGP achieve much better performance as compared to other methods in terms of accuracy and computational efficiency under various scenarios. The effects of uncertain loads and PVs on power system dynamics and system stability are also investigated with the consideration of nonlinear correlation among them. Data quality and quantity issues are also investigated and the limitation of SGP is discussed. SGP is able to capture the peak and tail information of the probabilistic TSI, which allows better visibility of system risks and informs preventive control. The future work would be on developing the robust SGP to handle data quality issues and exploring other physics-informed frameworks to reduce the required number of samples.

Acknowledgment

The authors would like to thank professor Joydeep Mitra for insightful discussions and suggestions to improve the paper.

References

[1] K. N. Hasan, R. Preece, and J. V. Milanović, "Existing approaches and trends in uncertainty modelling and probabilistic stability analysis of power systems with renewable generation," *Renew. Sustain. Energy Rev.*, vol. 101, pp. 168–180, Mar. 2019.

[2] J. V. Milanović, "Probabilistic stability analysis: The way forward for stability analysis of sustainable power systems," *Philos. Trans. Roy. Soc. A: Math., Phys. Eng. Sci.*, vol. 375, no. 2100, May 2017, Art. no. 20160296.

[3] W. Wangdee and R. Billinton, "Bulk electric system well-being analysis using sequential Monte Carlo simulation," *IEEE Trans. Power Syst.*, vol. 21, no. 1, pp. 188–193, Feb. 2006.

[4] H. Huang, C. Y. Chung, K. W. Chan, and H. Chen, "Quasi-Monte Carlo based probabilistic small signal stability analysis for power systems with plug-in electric vehicle and wind power integration," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 3335–3343, Aug. 2013.

[5] M. Perninge and L. Söder, "Analysis of transfer capability by Markov chain Monte Carlo simulation," in *Proc. IEEE Int. Conf. Intell. Energy Power Syst.*, 2010, pp. 232–237.

[6] Y. Xu, L. Mili, A. Sandu, M. R. V. Spakovsky, and J. Zhao, "Propagating uncertainty in power system dynamic simulations using polynomial chaos," *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 338–348, Jan. 2019.

[7] H. Choi, P. J. Seiler, and S. V. Dhople, "Propagating uncertainty in power-system DAE models with semidefinite programming," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 3146–3156, Jul. 2017.

[8] T. Tian, X. Kestelyn, O. Thomas, H. Amano, and A. R. Messina, "An accurate third-order normal form approximation for power system nonlinear analysis," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 2128–2139, Mar. 2018.

[9] R. Preece, K. Huang, and J. V. Milanović, "Probabilistic small-disturbance stability assessment of uncertain power systems using efficient estimation methods," *IEEE Trans. Power Syst.*, vol. 29, no. 5, pp. 2509–2517, Sep. 2014.

[10] A. Vaccaro, C. A. Cañizares, and K. Bhattacharya, "A range arithmetic-based optimization model for power flow analysis under interval uncertainty," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1179–1186, May 2013.

[11] J. R. Hockenberry and B. C. Lesieutre, "Evaluation of uncertainty in dynamic simulations of power system models: The probabilistic collocation method," *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1483–1491, Aug. 2004.

[12] K. Ye, J. Zhao, R. Yang, Y. Zhang, and X. Liu, "A generalized copula-polynomial chaos expansion for probabilistic power flow considering nonlinear correlations of PV injections," in *Proc. 52nd North Amer. Power Symp.*, 2021, pp. 1–6.

[13] Y. Weng and R. Rajagopal, "Probabilistic baseline estimation via gaussian process," in *Proc. IEEE Power Energy Soc. Gen. Meet.*, 2015, pp. 1–5.

[14] K. Ye, J. Zhao, C. Huang, N. Duan, Y. Zhang, and T. Field, "A data-driven global sensitivity analysis framework for three-phase distribution system with PVs," *IEEE Trans. Power Syst.*, vol. 36, no. 5, pp. 4809–4819, Sep. 2021.

[15] P. Pareek and H. D. Nguyen, "Gaussian process learning-based probabilistic optimal power flow," *IEEE Trans. Power Syst.*, vol. 36, no. 1, pp. 541–544, Jan. 2021.

[16] M. Hoffman, D. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *J. Mach. Learn. Res.*, vol. 14, pp. 1303–1347, 2013.

[17] X. Yang, D. Barajas-Solano, G. Tartakovsky, and A. M. Tartakovsky, "Physics-informed CoKriging: A Gaussian-process-regression-based multifidelity method for data-model convergence," *J. Comput. Phys.*, vol. 395, pp. 410–431, Oct. 2019.

[18] WECC Renewable Energy Modeling Task Force, "WECC Solar plant dynamic modeling guidelines," Apr. 2014. [Online]. Available: https://www.wecc.biz/Reliability/WECC%20Solar%20Plant%20Dynamic%20Modeling%20Guidelines.pdf

[19] L. Chávarro-Barrera, S. Pérez-Londoño, and J. Mora-Flórez, "An adaptive approach for dynamic load modeling in microgrids," *IEEE Trans. Smart Grid*, vol. 12, no. 4, pp. 2834–2843, Jul. 2021.

[20] R. Machlev, Z. Batushansky, S. Soni, V. Chadliev, J. Belikov, and Y. Levron, "Verification of utility-scale solar photovoltaic plant models for dynamic studies of transmission networks," *Energies*, vol. 13, no. 12, Jun. 2020, Art. no. 3191.

[21] W. Wu, K. Wang, B. Han, G. Li, X. Jiang, and M. L. Crow, "A versatile probability model of photovoltaic generation using pair copula construction," *IEEE Trans. Sustain. Energy*, vol. 6, no. 4, pp. 1337–1345, Oct. 2015.

