A Data-Driven Load Fluctuation Model for Multi-Region Power Systems

Zhen Dai, *Student Member, IEEE*, and Joseph Euzebe Tate, *Member, IEEE*

***Abstract*—In this paper, we propose a data-driven load fluctuation model, based on high-resolution historical demand data from multi-regional systems, that can be used for research such as power system generation control studies and probabilistic load flow studies. As in previous studies, the random load fluctuations are modeled as independent Gaussian random variables; however, unlike in previous studies, we do not assume the relationship between the standard deviation and the base demand in each region is known *a priori*. Instead, we propose a framework for determining the relationship between the base demand level and short-term demand uncertainty. The developed framework has been tested using actual 5-minute demand data from the New York and New Zealand power systems. The results demonstrate that the proposed models outperform those used in previous work. Coefficients of the example cases are included, the parameters of which can be applied to similar multi-region systems.**

***Index Terms*—Load modeling, Gaussian distribution, power system simulation.**

I. Introduction

LOAD fluctuation is determined by many factors: weather conditions, time (e.g. daily and seasonal peaks), types of load (e.g. residential, industrial, agricultural), electricity prices, etc. [1]. Modeling load variation is not only needed for economic dispatch and planning [2], it is also necessary for automatic generation control studies [3]–[5], stability analyses [6], probabilistic load flow [7], [8], dynamic state estimation [9] and parameter estimation (e.g., evaluating Thevenin equivalents [10]), since the mismatch between load and generation presents challenges to the robustness and accuracy of control and estimation methods. For many of these studies, particularly those based on publicly-available test systems, it is unlikely that the load model with required economical and environmental factors is available.

Probabilistic models, in particular the Gaussian distribution, are often used to represent load uncertainty [5], [11], [12]. These methods are simple to implement, especially when repeated simulation is needed, since they do not depend on comprehensive knowledge of the test systems and only require one parameter (the standard deviation σ of the load variation) to be specified for each region. Although some formulas relating σ to the base demand level in each region have been proposed in prior work (e.g., σ is proportional to the demand [2], [7], [8], [12] or to the square root of system capacity [5]), no evidence is provided to indicate these formulas result in useful models of load fluctuations.

The main alternative model that has been proposed is to utilize stochastic differential equations (SDEs), a concept first introduced in [13]. The main advantages of the SDE approach are that it allows for explicit modeling of mean reversion and it can fit the demand behavior using a larger set of potential random processes. For example, in [14], the best performance is obtained by fitting the historical data to a Normal-Inverse Gaussian process combined with a Poisson process. While potentially improving accuracy, SDE-based methods require large amounts of historical data at each load bus to determine the appropriate bus-specific random process model and mean reversion rate. For most study systems, particularly those used in public research, this detailed historical data is not available. Another drawback of using SDEs is that there is no obvious way to scale the mean reversion rate and diffusion parameters when there are substantial changes in the demand level, as is often the case in planning studies. Therefore, our research focuses on the simpler, Gaussian model of load fluctuation.

This paper proposes a framework to determine the probabilistic model parameters of load fluctuation. A polynomial model is developed to approximate the standard deviation based on the base power demand within each region. In Section II, we first review the adopted load fluctuation model. Then in Section III, the polynomial model to approximate the standard deviation of load variation is presented, along with the framework for determining and evaluating potential models. We demonstrate the framework in Section IV utilizing actual load data from the New York and New Zealand systems. Model parameters of the test cases are presented, followed by a discussion of model selection including a comparison of the proposed model and models proposed in previous studies. Concluding remarks and potential avenues for future work are provided in Section V.

II. Load Fluctuation Model

Load fluctuation happens constantly in power systems, which can be represented by probabilistic models. One common choice is to assume that the load fluctuations follow a Gaussian distribution [3], [5], [6], [11]. In this paper, we focus on change in load

Manuscript received April 30, 2018; revised September 25, 2018; accepted November 11, 2018. Date of publication November 21, 2018; date of current version April 17, 2019. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Discovery Grant NSERC RGPIN-2016-06674. Paper no. TPWRS-00656-2018. *(Corresponding author: Zhen Dai.)*

The authors are with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: zhen.dai@mail.utoronto.ca; zeb.tate@utoronto.ca).

