Probabilistic Modeling for Optimization of Resource Mix with Variable Generation and Storage

Weixuan Gao, *Student Member, IEEE,* and Dimitry Gorinevsky, *Fellow, IEEE,*

***Abstract*—Renewables, such as solar and wind generation, combined with storage are becoming a key part of modern grid. This paper develops probabilistic tools for analysis of grid reliability with such variable generation resources. The developed tools improve speed and accuracy of the reliability analysis compared to usual Monte Carlo methods. This is achieved by using an extension of well known convolution method applicable to interdependent variables. The interdependent distributions are obtained from historical data using Machine Learning of quantile models. The paper presents a novel approach to the analysis of reliability contribution of storage based on these models and related to Information Theory. The developed tools are demonstrated in several example scenarios for ISO-New England service area.**

***Index Terms*—LOLE, Machine Learning, Optimization, Power Grid, Renewables, Resource Mix, Ramp Rate, Storage**

I. Introduction

Renewable energy has large and increasing impacts on the grid. California plans to go carbon-free by 2045, Massachusetts considers similar legislation. Solar and wind resources are non-dispatchable and bring random variability. To compensate, more energy storage gets connected to the grid. Managing risk and cost of the grid with storage and variable generation at scales beyond anything seen before requires new probabilistic analysis tools. This paper presents tools for planning and market design using machine learning of probabilistic models from historical data. The tools can support capacity investments and other planning studies and enable optimization of resource mix scenarios supporting NERC 1-in-10 reliability requirement (one day of outage in ten years).

Probabilistic reliability analysis of grid can be done by Monte Carlo or analytical methods [1]. Predominantly, Monte Carlo methods are used in practice because they allow to encode complex analysis scenarios. Some examples include tools like StorageVet, DER-CAM [2], and SERVM [3]. One issue with Monte Carlo methods is that heavy computing is required to evaluate risk for distribution tails (peak events). This impedes their use for iterative resource mix optimization, especially in scenarios far different from historical data.

Analytical methods are based on discrete convolution, e.g., see [1]. Prior work addresses several subproblems of probabilistic risk analysis for grid with renewables and storage. Data-driven models for each hour of the year are discussed in [4]. For a given hour, load and variable resources are assumed independent and respective empirical distributions convolved. Such approach is used in operational forecasting, e.g., see [5] and [6]. Other prior work combined analytical and Monte Carlo methods, e.g., see [7], [8].

Existing methods need to be extended in two important ways. First issue is accounting for the interdependence of load with solar and wind generation, e.g., see [9], [10]. Machine learning of such interdependence is discussed below. Second need is rigorous probabilistic analysis of energy storage contribution. Prior work on grid reliability with energy storage is mostly based on deterministic simulation analysis of storage dispatch, e.g., see [11]. Analytical methods for energy adequacy and reliability with storage seem to be limited to the earlier work of the authors [12], which is extended in this paper.

Accuracy of both analytical and Monte Carlo methods is predicated or careful modeling of underlying probability distributions. To model the distributions, this paper uses Machine Learning methods related to quantile regression. Quantile regression for conditional quantiles of dependent variable was introduced in [13] and is now widely used in statistics. Multiple quantile regression and related issues of quantile level crossing are discussed in [14]–[16]. Quantile regression modeling of distribution tails is discussed in [16], [17]. Multivariate quantile regression was introduced in [12], [18]. Multiple quantile regression was used to forecast demand and renewable energy in [19], [20].

Analytical methods in this paper use probability models conditional on time regressors (such as hour and month). These quantile models are related to the load (or variable generation) duration curves for a given hour. Machine Learning methods used to obtain such quantile models from limited historical data are based on convex optimization and Extreme Value Theory for the tails; they are described in earlier work of the authors [12], [21]–[23].

This archival paper incorporates results from conference papers [12], [21] to provide for self-contained presentation with a view of practical use. Over half of this paper is new material. Section II-C, Section III with exception of III-C and III-E, Section IV, Section V-C, and Section VI have not been published earlier. The two main contributions of the paper are as follows. First is a machine learning method for modeling of interdependent random variables that also depend on time. It is used for probabilistic modeling of demand, solar, and wind generation. Using these models, adequacy of the resource mix can be assessed with less effort and more accurately than with existing approaches. Second, this paper provides probabilistic assessment of energy storage in grid reliability using approach

D. Gorinevsky is with the Department of Electrical Engineering, Stanford University, CA, 94305, e-mail: gorin@stanford.edu.

W. Gao is with Civil and Environmental Engineering Department, Stanford University, CA, 94305, e-mail: gaow@stanford.edu.

This work was supported by Seed Grant of Stanford Bits and Watts Initiative
Manuscript received in OCTOBER 2019.

connected to Information Theory.

Section III of this paper describes Machine Learning for component models using historical data. Section IV provides statistical validation of the developed models demonstrating their predictive power and improvements over a simple Monte Carlo approach. For a given resource mix, the component models are combined to get the distribution of reserve margin and compute the risk; this is introduced in Section II. Section V provides several grid planning examples. For each example, the analysis provides scenario with desired level of reliability (risk).

II. Risk Estimation

This section introduces analytical methods for grid reliability. The reserve margin distribution is evaluated from component models for demand, solar, and wind generation. The known convolution approach is extended to use of interdependent conditional models.

A. Sampled Distribution

Consider random variable u with known probability distribution. The probabilistic analysis is based on sampling the Cumulative Distribution Function (CDF) F_u as

$$F_u[k] = \mathbf{P}(u \le k\Delta h), \quad p_u[k] = F_u[k] - F_u[k-1], \tag{1}$$

where k is an integer sample number and p_u is the Probability Mass Function (PMF) for the sampled CDF F_u. If sampling step Δh is small enough, the sampled model is sufficiently accurate. Continuous distribution CDF can be recovered by interpolating F_u between the samples.

If independent random variables are sampled on the same grid, the probability mass function (PMF) of their sum is a convolution of the individual PMFs. As an example, outage distributions for a set of power generating units are described by independent Bernoulli distributions for each unit. The distribution of the total outage power can be computed as a chain convolution of these Bernoulli distributions sampled in accordance with (1), see [12]. Such convolution method has been used for grid reliability analysis for over two decades, see [1]. This paper extends the convolution method applications to include solar and wind generation as well as storage.