[22] P. W. Sauer, M. A. Pai, and J. H. Chow, *Power System Dynamics and Stability: With Synchrophasor Measurement and Power System Toolbox*. Hoboken, NJ, USA: Wiley, 2017.

[23] C. Rasmussen and Christopher K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[24] J. Q. Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *J. Mach. Learn. Res.*, vol. 6, pp. 1939–1959, 2005.

[25] M. K. Titsias, "Variational learning of inducing variables in sparse gaussian processes," in *Proc. 12th Int. Conf. Artif. Intell. Statist.*, 2009, pp. 567–574.
[26] H. Salimbeni and M. Deisenroth, "Doubly stochastic variational inference for deep Gaussian processes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4588–4599.
[27] M. P. Wand and M. C. Jones, *Kernel Smoothing*. London, U.K.: Chapman & Hall, 1995.
[28] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Grid structural characteristics as validation criteria for synthetic networks," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 3258–3265, Jul. 2017.
[29] IEEE Power Energy Society Task Force on Benchmark System for Stability Controls, "Benchmark systems for small-signal stability analysis and control," IEEE Power Energy Soc., Piscataway, NJ, USA, Tech. Rep. PES-TR18, Aug. 2015.
[30] E. Torre, S. Marelli, P. Embrechts, and B. Sudret, "A general framework for data-driven uncertainty quantification under complex input dependencies using vine copulas," *Probabilistic Eng. Mech.*, vol 55, pp. 1–16, 2019.

Ketian Ye (Student Member, IEEE) received the B.S. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 2017, and the M.S. degree in electrical engineering from the University of Florida, Gainesville, FL, USA, in 2019. He is currently working toward the Ph.D. degree with the University of Connecticut, Storrs, CT, USA. His current research interests include uncertainty quantification, machine learning, and data analytics for power systems.

Junbo Zhao (Senior Member, IEEE) received the Ph.D. degree from the Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA, in 2018. He is currently an Associate Director of the Eversource Energy Center for Grid Modernization and Strategic Partnerships and an Assistant Professor with the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA. He was an Assistant Professor and Research Assistant Professor with Mississippi State University, Mississippi State, MS, USA, and Virginia Tech, from 2019 to 2021 and 2018 to 2019, respectively. He did the summer internship with Pacific Northwest National Laboratory in 2017. He is the Principal Investigator for a multitude of projects funded by the National Science Foundation, the Department of Energy, National Laboratories, and Eversource Energy. He is currently the Chair of IEEE Task Force on Power System Dynamic State and Parameter Estimation and IEEE Task Force on Cyber-Physical Interdependency for Power System Operation and Control, Co-Chair of the IEEE Working Group on Power System Static and Dynamic State Estimation, the Secretary of IEEE PES Bulk Power System Operation Subcommittee and IEEE Task Force on Synchrophasor Applications in Power System Operation and Control. He has authored or coauthored three book chapters and more than 140 peer-reviewed journal and conference papers, where more than 70 appeared in IEEE Transactions, six of his papers have been listed as highly cited papers that reflect the top 1% (one hot paper with Top 0.1%) of papers by field according to the Web of Science. His research interests include renewable energy integration, cyber-physical power system modeling, monitoring, uncertainty quantification, learning, dynamics, stability control, and cyber security. He is an Associate Editor for IEEE TRANSACTIONS ON POWER SYSTEMS, IEEE TRANSACTIONS ON SMART GRID, *International Journal of Electrical Power & Energy Systems*, North America Regional Editor of the *IET Renewable Power Generation*, and Subject Editor of *IET Generation, Transmission & Distribution*. He has been listed as the 2020 and 2021 World's Top 2% Scientists released by Stanford University in both Single-Year and Career tracks. He was the recipient of the best paper awards of 2020, 2021, and 2022 IEEE PES General Meeting (five papers), 2021 IEEE TRANSACTIONS ON POWER SYSTEMS, 2021 IEEE Sustainable Power and Energy Conference, IEEE I&CPS Asia 2021, and 2020 *Journal of Modern Power Systems and Clean Energy*, Top three Associate Editor Award of IEEE TRANSACTIONS SMART GRID in 2020, 2020 IEEE PES Chapter Outstanding Engineer Award, 2022 IEEE PES Technical Committee Working Group Recognition Award for Outstanding Technical Report, 2021 IEEE PES Outstanding Volunteer Award, and 2022 IEEE PES Outstanding Young Engineer Award.

Nan Duan (Senior Member, IEEE) received the B.S. degree in automation from the Beijing University of Technology, Beijing, China, in 2010, the M.Eng. degree in control engineering from Beihang University, Beijing, China, in 2013, and the Ph.D. degree in electrical engineering from the University of Tennessee, Knoxville, TN, USA, in 2018. He is currently a Power Systems Engineer with Lawrence Livermore National Laboratory, CA, USA. His research interests include power system modeling, high-performance computing, machine learning, and synchrophasor applications.

Yingchen Zhang (Senior Member, IEEE) received the B.S. degree from Tianjin University, Tianjin, China, in 2003, and the Ph.D. degree from Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, in 2010. He is currently a Chief Scientist and Manager with National Renewable Energy Laboratory and also an Adjunct Faculty Member of Colorado State University, Fort Collins, CO, USA, and a Lecturer with the University of Colorado-Boulder, Boulder, CO, USA. During his tenure with NREL, he was a Principal Investigator and Co-Principal Investigator of multiple projects sponsored by the Department of Energy and other entities, such as Hawaiian Electric, Pacific Gas and Electric, and Eaton. His research interests include large-scale renewable integration planning and operation, artificial intelligence and predictive analytics, and advanced energy management system for future grids.