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

Digital Object Identifier 10.1109/TPWRS.2018.2882560

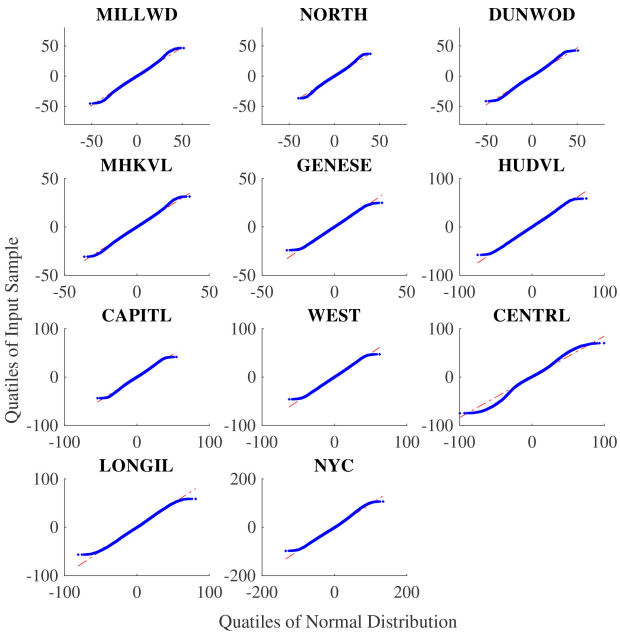


Fig. 1. Q-Q plots of load deviations in June, July, and August of 2012 for the 11 zones of NYCA (blue) vs. the best Gaussian fit (red).

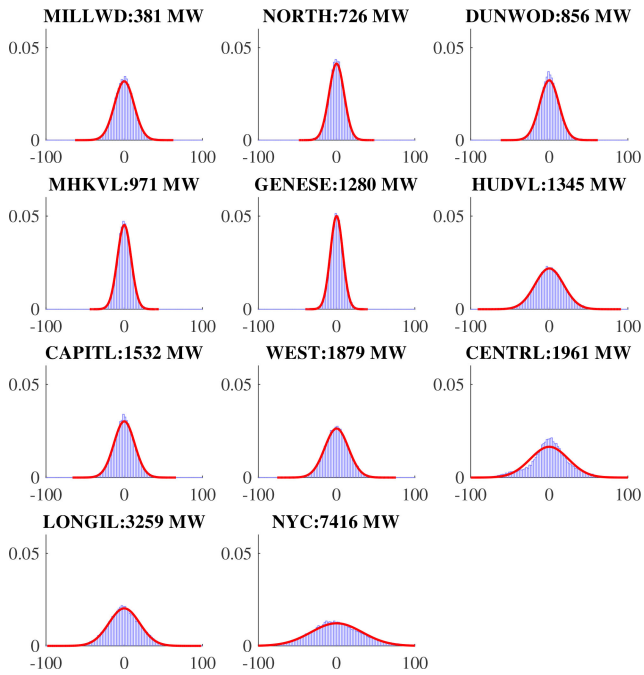


Fig. 2. Histogram of NYCA load deviations (blue) vs. Gaussian distribution with parameters given by (1) and (2) (red). The plots are ordered from the smallest mean load (top-left) to largest mean load (bottom-right).

values within a short time interval (e.g., 5 minutes). When given active power load measurements P at discrete time k and $k+1$, the difference can be computed as $\Delta P_k = P(k+1) - P(k)$.

For a normal distribution, two parameters need to be determined: mean and standard deviation. At each discrete moment, we consider ΔP_k as a random variable that has mean $\Delta \bar{P}_k$ and standard deviation σ_k. Similar to [5] and [6], we assume ΔP_k is independent of ΔP_l, where $l < k$ for all l and k. Another assumption is the standard deviation is the same for all k, which means the load changes are independent and identically distributed random variables.

Q-Q (quantile-quantile) plots were used to verify that the sample data follows a normal distribution (see [15] for details). If the normality assumption holds, then the Q-Q plot generated by the sample data and its corresponding theoretical distribution traces the same straight line. Fig. 1 shows an example of Q-Q plots for 11 zones of the New York control area (NYCA) based on the summer demands in 2012. To eliminate the impact of outliers, a total of 0.5% of the sample data were trimmed (0.25% maximum and 0.25% minimum values). For most regions, the Q-Q plots trace their respective reference lines, implying ΔP may follow a normal distribution. However, the deviations in these plots indicate the sample data have thinner tails on both sides than anticipated by a normal distribution. This characteristic indicates that assuming Gaussian behavior will lead to a slight overestimation of load fluctuations.