B. Conditional Distribution

Consider a sampled conditional random variable $u|Z$, where explanatory variables (regressors) Z are in a finite state space with states Z_j that have probabilities $\mathbf{P}(Z_j)$. For example, the model for the load (electricity demand) is conditional on vector

$$Z = \mathrm{col}(Z_M, Z_W, Z_H, Z_{Hol}) \in \Re^{45}, \tag{2}$$

which includes 12 binary regressors indicating calendar months Z_M, 7 weekday indicators Z_W, 24 indicators for hours of the day Z_H, and 2 indicators for holiday or not Z_{Hol}. There are a total of $12 \times 7 \times 24 \times 2 = 4,032$ different states Z_j. Such regressors have been used in many papers on the subject, e.g., see [23], [24] where further references can be found. In what follows, conditional distribution for wind and solar are modeled using Months and Hours as regressors; wind and solar do not depend on Weekday and Holiday. For each state Z_j, there is a conditional variable with PMF

$$u_j = u|Z_j, \quad u_j \sim p_{u_j}[\,\cdot\,]. \tag{3}$$

For two conditionally independent variables $u|Z$ and $v|Z$, distribution of $(u+v)|Z$ can be computed as convolution

$$p_{u+v|Z_j}[\,\cdot\,] = p_{u|Z_j}[\,\cdot\,] * p_{v|Z_j}[\,\cdot\,]. \tag{4}$$

The sum of the two variables can be described by conditional distributions for all N states Z_j computed as the convolutions.

C. Linear Interdependent Model

Consider two interdependent variables Y and X. By Bayes's Theorem, we have $\mathbf{P}(Y,X) = \mathbf{P}(Y|X)\mathbf{P}(X)$. Assume linear quantile model for the variable $Y|X$. (Estimating such models from data is described in Section III). The model is

$$\mathbf{P}_{Y|X}(y \le \alpha_q + \gamma x | x) = q, \tag{5}$$

where α_q is the quantile determined by q and γ is a coefficient. This is equivalent to

$$\mathbf{P}_{\tilde{Y}}(\tilde{y} \le \alpha_q) = q, \tag{6}$$

where $\tilde{Y} = Y - \gamma X$ is a new random variable. The joint distribution of variables X and Y can be represented as

$$\mathbf{P}(Y,X) = \mathbf{P}(\tilde{Y},X) = \mathbf{P}(\tilde{Y})\mathbf{P}(X), \tag{7}$$

This means random variables $\tilde{Y}$ and X are independent.

This derivation can be extended to multiple dependent variables by considering vectors $X = \mathrm{col}(X_1, X_2, ..., X_n)$ and $Y = \mathrm{col}(Y_1, Y_2, ..., Y_n)$. By using chain rule, (7) is extended to a product of n independent distributions,

$$\mathbf{P}(Y) = \mathbf{P}(\tilde{Y}_1)\mathbf{P}(\tilde{Y}_2)...\mathbf{P}(\tilde{Y}_n), \tag{8}$$

$$\tilde{Y}_k = Y_k - \sum_{i=k+1}^{n} \gamma_{k,i} Y_i. \tag{9}$$

Using vector notation $\tilde{Y} = \mathrm{col}(\tilde{Y}_1, \tilde{Y}_2, ..., \tilde{Y}_n)$, we get $\tilde{Y} = Y - \Gamma Y$ with upper diagonal coefficient matrix Γ,

$$\Gamma = \begin{bmatrix} 0 & \gamma_{12} & \gamma_{13} & \cdots & \gamma_{1n} \\ 0 & 0 & \gamma_{23} & \cdots & \gamma_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}. \tag{10}$$

Solving $\tilde{Y} = Y - \Gamma Y$ for Y yields

$$Y = (1 - \Gamma)^{-1} \tilde{Y}. \tag{11}$$

This allows computing distributions for sums of dependent variables Y_j through sums of independent variables $\tilde{Y}_j$,

$$\sum_j Y_j = \sum_j \tilde{\gamma}_j \tilde{Y}_j, \tag{12}$$

Computing (12) as $\mathbb{1}^T Y$ and using (11) yields

$$\mathrm{col}(\tilde{\gamma}_1, \tilde{\gamma}_2, ..., \tilde{\gamma}_n) = \mathbb{1}^T (1 - \Gamma)^{-1}. \tag{13}$$

A distribution of sum in (12) can be computed by convolving the PMFs of independent variables $\tilde{\gamma}_j \tilde{Y}_j$, see (12).

As an application example, consider three interdependent variables: demand (load) L, wind generation W, and solar generation S. To describe dependency between L, W, S assume quantile models of the form (5)

$$\mathbf{P}_{L|W,S}(L \le \alpha_{L,q} + \gamma_{LW}W + \gamma_{LS}S|W,S) = q, \tag{14}$$

$$\mathbf{P}_{W|S}(W \le \alpha_{W,q} + \gamma_{WS}S|S) = q, \tag{15}$$

$$\mathbf{P}_{S}(S \le \alpha_{S,q}) = q. \tag{16}$$

Section III describes how quantile models (14), (15), (16) can be estimated from historical data. Independent variables can be obtained through interdependent variables in (14)-(16) similar to (9)

$$\tilde{L} = L - \gamma_{LW}W - \gamma_{LS}S, \tag{17}$$

$$\tilde{W} = W - \gamma_{WS}S, \tag{18}$$

$$\tilde{S} = S. \tag{19}$$

Conversely, dependent variables W, S, and $-L$ can be expressed through the independent variables following the logic of (12). This yields an expression for a 'net injection' sum

$$W + S - L = \tilde{\gamma}_W\tilde{W} + \tilde{\gamma}_S\tilde{S} + \tilde{\gamma}_L\tilde{L}, \tag{20}$$

$$\tilde{\gamma}_S = 1 - \gamma_{LS} + \gamma_{WS} - \gamma_{LS}\gamma_{WS}, \tag{21}$$

$$\tilde{\gamma}_W = 1 - \gamma_{WS}, \tag{22}$$

$$\tilde{\gamma}_L = -1. \tag{23}$$

The distribution of $W+S-L$ can be computed by convolving PDFs of the independent variables $\tilde{\gamma}_L\tilde{L}$, $\tilde{\gamma}_W\tilde{W}$, and $\tilde{\gamma}_S\tilde{S}$ as

$$p_{W+S-L}[\cdot] = p_{\tilde{\gamma}_W\tilde{W}}[\cdot] * p_{\tilde{\gamma}_S\tilde{S}}[\cdot] * p_{\tilde{\gamma}_L\tilde{L}}[\cdot]. \tag{24}$$

Section III presents Machine Learning approach to estimating the joint distribution of the three variables L, W and S conditional on Z. The distribution of sum (20) can be then computed through the convolutions (24) for each of regressor states Z_j. More complex copula models for similar purpose are proposed in [10]; such models are much harder to estimate from data, however.

D. Evaluate Risk

This paper computes Loss of Load Hours (LOLH) reliability index, which is the expected number of hours for loss of load per year [25]. The 1-in-10 NERC requirement corresponds to having LOLH ≤ 2.4. Other risk analysis indexes for the power grid, such as Expected Energy Not Served, can be computed from the same distributions as LOLH is in this paper.

