Data-Driven Probabilistic Voltage Risk Assessment of MiniWECC System With Uncertain PVs and Wind Generations Using Realistic Data

Ketian Ye, Junbo Zhao, *Senior Member, IEEE*, Hongming Zhang, and Yingchen Zhang, *Senior Member, IEEE*

***Abstract*—It is found from actual data that due to generation dispatch and uncertain renewable generations and loads with complicated correlations, inferring the probabilistic distributions for uncertain inputs is challenging. Many probabilistic power flow approaches have been developed in the literature but their validations using realistic systems and data are lacking. This paper proposes a data-driven probabilistic analysis approach for system risk assessment of the miniWECC system using actual data. The sparse Gaussian process (SGP) is advocated to quantify the impacts of uncertain inputs on voltage security. SGP does not need the probability distribution function of uncertain inputs, can handle correlations and is highly computationally efficient. Results on the miniWECC system using realistic data show that SGP outperforms existing approaches and is able to quantify the voltage violation risks.**

***Index Terms*—MiniWECC, probabilistic power flow, renewable energy, sparse Gaussian process, uncertainty quantification.**

I. Introduction

As MANY uncertain sources are connected, such as flexible loads, solar and wind, power system responses are subject to stochasticity. These uncertainties may lead to voltage violations and need to be carefully monitored. To this end, several probabilistic power flow approaches have been developed. Monte Carlo (MC)-based sampling methods and their variants [1], [2] are the basic probabilistic analysis tools. But they need the accurate probability distribution function (PDF) of uncertain inputs and are computationally expensive. Other analytical methods, such as cumulants [3] derive PDF by simplification and linearization, yielding large errors in the presence of large uncertainties. Polynomial chaos expansion (PCE) [4] and Gaussian process (GP) modeling [5] have been recently demonstrated to have excellent performances in dealing with uncertainty quantification problems. PCE approximates the model with orthogonal polynomials but requires the knowledge of uncertain input distributions and correlations. GP is data-driven but its computational cost is high for high-dimension inputs.

It is found from the realistic data in Western Electricity Coordinating Council (WECC) systems that the uncertainty patterns for renewable energy resources are rather stochastic and difficult to be represented by standard PDFs. There are also complicated correlation relationships among uncertain resources, which cannot be simply approximated by existing linear correlations. Due to the scale of the system, the dimension of the uncertain inputs is high, causing computational challenges. Although it is possible to leverage the sparsity for improving the computational efficiency of the sampling-based and PCE-based approaches [6], the curse of dimensionality issue still exists, making them not scalable to large-scale systems. Furthermore, these approaches need accurate PDFs of uncertain inputs that are difficult to obtain in practice. This letter addresses these concerns and develops a scalable data-driven sparse GP (SGP) for realistic miniWECC system voltage violation risk assessment. This is achieved via the stochastic variational inference that integrates the variational inference and inducing variables into the GP framework. Unlike the original GP method that has cubic time complexity [7], SGP uses a small number of inducing variables for variational inference approximation without loss of accuracy, significantly enhancing the computational efficiency. Note that the proposed approach has been tested in the miniWECC system, a reduced-order model of the original WECC system. Actual historical data have been used to test the performance of various approaches, another major contribution. It is observed that the proposed method can more accurately capture the voltage violation risk while being computationally more efficient than other existing approaches.

II. Problem Statement

Probabilistic analysis aims to provide statistical information of power system with uncertainties. The power flow model can be described as: $y = \mathcal{M}(\boldsymbol{x})$, where the input $\boldsymbol{x}$ contains uncertainties, such as uncertain loads, wind generations and PVs; $\mathcal{M}$ is the nonlinear function between inputs and outputs; the model output y is the output of interest, such as voltage magnitude or line flow. It is worth pointing out that existing probabilistic power flow approaches are typically tested using simulation data under standard test systems and may not reflect all characteristics of the actual systems. For example, it is found from the realistic data in the miniWECC system that

Manuscript received 15 January 2022; revised 10 April 2022; accepted 14 June 2022. Date of publication 17 June 2022; date of current version 19 August 2022. This work was supported in part by the U.S. Department of Energy Wind Energy Technology Office and National Science Foundation under Grant ECCS-1917308. Paper no. PESL-00017-2022. *(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).

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

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPWRS.2022.3184124.