The sample mean and the standard deviation can be estimated given historical data as follows [16]:

$$\Delta \bar{P}_k = \frac{1}{N} \sum_{k=1}^{N} \Delta P_k \tag{1}$$

$$\tilde{\sigma}^2 = \frac{1}{N-1} \sum_{k=1}^{N} (\Delta P_k - \Delta \bar{P}_k)^2 \tag{2}$$

We noticed in our case (three months of data) that the sample mean of ΔP is around zero. Since most load change follows a daily cycle, zero mean can be observed given samples from complete cycles. One-sample student t-tests were then conducted to further test the hypothesis. The test fails to reject the null hypothesis (i.e., zero mean, unknown variance) with a typical p value around 0.7.

Fig. 2 shows a comparison between the histogram versus the probability density function (pdf) of the Gaussian distribution defined by applying (1) and (2) given the same data used in Fig. 1. The mean demands of each region are also provided above each subplot. The histogram and the pdf agree with each other indicating approximate normality for most cases with an exception of zone CENTRL. From the corresponding Q-Q plot in Fig. 1, the same conclusion can be drawn: the frequency that ΔP falls in the range -100 MW to -20 MW is lower than the normal distribution predicts. The fatter tail also implies significant load decrease is more likely to happen in this zone, which might be useful for load profiling.

Intuitively, larger demand in an area should correspond to a higher rate of load change, in terms of absolute quantities (e.g., MW). The trend is corroborated by Fig. 2, which shows that as mean demand increases, the corresponding distribution of ΔP tends to spread over a wider range, i.e., has a larger standard deviation.

III. Polynomial Model of σ

As discussed in the previous section, there is an observed correlation between load size and standard deviation of load variation. If there exists a simple function to represent σ using known load information (e.g., the peak or average load), then σ

of similar regions and systems can be estimated. To approximate σ, we propose a polynomial function in the base power ${P_b}^{n_e}$ with degree n_p:

$$\begin{aligned}\sigma &= \sum_{i=0}^{n_p} a_i {P_b}^{n_e \cdot i} \\ &= \begin{bmatrix} a_0 & a_1 & \cdots & a_{n_p} \end{bmatrix} \begin{bmatrix} 1 \\ P_b^{n_e} \\ \vdots \\ (P_b^{n_e})^{n_p} \end{bmatrix} \\ &= \underline{\alpha}^T \underline{\mathcal{P}}_b, \text{ with } \underline{\alpha},\ \underline{\mathcal{P}}_b \in \mathbb{R}^{n_p+1}. \end{aligned} \tag{3}$$

Three parameters—P_b, n_e and n_p—determine the set of candidate models under consideration. After a choice is made, the corresponding coefficient vector $\underline{\alpha}$ is computed through an estimation technique that minimizes the difference in the observed and estimated σ.

If the polynominal coefficients in (3) are the same for all regions, with base powers $P_{b,k}$ ($k = 1, 2 \ldots, n$ regions), $\underline{\alpha} = [a_0, \ldots, a_{n_p}]^T$ can be estimated using linear regression. Considering all n regions simultaneously, the vectorized form of (3) is:

$$\tilde{\underline{\sigma}} = \mathcal{P}\underline{\alpha}, \tag{4}$$

with

$$\mathcal{P} = \begin{bmatrix} \underline{\mathcal{P}}_{b,1}^T \\ \underline{\mathcal{P}}_{b,2}^T \\ \vdots \\ \underline{\mathcal{P}}_{b,n}^T \end{bmatrix} = \begin{bmatrix} 1 & P_{b,1}^{n_e} & \ldots & P_{b,1}^{n_e n_p} \\ 1 & P_{b,2}^{n_e} & \ldots & P_{b,2}^{n_e n_p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & P_{b,n}^{n_e} & \ldots & P_{b,n}^{n_e n_p} \end{bmatrix}, \ \tilde{\underline{\sigma}} = \begin{bmatrix} \tilde{\sigma_1} \\ \tilde{\sigma_2} \\ \vdots \\ \tilde{\sigma_n} \end{bmatrix}.$$

Among different estimation techniques, the ordinary least squares (OLS) estimator, which solves the minimization problem

$$\min_{\underline{\alpha}} \left\| \tilde{\underline{\sigma}} - \mathcal{P}\underline{\alpha} \right\|_2^2, \tag{5}$$

is commonly used for its computational efficiency and closed form solution. The estimated coefficients using the OLS method are given by

$$\hat{\underline{\alpha}} = \left(\mathcal{P}^T \mathcal{P}\right)^{-1} \mathcal{P}^T \tilde{\underline{\sigma}}. \tag{6}$$