Reserve margin R measures the generation capacity over and above the demand. Negative reserve margin implies that the generation cannot meet the demand. Assume that sampled conditional distribution for R is available in form (3). The conditional distribution of the reserve margin $R|Z$ allows computing Loss of Load Probability (LOLP) as

$$\text{LOLP}(Z_j) = \mathbf{P}(R < 0|Z_j) = \sum_{R_j \le 0} p_{R_j}[k]. \tag{25}$$

Using (25), we can compute LOLH = $\mathbf{P}(R < 0)$ as

$$\text{LOLH} = \mathbf{E}(\mathbf{P}(R < 0|Z)) = \sum_{j=1}^{m}\sum_{R_j \le 0} p_{R_j}[k] \cdot \mathbf{P}(Z_j), \tag{26}$$

where probability of states Z_j with Holiday indicated is $\mathbf{P}(Z_j) = p_H/2016$; the rest, $\mathbf{P}(Z_j) = (1 - p_H)/2016$; and holiday probability is $p_H = 9/365$ is the holiday probability; see Subsection II-B.

III. Machine Learning Methods

Section II discusses computation of risk given the distribution models. This section shows how such models can be built from historical data. The models are conditional distribution versions of (14)-(16). A starting point of the method is learning a distribution for $u|Z$, where u is the random variable and Z is explanatory variable (regressor) vector, from historical data

$$\mathscr{D} \equiv \{u_i, Z_i\}_{i=1}^{N}. \tag{27}$$

A. Modeling Data

The examples illustrating the proposed approach use data for ISO New England (ISO-NE) service area. Publicly available ISO-NE data include time series for load [26] and wind generation [27]. Historical solar generation was obtained from Renewables Ninja [28] based on MERRA-2 dataset assuming non-tracking solar panels with 10% loss, tilt angle of 35°, and azimuth of 180° spread over four locations in New England. We modeled load intensity, wind intensity, and solar intensity as non-dimensional variables on $[0,1]$ interval. Since the annual demand is decreasing around 2.3% per year from 2015 to 2017, load intensity scale factor decreases the same amount. The wind intensity is scaled by wind generation nameplate: 750MW, 880MW, and $1,005$MW in years 2015–2017.

B. Quantile Bins Modeling

Quantile Bins is the simplest quantile model formulation. It also requires the most data to estimate. This approach works well for solar generation modeling, e.g., see (16). For solar (and wind) modeling, the weekday and holiday regressors in (2) are irrelevant. The remaining regressors are $Z_M \in \mathfrak{R}^{12}$, binary variable indicating calendar month, and $Z_H \in \mathfrak{R}^{24}$, binary indicator of hour. The regressor state is then described by the direct product vector

$$Z_{MH} = Z_M \otimes Z_H \in \mathfrak{R}^{288}. \tag{28}$$

For any regressor vector Z_j, state vector $Z_{MH,j}$ (28) indicates the month and hour. Let $\mathbf{J}$ be a set of indexes i in data set (27) such that $Z_i = Z_j$, The quantiles of the distribution $u|Z_{MH,j}$ can be found by solving m separate Linear Program (LP) problems for each quantile level q_k (e.g., see [13])

$$\alpha_{k,j} = \arg\min \sum_{i \in \mathbf{J}} h(u_i - \alpha_{k,j}; q_k) \tag{29}$$

$$h(y;q) = \tfrac{1}{2}|y| + (q - \tfrac{1}{2})y, \tag{30}$$

where $h(\cdot;q)$ is the 'pinball' function. Model parameters $\alpha_{k,j}$ are quantiles that correspond to quantile levels $q_k \in (0,1)$, $(k = 1,...,m)$. Separate multiple quantile models $\{\alpha_{k,j}\}_{k=1}^{m}$ are estimated for all 288 distinct combinations of month and hour $Z_{MH,j}$. Sufficient data is needed to achieve the accurate estimation of the $288 \times m$ parameters $\alpha_{k,j}$.

We used 5 years of hourly data from 2011 to 2016 to train quantile bins model for solar generation. This provides around 152 data points to estimate multiple quantile model for each regressor state $Z_{MH,j}$. Figure 1 shows the predicted quantiles of solar intensity and example data for the day with peak solar generation in year 2016. In Figure 1, Z_M is fixed and corresponds to month of August, while Z_H is defined by the hour shown as plot argument. For each hour, $\alpha_{k,j}$ is computed and shown, where k indexes quantile q_k in (29) and j corresponds to $Z_{MH,j} = Z_M \otimes Z_H$.

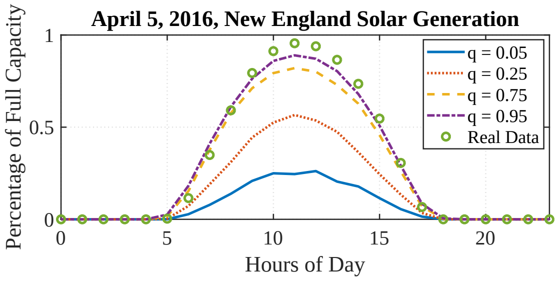

Fig. 1: Solar Generation of a Sample Day

C. Multiple QR Model

As a starting point for Machine Learning of interdependent conditional distributions (14) and (15), consider Quantile Regression (QR) model for random variable u conditional on regressor vector Z, of the form

$$\mathbf{P}(y \le y(q)|Z) = q, \quad y(q) = Z\beta(q) + \alpha(q), \tag{31}$$

where $q \in (0,1)$ is quantile level; β is parameter vector in the dot product $Z\beta$ and α is scalar. For a given quantile level q, model (31) can be found by solving an LP problem, see [9].

To estimate distribution at multiple quantile levels q_k, Multiple QR problem was formulated in [29] as an optimization problem in Second Order Cone Program (SOCP) form. Model parameters, scalars α_k and vectors β_k, for quantiles q_k ($i = 1,\dots,m$) are minimizers for the following SOCP

$$\{\alpha_k, \beta_k\}_{k=1}^m = \arg\min \sum_{k=1}^m \sum_{i=1}^N h(u_i - Z_i\beta_k - \alpha_k; q_k) \tag{32}$$
$$+ \lambda \sum_{l=2}^m \|\beta_l - \beta_{l-1}\|^2 + \mu \sum_{j=2}^{m-1} (\alpha_{j+1} - 2\alpha_j + \alpha_{j-1})^2.$$

The solution of (32) and selection of regularization parameters λ and μ are discussed in [12].

D. Linear Dependent Model Estimation

Power demand, solar, and wind generation are interdependent. Sun influences the wind, while both sun and wind affect the demand. Subsection II-C introduced linear interdependent model for demand, solar, and wind.

Machine Learning method below estimates such model for (14), (15) from historical data of the form

$$\mathcal{D}_W \equiv \{l_i, w_i, s_i, Z_i\}_{i=1}^N, \tag{33}$$

where i is the index of hourly data sample; l_i, w_i, s_i are samples of demand, wind, and solar; Z_i are samples of explanatory variable (regressor) Z.