Digital Object Identifier 10.1109/TPWRS.2022.3184124

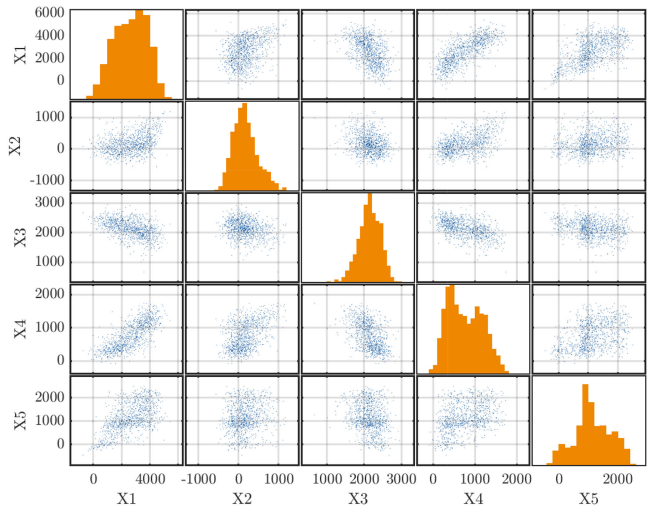

Fig. 1. Complex dependence among selected 5 uncertain sources, where the units for x and y axis are MW.

the generation patterns are rather stochastic and have some jumps due to generation dispatch, which make it difficult to infer the PDFs for existing none data-driven approaches; there are nonlinear correlations among uncertain PVs that are geographically close, see Fig. 1; the number of uncertain inputs is high, causing challenges for data-driven GP approaches. For the sampling-based and the PCE-based approaches that require exact joint distribution of the uncertain inputs, the kernel density estimation approach and copula statistics [6] can be utilized to infer the marginal probability distribution and model the nonlinear correlation among inputs, respectively. However, the probability distribution for uncertain inputs in the realistic system is typically not the standard distributions and the inference can bring additional errors for uncertain input modeling. This letter develops a data-driven SGP approach to address the aforementioned problems. By leveraging data-driven feature of GP and the sparsity technique via variational inference and inducing points, SGP overcomes the dimensionality problem of GP with further improved accuracy. More details are shown in the next section.

III. SGP for MiniWECC System

GP-based method has been used for probabilistic power flow but has difficulties in handling high-dimensional uncertain inputs. This letter advocates the SGP with stochastic variational inference approach without the need of PDFs of uncertain inputs while being computationally efficient.

GP modeling attempts to describe the model output into $\boldsymbol{y}=f(\boldsymbol{X})+\boldsymbol{\epsilon}$, where $\boldsymbol{X}\in\mathcal{R}^{N\times d}$ and $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma_\epsilon^2\boldsymbol{I})$; f is defined by the mean function and the covariance function [7]:

$$f\sim\mathcal{GP}(m(\boldsymbol{x}),k(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\theta})) \tag{1}$$

where $f(\cdot)$ is the nonlinear mapping function; the mean function $m(\boldsymbol{x})$ is commonly expressed in the form of polynomial function and the covariance function $k(\cdot)$ sets the covariance between points between $\boldsymbol{x}$ and $\boldsymbol{x}'$; this covariance function allows us to capture the correlations among inputs, being linear or nonlinear dependence; $\boldsymbol{x}'$ represent the samples excluding $\boldsymbol{x}$ and $\boldsymbol{\theta}$ represents the GP parameters. The parameters of GP model are learned by maximum likelihood estimation as the posterior is derived from the prior and the likelihood based on Bayes' rule. As a non-parametric method, GP accomplishes the regression in a data-driven manner. The performance is generally better than than sampling-based and PCE-based methods with less needed number of samples for uncertain inputs. However, GP-based method suffers from scalability issue as its model complexity reaches $\mathcal{O}(n^3)$, where n is the dimension of the uncertain inputs [7].

To address the scalability issues of GP in the presence of high dimensional uncertain inputs, some SGP have been developed in computer science field [8], [9] but not been applied for addressing power system applications. Specifically, a set of inducing points $\boldsymbol{Z}=\{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_l\}$ is introduced with function value $\mathbf{u}=f(\boldsymbol{Z})$. Under the assumption that $\mathbf{f}$ and $\mathbf{f}_*$ are conditionally independent given inducing points $\mathbf{u}$, the joint distribution $p(\cdot)$ can be approximated by the inferred probability distribution $q(\cdot)$:

$$p(\mathbf{f},\mathbf{f}_*)\simeq q(\mathbf{f},\mathbf{f}_*)=\int q(\mathbf{f}|\mathbf{u})q(\mathbf{f}_*|\mathbf{u})\,\mathrm{d}\mathbf{u} \tag{2}$$

The model complexity is reduced to $\mathcal{O}(nl^2)$ by using inducing points, where the dimension of the inducing points is l, a much smaller number as compared to n. This explains theoretically SGP is much more scalable to larger-scale power systems as compared to GP. The true posterior of GP is approximated by minimizing the Kullback-Leibler (KL) divergence $\mathrm{KL}[q(\mathbf{f},\mathbf{u})||p(\mathbf{f},\mathbf{u}|\boldsymbol{X})]$.