We can apply the coefficients to other similar regions or the same region at a different time. Thus for any region k with known $P_{b,k}$, the estimated standard deviation of the load fluctuation is given by

$$\hat{\sigma}_k = \hat{\underline{\alpha}}^T \underline{\mathcal{P}}_{b,k}. \tag{7}$$

This estimate requires no prior knowledge of a region except the base power P_b. Fig. 3 shows the framework to estimate and then evaluate the polynomial model. First, coefficients are estimated for a model with a set of parameters P_b, n_e and n_p and a set of training data from all regions. Next, the model is evaluated using a validation data set based on percentage error

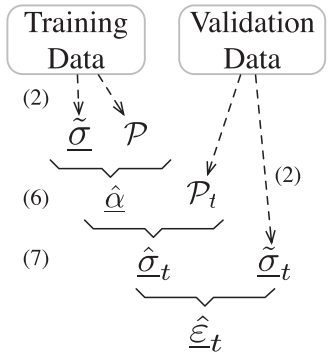


Fig. 3. Procedure for the estimation and verification of polynomial models.

in predicted σ with respect to the sample σ calculated using (2). The percentage error $\hat{\varepsilon}_{k,t}$ of region k and test set t is defined as:

$$\hat{\varepsilon}_{k,t} = \frac{|\hat{\sigma}_{k,t} - \tilde{\sigma}_{k,t}|}{\tilde{\sigma}_{k,t}}. \tag{8}$$

The modified mean of regions and test sets can be computed as follows:

$$\varepsilon_m = \frac{1}{n-2}\left(\sum_{k=1}^{n} \bar{\varepsilon}_k - \min_j \bar{\varepsilon}_j - \max_j \bar{\varepsilon}_j\right), \tag{9}$$

where

$$\bar{\varepsilon}_k = \frac{1}{n_t}\sum_{t=1}^{n_t} \tilde{\varepsilon}_{k,t}$$

is the mean percentage error over all test sets.

The procedure is repeated with different selections of P_b, n_e and n_p, and the best model type is determined based on ε_m. Inspired by [11], in which load is modeled as a Gaussian distribution with mean being either the peak load or the average load in a given period, we consider peak, average and median load as candidates for the base load quantity P_b. Further discussion of model selection is presented in the subsequent case study section, with accompanying experimental results.

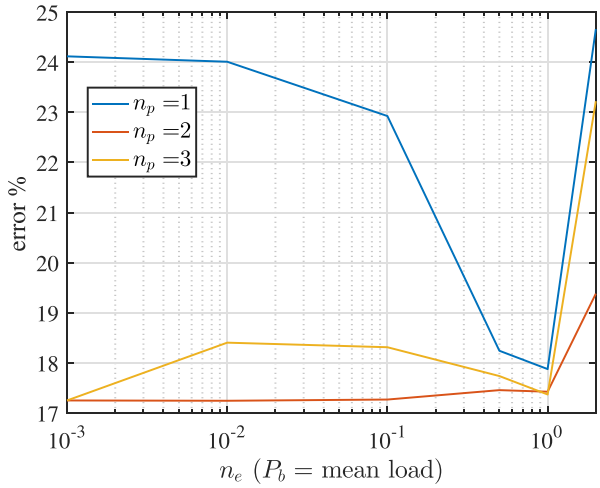
IV. Case Studies

Data from the NYISO (New York Independent System Operator) [17] and Transpower New Zealand [18] was used to verify the proposed estimation method. Both provide regional demand data with 5 minute intervals to the public. The experiments were conducted in MATLAB on a computer with Intel Xeon E5-1607 processor and 8 GB RAM. Steps (1) and (2) take 0.0036 s on average for one region while the parameter estimation (6) takes under 0.001 s for a given polynomial model.

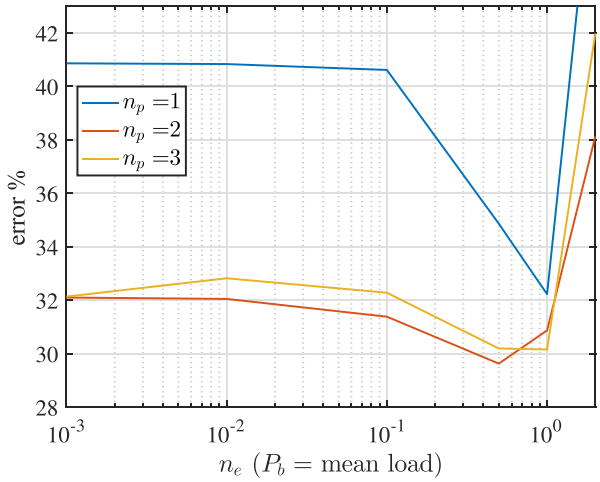
In this section, we first discuss the choice of different model parameters. Then, a detailed analysis of the experimental results is presented, including a comparison to legacy models that have appeared in prior work.