For a given vector Z (2), model (14)-(23) expresses interdependent random variables $-L$, W, S through independent random variables $\tilde{L}$, $\tilde{W}$, $\tilde{S}$ using model weights $\tilde{\gamma}_S(Z)$, $\tilde{\gamma}_W(Z)$, $\tilde{\gamma}_L(Z)$. This subsection assumes that dependence on Z is fully described by the month and hour indicators Z_{MH} (28). We group all unknown model weights for the 288 distinct states $Z_{MH,j}$ in parameter vectors

$$\Gamma_{LW} = \text{col}(\gamma_{LW}^{(1)}, \dots, \gamma_{LW}^{(288)}) \in \mathfrak{R}^{288,1} \tag{34}$$
$$\Gamma_{LS} = \text{col}(\gamma_{LS}^{(1)}, \dots, \gamma_{LS}^{(288)}) \in \mathfrak{R}^{288,1} \tag{35}$$
$$\Gamma_{WS} = \text{col}(\gamma_{WS}^{(1)}, \dots, \gamma_{WS}^{(288)}) \in \mathfrak{R}^{288,1} \tag{36}$$

Multiple QR problem for interdependent model (14) can be estimated by solving an SOCP extending (32),

$$\{\{\alpha_i, \beta_i\}_{i=1}^m, \Gamma_{LW}, \Gamma_{LS}\} = \arg\min$$
$$\sum_{k=1}^m \sum_{i=1}^N h(l_i - Z_i\beta_k - \alpha_k - Z_{MH,i}(\Gamma_{LW} w_i + \Gamma_{LS} s_i); q_i)$$
$$+ \lambda \sum_{l=2}^m \|\beta_l - \beta_{l-1}\|^2 + \mu \sum_{j=2}^{m-1} \|\alpha_{j+1} - 2\alpha_j + \alpha_{j-1}\|^2$$
$$+ \nu_1 \|D_M^2 \Gamma_{LW}\|^2 + \nu_2 \|D_M^2 \Gamma_{LS}\|^2$$
$$+ \nu_3 \|D_H^2 \Gamma_{LW}\|^2 + \nu_4 \|D_H^2 \Gamma_{LS}\|^2, \tag{37}$$

where k is quantile index and i is the index of data point in data set (27). The last four terms in (37) are the regularizations for Γ_{LW} and Γ_{LS} dependencies on month and hour with D_M^2 being a circulant second difference operator for dependence on month and D_H^2 for dependence on hour. The regularization parameters λ and μ in (37) are selected similar to (32); parameters ν_1, ν_2, ν_3, ν_4 can be selected by cross-validation, see Section IV. The SOCP (37) was set up in CVX and solved using Gurobi. For 2 years of data, there are $N = 17,520$ data points and $m = 10$ quantiles, and $1,036$ decision variables; the solution takes about 4 minutes. Figure 2 shows 3-D plots of Γ_{LW}, Γ_{LS}, Γ_{WS} vs hour and month in Z_{MH} obtained by solving (37).

Multiple QR model for wind (15) dependent model can be estimated by solving the SOCP optimization problem similar to how (37) estimates model (14). The differences are that data w_i is used in place of l_i, vector Γ_{LS} is replaced by Γ_{WS}, and vector Γ_{LW} is absent from the problem. Estimation of model for solar (16) is described in Subsection III-B.

E. Tail Modeling

For the NERC 1-in-10 requirement, the probability of outage for each hour is of the order of 10^{-4}. This means quantiles of interest have just 2-3 hourly samples per year available for modeling. The distribution tails (for $0 < q \ll 1$ or $0 < 1-q \ll 1$), can be modeled using Extreme Value Theory (EVT), see [24]. Power system peak data usually follow Exponential or Generalized Pareto distributions predicted by the EVT [22], [30]. Estimating 2 or 3 parameters of these distributions allows extrapolating the tail into the quantiles where data are scarce.

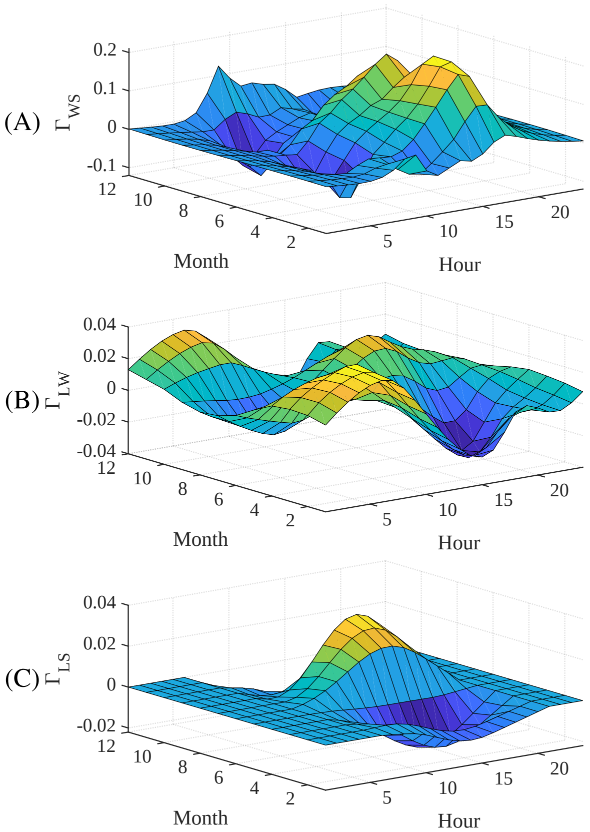

Fig. 2: (A) Solar Impact on Wind Coefficients Γ_{WS}
(B) Wind Impact on Load Coefficients Γ_{LW}
(C) Solar Impact on Load Coefficients Γ_{LS}

Quantile sampling used in Multiple QR model (37) covered quantile levels $q_1 = 0.05$ to $q_m = 0.95$ with pitch 0.1. For two years worth of hourly data, there are about 800 demand points with $q \geq 0.95$. The distribution tail was modeled using Peaks Over Threshold (POT) approach of EVT. For the right tail, solution of (37) was used to compute

$$e_{R,k} = y_{j_k} - Z_{j_k}\beta_m - \alpha_m, \tag{38}$$

where POT data indexes j_k are such that $e_{R,k} > 0$. Fitting EVT Pareto tail model to the POT demand data is described in [22], [24]. The tail model can be then sampled beyond quantile level q_m. Similar to (38), the left tail is modeled based on POT data

$$e_{L,k} = y_{j_k} - Z_{j_k}\beta_1 - \alpha_1, \tag{39}$$

where indexes j_k are such that $e_{L,k} < 0$. Figure 3 illustrates the QQ plots for the tail, QQ is a method to visualize the fitness of empirical distribution and theoretical distribution. QQ plot, know as quantile versus quantile plot, is the graph of of the quantiles of empirical CDF F_E versus corresponding quantiles of a hypothesized CDF F_0 [31]. The QQ plot is $[F_E^{-1}(q), F_0^{-1}(q)]$ for $0 < q < 1$. In Figure 3, the left and right blocks are using QQ plots to test the goodness-of-fit for left and right tails of the distribution of demand, respectively.