This letter advocates to embed the stochastic variational inference into the probabilistic framework to further speed up the inference. The stochastic variational inference is developed based on the combination of variational inference and inducing variables [9]. Formally, the variational posterior $q(\mathbf{f},\mathbf{u})=p(\mathbf{f}|\mathbf{u})q(\mathbf{u})$ becomes the following equation after marginalizing $\mathbf{u}$:

$$q(\mathbf{f}|\mathbf{m}_q,\mathbf{S})=\int p(\mathbf{f}|\mathbf{u})q(\mathbf{u})\,\mathrm{d}\mathbf{u} \tag{3}$$

in which $q(\mathbf{u})=\mathcal{N}(\mathbf{u}|\mathbf{m}_q,\mathbf{S})$. The mean and variance are obtained as

$$\begin{cases}\mu_*=\mathbf{m}_q+\boldsymbol{P}\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{K}_{\boldsymbol{ZX}}(\mathbf{m}_q-\mathbf{m}_{\boldsymbol{Z}})\\ \sigma_*^2=\boldsymbol{K}-\boldsymbol{P}^{\mathrm{T}}(\boldsymbol{K}_{\boldsymbol{ZZ}}-\mathbf{S})\boldsymbol{P}\end{cases} \tag{4}$$

where $\boldsymbol{P}=\boldsymbol{K}_{\boldsymbol{XZ}}\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}$; $\boldsymbol{K}_{\boldsymbol{XZ}}$ and $\boldsymbol{K}_{\boldsymbol{ZZ}}$ represent the cross covariance matrix between inducing points and training points and self-covariance matrix of the inducing points, respectively. As shown in [10], minimizing the KL divergence between variational posterior q and the true posterior p is equivalent to maximizing the lower bound of the true log marginal likelihood shown below:

$$\mathcal{L}=\sum_{i=1}^{N}\mathbb{E}_{q(f_i|\mathbf{m}_q,\mathbf{S})}\left[\log p(x_i|f_i)\right]-\mathrm{KL}\left[q(\mathbf{u})||p(\mathbf{u})\right] \tag{5}$$

where $\mathbb{E}(\cdot)$ is the expectation operator. Maximizing (5) yields the estimated variational parameters $(\boldsymbol{Z}, \mathbf{m}_q, \mathbf{S})$. To enhance robustness with limited samples, r-fold cross-validation is utilized for parameter estimation. Thanks to the significantly reduced computational complexity, the proposed SGP allows us for voltage risk assessment of large-scale power systems with high dimension uncertain inputs.

IV. Test Results

The effectiveness of the proposed method is tested on the miniWECC system using realistic data. The miniWECC system is a reduced-order system of WECC system and it contains 243 buses, 146 generating units (including 109 synchronous machines and 37 renewable generators), 329 transmission lines, 122 transformers, 7 switched shunts and 139 loads [11]. Note that the miniWECC system is a reduced model for WECC by aggregating nodes that are below certain voltage level. National Renewable Energy Lab has developed an approach to map the original WECC system historical data into this miniWECC system. The "real" data for the miniWECC system refers to true historical data mapped from the original WECC system. The uncertain sources include all loads and renewable generators. This paper specifically focuses on the voltage violation risk, a concern for WECC during operations. A total number of 720 samples are obtained by security-constrained economic dispatch at 1 h interval over a 30 d period. Various types of methods are applied for comparisons, including sparse PCE (SPCE) [6] and GP [5], Latin hypercube sampling (LHS), and quasi-MC [1]. The model error (e_M) is evaluated using the mean absolute percentage error (MAPE) index:

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i^* - \hat{y}_i}{y_i^*}\right| \times 100\% \tag{6}$$

where y^* and $\widehat{y}$ are the true and estimated voltage magnitudes, respectively. For statistical estimates, MAPE over sample mean and variance are used, denoted as e_μ and e_{σ^2}. The error indices are computed based on all bus voltage magnitudes but only the voltage magnitude at bus 2202 is illustrated due to the lack of space. The benchmark is obtained using Monte Carlo simulations with all historical data. All simulations are carried out in MATLAB with 2.60 GHz Intel Core i7-6700HQ.