Distributions for solar and wind generation are bounded between zero and nameplate capacity. In these cases, Multiple QR formulation (37) includes quantile level samples at $q = 0$ and $q = 1$.

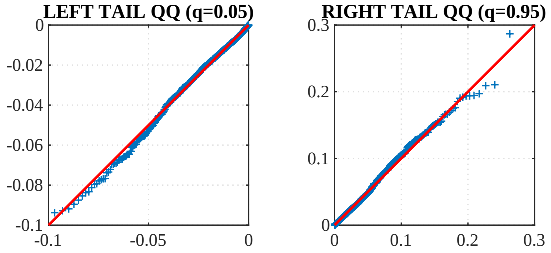

Fig. 3: QQ Plots on the Tail of Demand

IV. Validation

Statistical tests were used to quantify the fitness of the estimated conditional model (14)-(16) and to help choosing the 10 hyper-parameters of the Machine Learning estimator (37) (regularization parameters).

A. Pearson's Chi-squared Test

Pearson's Chi-squared Test [32] is used to test whether a set of data is coming from a certain distribution. The test divides probability space into k bins with probabilities $\{p_i\}_{i=1}^k$. The χ^2 statistic is computed from bin sample counts n_i as

$$\chi^2 = N \sum_{i=1}^{n} \frac{n_i/N - p_i}{p_i}, \tag{40}$$

For large enough n_i, statistics (40) follows a chi-square distribution with $k-1$ degrees of freedom (DOF). The hypothesis that data comes from the assumed distribution is rejected if

$$\chi^2 > \chi^2_{1-\alpha,k-1}. \tag{41}$$

Load, wind, and solar data from 2015 to 2016 were used to train the models, and 2017 data to test them. For wind distribution, data from 2015 to 2016 were used for the training and 2017 data for the testing. For the solar, five years of data from 2012 to 2016 were used for training and 2017 data for the testing. To validate model predictive power for the net injection $W + S - L$ (see, (24)), data from 2015 to 2016 was used to train, and data in 2017 to test the model. The scales for load, wind, and solar were taken from Scenario F in Table II.

The estimated distribution models were validated using Pearson's Chi-square test with 11 bins defined by quantile boundaries $q_1 = 0.05$, $q_2 = 0.15$, ..., $q_{10} = 0.95$. The bin sample counts are

$$n_i = \sum_{j=1}^{N} I(u_j \in (Q_{i-1}, Q_i)|Z_j), \quad (i = 1, 2, ..., 11), \tag{42}$$

where $I(\cdot)$ is the indicator function, and Q_i is the quantile for q_i; it is assumed that $Q(i = 0) = -\infty$ and $Q(i = 11) = +\infty$. The χ^2 statistics are calculated according to (40).

The linearly dependent model in Subsection II-C involves three variables: demand L, wind W, and solar S. Table I summarizes the results of the goodness-of-fit test for estimated models of L, W, S, and convolution model $W + S - L$. All chi-square statistics are below 95% level for both the train set and test set, which means all trained models fit the test data well. The model for $W + S - L$ was based on Scenario F, where renewables support over 59% of demand.

Variable	*DOF*	95% *Level*	*Pearson's* χ^2 *Statistics*	
			Training Set	***Test Set***
L(Load)	10	18.31	0.7569	17.9580
W(Wind)	10	18.31	16.8932	9.8689
S(Solar)	10	18.31	0.4973	4.0557
$W + S - L$	10	18.31	9.4982	9.7098

TABLE I: Grid Planning Scenarios

B. Comparison with Monte Carlo Analysis

This subsection compares the proposed analytical method with a Monte Carlo approach. The examples below illustrate two fundamental issues with the Monte Carlo. First, Monte Carlo tools for power grid reliability analysis usually simulate limited number of samples. As a result, the tails of the estimated distributions are not accurately characterized. Second, the tools used in practice usually ignore dependencies between the underlying distributions to reduce the effort required for setting up the analysis. We show that the results then could be substantially biased for large penetration of renewables.

As an example, reserve margin distribution was computed through the mix of load, wind, solar and generation capacity for $R = C - L + S + W$. The example follows Scenario F in Table II with 59% of renewables. Capacity C in the example is a constant parameter (generator outage is not modeled). The analytical method for computing distribution of R described in Section III-D estimated interdependent models for L, S, W from the data. The PMFs for $R = C + W + S - L|Z_j$, were computed by convolutions. The Monte Carlo analysis of R was based on two years of historical data. For the peak hour/week (no holiday), the two years of data contain 9 samples. Assuming that variables L, S, and W are independent, $9 \times 9 \times 9 = 729$ Monte Carlo samples for R combine the raw data samples for L, W, and S to estimate the empirical CDF. Figure 4 shows the CDFs for R computed by the two methods for Z_j corresponding to the peak hour/week (for August, Tuesday, 5pm, no holiday) and $C = 25.2$GW.

The convolved CDF was validated to be accurate, see Subsection IV-A. The Monte Carlo estimation deviates from this CDF. The difference in reserve margin at the 10^{-4} level can be as high as 5GW, over 15% of the total capacity. The CDFs, such as shown in Figure 4, were used to compute LOLP and then LOLH as described in Section II-D. Figure 5 illustrates the computed LOLH vs fixed generation capacity C for the two methods. Monte Carlo underestimates capacity C required to achieve LOLH = 2 by roughly 2GW, which is about 6%. This result illustrates the importance of the accurate estimation of interdependent distributions for demand, solar, and wind generation achieved by the proposed method.

V. Risk Estimation Examples

This section provides examples where the tools presented in Section II and Section III aree applied to ISO-NE service area data. We consider scenarios with solar and wind capacities higher than current ISO-NE values. We also consider large battery storage added to the actual pumped storage (2GW of power, 12GW-hours of energy).

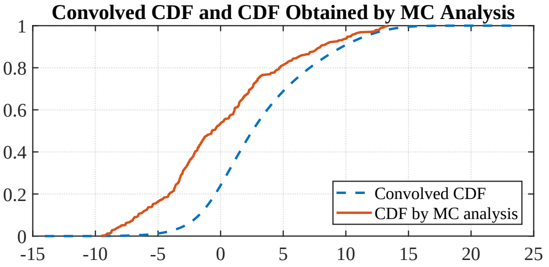

Fig. 4: Reserve CDFs: Monte Carlo and Analytical Method

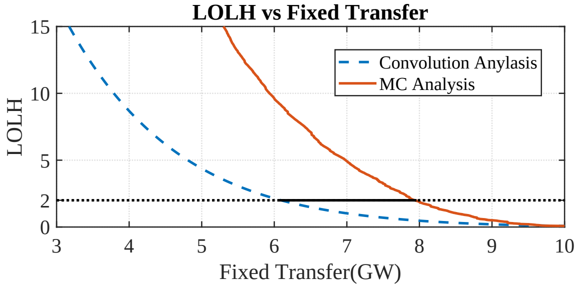

Fig. 5: LOLH for Monte Carlo and Analytical Method

A. Power Balancing

Reserve margin R is a random variable computed as

$$R = F + C - O - L + W + S, \tag{43}$$

where the first two terms in the r.h.s. are constants

F is the fixed transfer capacity of 1.7GW,
C is the dispatchable generation capacity of 36.4GW.

The last four terms in (43) are random variables with distributions estimated from historical data as described below.

O is the outage capacity,
L is the load (power demand),
W is the wind generation,
S is the solar generation.

Outaged Generation: Capacity and outage data for ISO-NE generators are stored in NERCs GADS database, same as for other ISOs. The total dispatchable capacity C is provided by 306 thermal generating units. GADS includes unit capacity levels h and outage probabilities q given as Equivalent Forced Outage Rate - demand (EFORd) values. Sampled Bernoulli distribution models were obtained from h, q and sampling step $\Delta h = 1$MW. Outage PMF is a convolution of the sampled Bernoulli distributions for the individual outages, see Subsection II-A. Figure 6 shows the PMF of outaged generation for the ISO-NE data.

Load, Wind, and Solar: The hourly load, wind and solar data for ISO-NE service area are described in Subsection III-A. The probabilistic models for demand, solar, and wind generation were estimated from data as described in Section III.

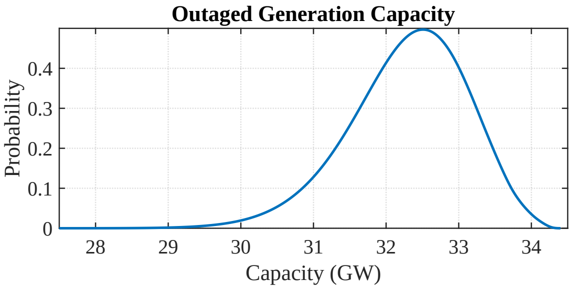

Fig. 6: Probability Distribution of Outaged Generation

To compute the distribution for $W + S - L$ in (43), the estimated distributions were convolved as described in Subsection II-C.

Computing LOLH: The LOLH risk index calculation described in Subsection II-D requires conditional distribution of reserve margin, $R|Z$ in (25) (26). Distribution of reserve margin R in (43) is modeled as conditional on calendar variables: Month, Weekday, Hour of a Day, and Holiday, see Subsection II-B. There are a total of $4,032$ different time regressor states Z_j. For each of these states, conditional distribution for $(S + W - L)|Z_j$ is computed as described in Subsection V-A. In (43), F and C are constants; random variable O is independent of Z_j and of L, W, and S. For each Z_j, the distribution for $R|Z_j = (F + C - O + S + W - L)|Z_j$ in (43), can be computed by convolving PMFs of independent variables $(F+C)$, $-O$, and $(S+W-L)|Z_j$, see Subsection II-B.

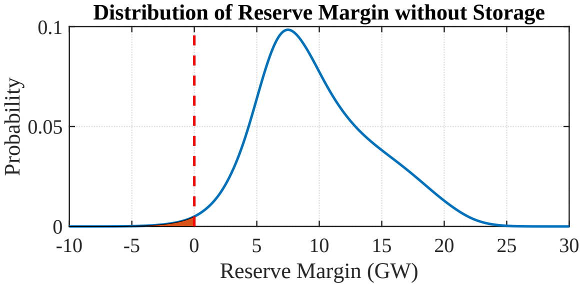

Fig. 7: Distribution of Reserve Margin at a Peak Hour

Given the conditional probability distributions of reserve margin $R|Z_j$ for all the regressor states, the LOLH risk index can be calculated as described in Subsection II-D. As an illustration, Figure 7 shows the reserve margin distribution of $R|Z_j$, where Z_j corresponds to a peak hour of the peak day. The shaded area under the curve corresponds to LOLP(Z_j) in (25).

B. Energy Balancing with Storage

The contribution of storage is analyzed through *energy margin* introduced in [12]. The battery can only stave off loss of load if the energy margin, the sum of reserve margins in n consecutive hours combined and battery capacity B_n for energy injection over these n hours, does not fall below zero. The energy margin can be computed as

$$R_n = F_n + C_n - O_n - L_n + W_n + S_n, \tag{44}$$

where the first two terms in the r.h.s. are constants

F_n is the energy transfer capacity over the n hours, $F_n = nF$,
C_n is the n-hours energy generation capacity, $C_n = nC$,

The remaining terms in (44) are random variables that can be estimated from the data as described below,

O_n is the total outage energy over the n hours,
W_n is the wind-generated energy over the n hours,
L_n is the total energy demand over the n hours,
S_n is the solar-generated energy over the n hours.

Outaged Energy: The outaged energy over n consecutive hours from $t_k - n + 1$ to t_k can be defined as

$$O_n(t_k) = O(t_k - n + 1) + \ldots + O(t_k). \tag{45}$$

Each of n terms $O(t_k)$ in the r.h.s. of (45) is assumed to be an independent random variable following the outage distribution computed as discussed in Subsection II-A and illustrated in Figure 6. The PMF of resulting distribution is given by convolution of the PMFs for the n summands, which is n-fold convolution of Figure 6 distribution for O with itself.

Net Energy for Load, Wind, and Solar: Variables L_n, W_n, and S_n in (44) at any hour t, can be expressed similar to (45). Data-driven models for these variables can be built from historical data. As an example, the data for the random variable S_n is obtained by taking a time series of hourly samples for $S(t)$ and running it through a convolution filter with rectangular window $[1, 1, ..., 1]$ of n ones. The convolution output samples can be used to build a probability distribution model for S_n as described in Subsection III-B. The same approach is used for computing W_n and L_n as rectangular window convolutions of wind and load data, respectively.

The estimation of interdependent probability distributions for L_n, W_n, and S_n from the data is similar to that for L, W, S in Subsection V-A. To compute the distribution for $W_n + S_n - L_n$ in (45), the distributions are combined as described in Subsection II-C. The r.h.s. of (44) is the sum of the constant $F_n + C_n$ and two conditionally independent random variables: O_n and $W_n + S_n - L_n$. The distribution for R_n is computed by convolutions, similar to R in Subsection V-A.

C. LOLH Bounds for Energy Balancing

For a given window width n and time regressor Z_j, consider a loss of load condition

$$\text{LOLP}_{n,j} = \mathbf{P}(R_n \leq -B_n - H_n|Z_j), \tag{46}$$

where B_n is energy provided by battery storage within n hours and H_n by hydro storage. The battery energy is limited by

$$B_n = \min(nB_M, B_C),$$

where B_M is max power and B_C is energy capacity of the battery. Similar expression holds for hydro storage energy H_n.