The PDF of voltage magnitude is obtained by kernel density estimator (KDE). The degrees of PCE and SPCE are set to be $n=3$ and 360 (50%) samples are used for model construction. For SGP, linear trend is used for mean function and Matérn kernel for covariance function. The parameters are estimated via cross-validation. The training dataset consists of 288 (40%) samples and 72 (10%) inducing points. For sampling methods, enough samples are used to reach similar accuracy with SGP and therefore the SGP's computational advantage will be emphasized.

The benchmark shows a low voltage at bus 2202 and there exists a case, where voltage magnitude is below 0.9, which may cause severe issues in system operation. This is reflected in the tail of PDF. From Fig. 2 and Table I, it can be observed that the original PCE has the worst performance as it cannot handle

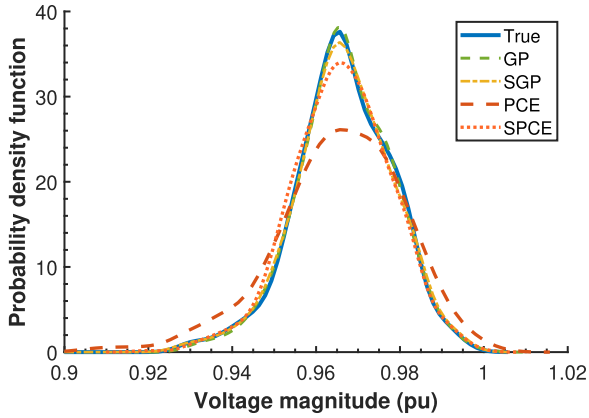

Fig. 2. Comparison results of different methods in estimating the PDF of voltage magnitude at bus 2202.

TABLE I
Comparison Results of Different Methods on MiniWECC System

Method	MAPE			CPU Time (s)
	$e_M(\%)$	$e_\mu(\times 10^{-2})$	$e_{\sigma^2}(\%)$	
PCE	8.4	6.29	59.92	735.66
SPCE	2.6	2.01	16.26	39.73
GP	0.11	0.72	2.47	125.68
SGP	0.13	0.76	2.73	13.96
LHS	0.19	0.95	2.98	90.50
QMC	0.18	0.88	2.87	89.68

complicated correlations among uncertin inputs. SPCE achieves better performance and lower computational cost than vanilla PCE but has large errors at the peak and tail distributions. Note that PCE and SPCE need accurate PDFs of uncertain inputs, a difficult task in practice. By contrast, GP is able to approximate the PDF with much better results, which demonstrates its advantages on accuracy. SGP has a similar performance but being much more computationally efficient with the help of sparse methods. SGP can also capture the tail distributions, which is important for successfully predicting the extreme case, such as voltage violations, i.e., voltage magnitude below 0.9 (around 0.87). Comparing the CPU times of two GP-based methods in Table I indicates that traditional GP is not scalable to handle high dimension uncertain inputs in the miniWECC system while SGP is adequate for high-dimensional data by reducing 90% of CPU time.

For sampling-based methods, LHS obtains acceptable accuracy with less samples as compared to the original MC and QMC is based on low-discrepancy sequence, i.e., Sobol sequence in this letter. It can be observed in Fig. 3 that QMC based on Sobol sequence slightly outperforms LHS. As shown in Table I, the efficiency of enhanced sampling methods is highly dependent on the used number of samples. However, both sampling-based methods can hardly catch extreme cases with small number of samples, i.e., a large number of samples is needed that can be very time consuming. In addition, the sampling-based methods also need accurate PDFs of uncertain inputs and this can be very

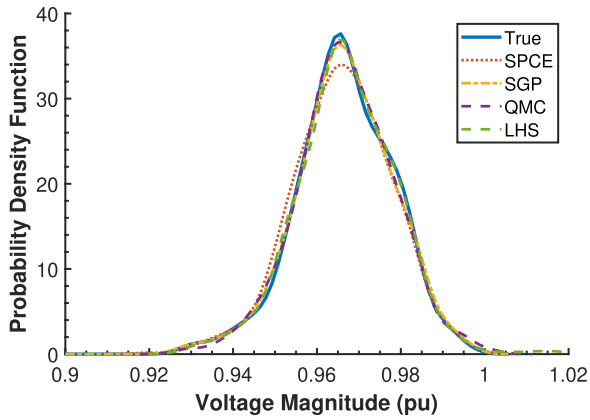

Fig. 3. Comparison results with sampling-based methods in terms of the PDF of voltage magnitude at bus 2202.

difficult if the number of historical data is small. By contrast, SGP achieves the best performance in terms of percentage error and computational efficiency as shown in Table I. It is worth noting that SGP is the only method that accurately depicts the tail of PDF, which is helpful for future risk analysis and preventive control.