In expression (46), $R_n + B_n + H_n < 0$ is a condition of load loss assuming that full energy capacity of storage system is available any time it is needed, the storage is fully charged. This expression yields the lower bound of actual LOLP; actual LOLP can be only higher. Using (46), one can calculate the lower bound LOLH$_n$ as described in Subsection II-D.

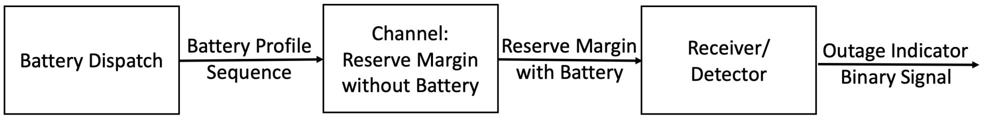
Battery Dispatch
Battery Profile Sequence
Channel: Reserve Margin without Battery
Reserve Margin with Battery
Receiver/ Detector
Outage Indicator Binary Signal

Fig. 8: Information Theory View of Storage Impact on Grid Reliability

Variable	*Scenario A*	*Scenario B*	*Scenario C*	*Scenario D*	*Scenario E*	*Scenario F*
Dispatchable Capacity	34.4GW	30.4GW	30.4GW	30.4GW	30.4GW	25.2GW
Wind Nameplate Capacity	0.95GW	6.41GW	9.79GW	6.41GW	6.41GW	23.6GW
Solar Nameplate Capacity	0.57GW	12.57GW	12.57GW	18.57GW	12.57GW	23.07GW
Battery Storage (5-hour)	-	1.5GW	1.5GW	1.5GW	2.7GW	5GW
Pumped Storage (8-hour)	1.5GW	1.5GW	1.5GW	1.5GW	1.5GW	1.5GW
Fixed Transfer	1.7GW	1.7GW	0.7GW	0.7GW	0.7GW	4GW
90/10 Load	24.94GW	24.94GW	24.94GW	24.94GW	24.94GW	24.94GW
LOLH Lower Bound	0.3328	0.5317	0.8337	1.0482	0.4427	0.1944
LOLH Upper Bound	2.37	2.23	2.19	2.16	2.24	2.21

TABLE II: Grid Planning Scenarios

More accurate lower bound is obtained by considering all possible energy windows with different lengths n. Assume that storage charges during the night and discharges during the day. If the storage runs out of energy, this means condition $R_n + B_n + H_n < 0$ holds for one of the energy windows during the day. Thus, the lower bound for LOLH is

$$\text{LOLH} = \max_n \{\text{LOLH}_n\}. \tag{47}$$

Additional discussion can be found in [12].

An upper bound of the LOLP can be obtained by assuming a fixed dispatch of the battery. With this assumption, storage generation just adds a deterministic term in power balance analysis of Subsection V-A. We assume storage dispatch is the same every day within a month. On each day, energy stored equals energy discharged and does not exceed storage capacity.

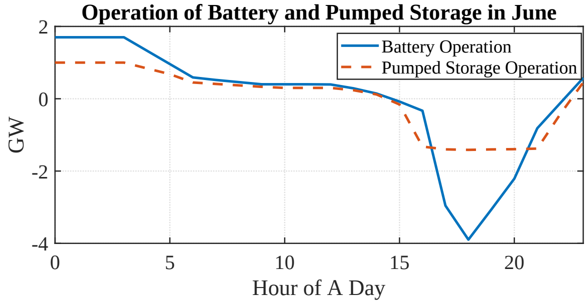

Fig. 9: Storage Power Dispatch

Figure 9 has an example of battery and pumped storage dispatch for Scenario F in Table II during a day in June. Storage charging is shown as positive value, discharging as negative. Scenario F has 59% renewables with large solar generation; storage is discharging in the evening and night.

Information Theory View: The presented energy balance analysis is related to Information Theory analysis of digital communications. Figure 8 depicts LOLP analysis problem as communication system. The input is B, the storage hourly operation profile. This signal is transmitted through a noisy channel $(R + B)|Z$, where R is given by (43), and produces an hourly reserve margin waveform. The Receiver/Detector determines Loss of Load events as $R + B < 0$. The decoded output is 1 if there is Loss of Load, otherwise 0 is decoded.

Random variable R represents channel noise. Code B (storage dispatch) is meant to be decoded as 0, no Loss of Load. LOLP is the probability of decoding 1 instead, an error. Performance of the code is determined by the probability of error, which is computed the LOLP upper bound in Section V-B. Capacity for error-free transmission in a noisy communication channel is described by Shannon's Limit, which has the meaning similar to the lower bound for the expected error probability computed in Section V-B.

There are two fundamental differences between the analysis in this paper and the standard usual Information Theory analysis. First, Shannon's analysis assumes Gaussian noise in the channel, while distributions of reserve margin in this paper are very much non-Gaussian. Second, Shannon's analysis assumes memoryless channel, while in this paper the probability distributions are dependent on time regressor Z and strongly vary through the day. Thus, despite the described analogy, Shannon's analysis is not directly applicable and more complicated analysis presented in this paper is necessary.

VI. Planning Grid with Renewables and Storage

This subsection provides LOLH analysis result for six scenarios in Table II. In each scenario, the upper bound of LOLH is around 2.2 in compliance with the 1-in-10 requirement. Scenario A is close to actual 2017 ISO-NE system, but the load is factor 1.4 higher. In Scenario B, wind and solar account for 10% of annual load and battery storage is added. Scenarios C, D, and E, replace 1GW of fixed transfer capacity in Scenario B by wind, solar, and battery storage, respectively. The results show capacity factor of 29.6% for wind, 16.67% for solar, and 83.3% for battery.

Scenario F is extreme case where solar and wind account for 58.7% of the total annual load. It assumes that all dispatchable generators over 525MW are retired. Battery capacity is raised to 4GW, and fixed transfer capacity is raised to 5GW. For most scenarios, the ratio of upper and lower bound of LOLH is less than 2. The ratio is larger for Scenarios C and F because of the large wind capacity. The wind is rather unpredictable and cannot be addressed by using the same battery profile everyday.

VII. Conclusion

This paper demonstrated efficient, practical tools for probabilistic reliability analysis of grid with very high penetration of renewables and storage. Machine learning models are built from historical data and accurately describe probability distributions for interdependent variables (solar, wind, load) including extreme (tail) events. The tools combine the models to compute probabilities, LOLP, and LOLH for a planning scenario with given variable generation and storage capacities. The accuracy of predicting the probabilities based on the previous year training data is statistically validated.

The demonstrated tools are substantially more accurate than Monte Carlo methods currently used in practice. They are also much faster. Complete risk analysis for a given scenario might take just a few seconds. This allows to analyze many scenarios, e.g., for resource mix optimization.

The developed analysis of a grid with storage provides a lower bound for the LOLH risk. Similar to Shannon's limit in Information Theory, this lower bound allows to evaluate room for improving a given battery schedule. A specific schedule provides an upper bound of the LOLH.

Probabilistic ramp rate analysis using developed tools will be demonstrated in a follow-on paper.

References

[1] R. N. Allan *et al.*, *Reliability evaluation of power systems*. Springer Science & Business Media, 2013.

[2] M. Shahidehpour, Z. Li, J. Wang, and C. Chen, "Applying DER-CAM for IIT microgrid explansion planning," ANL, Tech. Rep. ESD-16/6-127680, 2016.

[3] "Servm software." [Online]. Available: https://www.astrape.com/servm/

[4] B. A. Frew, W. J. Cole, Y. Sun, T. T. Mai, and J. Richards, "8760-based method for representing variable generation capacity value in capacity expansion models," NREL, Tech. Rep. PR-6A20-68870, 2017.

[5] M. A. Matos and R. J. Bessa, "Setting the operating reserve using probabilistic wind power forecasts," *IEEE Trans. on Power Systems*, vol. 26, no. 2, pp. 594–603, 2011.

[6] L. Hirth and I. Ziegenhagen, "Balancing power and variable renewables: Three links," *Renewable Sustain. Energy Rev.*, vol. 50, pp. 1035–1051, 2015.

[7] M. Deshmukh and S. Deshmukh, "Modeling of hybrid renewable energy systems," *Renewable and Sustain. Energy Rev.*, vol. 12, pp. 235–249, 2008.

[8] Y.-Y. Hong and R.-C. Lian, "Optimal sizing of hybrid wind/pv/diesel generation in a stand-alone power system using markov-based genetic algorithm," *IEEE Trans. Power Del.*, vol. 27, no. 2, pp. 640–647, 2012.

[9] N. Y. Krakauer and D. S. Cohan, "Interannual variability and seasonal predictability of wind and solar resources," *Resources*, vol. 6, no. 3, p. 29, 2017.

[10] N. Zhang, C. Kang, C. Singh, and Q. Xia, "Copula based dependent discrete convolution for power system uncertainty analysis," *IEEE Trans. Power Systems*, vol. 31, no. 6, pp. 5204–5205, 2016.

[11] R. Sioshansi, S. H. Madaeni, and P. Denholm, "A dynamic programming approach to estimate the capacity value of energy storage," *IEEE Trans. on Power Systems*, vol. 29, no. 1, pp. 395–403, 2014.

[12] W. Gao and D. Gorinevsky, "Probabilistic balancing of grid with renewables and storage," in *IEEE Int. Conf. on Probabilistic Methods Applied to Power Systems*, Boise, ID, June 2018.

[13] R. Koenker and G. Bassett Jr, "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.

[14] H. D. Bondell, B. J. Reich, and H. Wang, "Noncrossing quantile regression curve estimation," *Biometrika*, vol. 97, pp. 825–838, 2010.

[15] V. Chernozhukov, I. Fernández-Val, and A. Galichon, "Quantile and probability curves without crossing," *Econometrica*, vol. 78, no. 3, pp. 1093–1125, 2010.

[16] S. Shenoy and D. Gorinevsky, "Estimating long tail models for risk trends," *IEEE Signal Proc. Letters*, vol. 22, no. 7, pp. 968–972, 2015.

[17] K. Wang and H. J. Wang, "Optimally combined estimation for tail quantile regression," *Statistica Sinica*, vol. 26, pp. 295–311, 2016.

[18] J.-P. Chavas, "On multivariate quantile regression analysis," *Statistical Methods & Applications*, vol. 27, no. 3, pp. 365–384, 2018.

[19] C. Sigauke, M. Nemukula, and D. Maposa, "Probabilistic hourly load forecasting using additive quantile regression models," *Energies*, vol. 11, no. 9, p. 2208, 2018.

[20] R. Juban, H. Ohlsson, M. Maasoumy, L. Poirier, and J. Z. Kolter, "A multiple quantile regression approach to the wind, solar, and price tracks of gefcom2014," *Int. J. Forecast.*, vol. 32, no. 3, pp. 1094–1102, 2016.

[21] W. Gao, D. Tayal, and D. Gorinevsky, "Probabilistic planning of minigrid with renewables and storage in western australia," in *IEEE PES GM*. IEEE, 2019, pp. 1–5.

[22] S. Shenoy and D. Gorinevsky, "Data-driven stochastic pricing and application to electricity market," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 6, pp. 1029–1039, 2016.

[23] ——, "Stochastic optimization of power market forecast using non-parametric regression models," in *IEEE PES GM*, Denver, CO, 2015.

[24] ——, "Risk adjusted forecasting of electric power load," in *Amer. Control Conf.*, Portland, OR, June 2014, pp. 914–919.

[25] NERC, *Methods to Model and Calculate Capacity Contributions of Variable Generation for Resource Adequacy Planning*, 2011.

[26] "Energy, load, and demand reports." [Online]. Available: https://www.iso-ne.com/isoexpress/web/reports/load-and-demand/-/tree/zone-info

[27] "Operations reports." [Online]. Available: https://www.iso-ne.com/isoexpress/web/reports/operations/-/tree/daily-gen-fuel-type

[28] "Renewables.ninja." [Online]. Available: https://www.renewables.ninja/

[29] S. Shenoy, D. Gorinevsky, and S. Boyd, "Non-parametric regression modeling for stochastic optimization of power grid load forecast," in *Amer. Control Conf.*, Chicago, IL, July 2015, pp. 1010–1015.

[30] L. De Haan and A. Ferreira, *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

[31] J. D. Gibbons and S. Chakraborti, *Nonparametric statistical inference*. Springer, 2011.

[32] G. W. Snedecor and W. G. Cochran, *Statistical methods*. Ames, IA: Iowa State Univ. Press, 1989.

Weixuan Gao is a Ph.D. candidate in Civil and Environmental Engineering at Stanford University. He received the B.E. degree in Statistics from Wuhan University, Wuhan, China, and the M.S. degree in Statistics from Washington University in St. Louis. His research interests include big data for extreme event, power system planning, machine learning and deep learning.

Dimitry Gorinevsky (M'91-SM'98-F'06) received a Ph.D. from Moscow (Lomonosov) University, and a M.Sc. from Moscow Institute of Physics and Technology (Phystech). He has been a Consulting Professor in Electrical Engineering with Information Systems Laboratory at Stanford University since 2003. His interests are in Industrial AI - analytical applications for Industrial Internet of Things (IIoT). He is a founder of Mitek Analytics, an Industrial AI company in Palo Alto, CA. Over the last two decades, he has been working on data analytics applications in aerospace systems, energy, computing, and other industries. He has authored a book, 180+ papers, and many patents. He received Control Systems Technology Award, 2002, and Transactions on Control Systems Technology Outstanding Paper Award, 2004, of the IEEE Control Systems Society. He received Best Paper Award (Senior Award), 2013, of the IEEE Signal Processing Society. He is a Fellow of IEEE.