In summary, in terms of better accuracy, both GP and SGP are data-driven approaches, and the probability distribution of uncertain inputs is not required. Thus, the sampling errors from the probability distribution of uncertain inputs are mitigated as compared to model-based and PCE-based approaches. The GP has issue of handling high dimensional uncertain inputs while SGP addresses this via sparse technique, yielding better accuracy. In terms of computational efficiency, the sampling-based methods need a large number of samplings from the probability distribution of uncertain inputs, and this leads to higher computational demand as compared to GP-based approaches. Theoretically, the computational complexity of the original GP method has cubic computational complexity. With variational inference and sparse techniques, the proposed method has linear computational complexity. Thanks to this, the method is scalable to larger-scale power systems and has high computational efficiency.

V. Conclusion

This letter shows several realistic challenges for power system probabilistic analysis, including strong stochasticity of generation patterns, complicated correlation relationships among uncertain inputs and high dimensional uncertain inputs. It is shown that many existing approaches perform well in standard test systems but obtain significantly degraded performance using actual data. This is mainly because some assumptions behind these algorithms may not always hold for true in practice. To this end, this letter develops a sparse GP for the probabilistic analysis of miniWECC system. Test results on the miniWECC system show that the proposed method can accurately capture the system risk in the tails of PDFs while being highly computationally efficient than other approaches. There can be other approaches that can be used together with sparse GP available to further reduce the computational burden at the cost of accuracy if the size of system keeps growing, such as sparse kernel. We will investigate this in our future work.

References

[1] B. R. Prusty and D. Jena, "A critical review on probabilistic load flow studies in uncertainty constrained power systems with photovoltaic generation and a new approach," *Renewable Sustain. Energy Rev.*, vol. 69, pp. 1286–1302, Mar. 2017.

[2] X. Xu and Z. Yan, "Probabilistic load flow calculation with quasi-Monte Carlo and multiple linear regression," *Int. J. Elect. Power Energy Syst.*, vol. 88, pp. 1–12, Jun. 2017.

[3] M. Fan, V. Vittal, G. T. Heydt, and R. Ayyanar, "Probabilistic power flow studies for transmission systems with photovoltaic generation using cumulants," *IEEE Trans. Power Syst.*, vol. 27, no. 4, pp. 2251–2261, Nov. 2012.

[4] J. Laowanitwattana and S. Uatrongjit, "Probabilistic power flow analysis based on partial least square and arbitrary polynomial chaos expansion," *IEEE Trans. Power Syst.*, vol. 37, no. 2, pp. 1461–1470, Mar. 2022.

[5] Y. Xu, Z. Hu, L. Mili, M. Korkali, and X. Chen, "Probabilistic power flow based on a Gaussian process emulator," *IEEE Trans. Power Syst.*, vol. 35, no. 4, pp. 3278–3281, Jul. 2020.

[6] K. Ye, J. Zhao, Y. Zhang, X. Liu, and H. Zhang, "A generalized computationally efficient copula-polynomial chaos framework for probabilistic power flow considering nonlinear correlations of PV injections," *Int. J. Elect. Power Energy Syst.*, vol. 136, Mar. 2022, Art. no. 107727.

[7] C. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, U.K.: MIT Press, 2006.

[8] E. Snelson and Z. Ghahramani, "Sparse Gaussian processes using pseudo-inputs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, vol. 18, pp. 1259–1266.

[9] J. Hensman, N. Fusi, and N. D. Lawrence, "Gaussian processes for Big Data," in *Proc. 29th Conf. Uncertainty Artif. Intell.*, 2013, pp. 282–290.

[10] M. K. Titsias, "Variational learning of inducing variables in sparse Gaussian processes," in *Proc. 12th Int. Conf. Artif. Intell. Statist.*, 2009, pp. 567–574.

[11] H. Yuan, R. Sen biswas, J. Tan, and Y. Zhang, "Developing a reduced 240-bus WECC dynamic model for frequency response study of high renewable integration," in *Proc. IEEE PES Transmiss. Distrib. Conf. Expo.*, 2020, pp. 1–5.