A. System Description

The NYISO provides load data for 11 zones dating back to 2001. The New York control area had a total coincident



(a) NYISO



(b) New Zealand

Fig. 4. Modified mean error of models with different n_p and n_e.

summer peak demand of 32076 MW and a winter peak demand of 24164 MW in 2016 [19], but the demands within the 11 zones differ significantly. The largest zone demand at a time can be 20 times that of the smallest zone, which makes the system a good data source for estimating and verifying the proposed polynomial model (3). The results were generated based on June-July-August load data from 2012 to 2017 with 2012 used as the training set.

The New Zealand system consists of two islands totaling 14 zones. In 2015, the total winter, shoulder and summer peak were 6088 MW, 5663 MW and 5092 MW respectively [20]. The demand also varies significantly from zone to zone. The maximum zonal demand (occurring in zone Southland) can be more than 10 times of the minimum (in zone West Coast) during the same time period. The system operator Transpower provides regional load data for the past month to the public. The experimental results were based on load data from mid November 2017 to mid January 2018. The data is equally divided with the first half as the training set and the remaining as the validation set.

B. Polynomial Model Order and Base Power Selection

As mentioned in the previous section, three parameters—P_b, n_e and n_p—are needed for the polynomial model. In order to determine the best candidate models, experiments were conducted with different parameters. Fig. 4 shows the semi-logarithmic plots of the modified mean error ε_m versus n_e with $n_p = 1, 2, 3$ and P_b = mean load. The modified mean ε_m, as defined in (9), was computed by discarding the best and the worst region but with all test years considered.

For the NYISO, Fig. 4(a) demonstrates that given a fixed n_e, increasing n_p does not always result in better accuracy. In fact, the best performance is achieved when $n_p = 2$ regardless of n_e, which indicates a higher degree polynomial model is unnecessary. Regarding n_e, we note that $n_e > 1$ consistently leads to a significant increase in error. For $n_e < 1$, we observe that the model becomes highly dependent on the training data. In particular, exceeding the training range (in terms of P_b) by a small amount can lead to rapid change in σ. For example, when $n_e = 0.1$ and optimal coefficients are used, if P_b decreases below the minimum P_b in the training set, the estimated σ will increase rather than decrease, i.e., the polynomial function in this case is not monotonic. Thus these models are not well-suited to studying systems with demand levels lower than those seen in the training data. Additionally, numerical instability occurs when n_p is large and/or n_e is small due to the large differences in demand between regions (resulting in ill-conditioning when computing the optimal coefficients (6)). Based on these experimental results, only n_e values in the range of 0.5 to 1 were considered. The analogous semi-logarithmic plot of the New Zealand system in Fig. 4(b) shows a similar trend except that $n_p = 2$ and $n_e = 0.5$ yields the global minimum error.

TABLE I
POLYNOMIAL COEFFICIENTS OF NY SYSTEM

P_b	n_p	n_e	a_0	a_1	a_2	ε_m
peak	1	0.5	1.29228	0.28504	0	18.58%
	1	1	9.32242	0.00213	0	18.15%
	2	0.5	5.59833	0.13005	0.00118	17.46%
	2	1	8.09106	0.00291	-6.55E-8	17.62%
mean	1	0.5	1.42252	0.35341	0	18.24%
	1	1	9.49018	0.00324	0	17.88%
	2	**0.5**	**5.44130**	**0.17459**	**0.001673**	**17.45%**
	2	**1**	**7.68245**	**0.00499**	**-2.20E-7**	**17.42%**
median	1	0.5	1.39892	0.35685	0	18.46%
	1	1	9.51206	0.00328	0	17.98%
	2	0.5	5.28067	0.18330	0.00163	17.59%
	2	1	7.53612	0.00521	-2.44E-7	17.53%

Data source: 2012 to 2017 summers (June, July, August) from the NYISO. Modified mean ε_m was calculated using (9). Bold is used to highlight the best combinations, in terms of minimal ε_m .

Based on Fig. 4 and the first-order models in prior work, we now focus on the performance when $n_p = 1$ or 2 and $n_e = 0.5$ or 1. Table I shows the percentage error obtained with different models in the last column and the corresponding fitted coefficients of the New York system. Based on the results, we make the following observations:

1) Using mean load yields the lowest error rate with other parameters fixed. Nevertheless, the choice of P_b does not

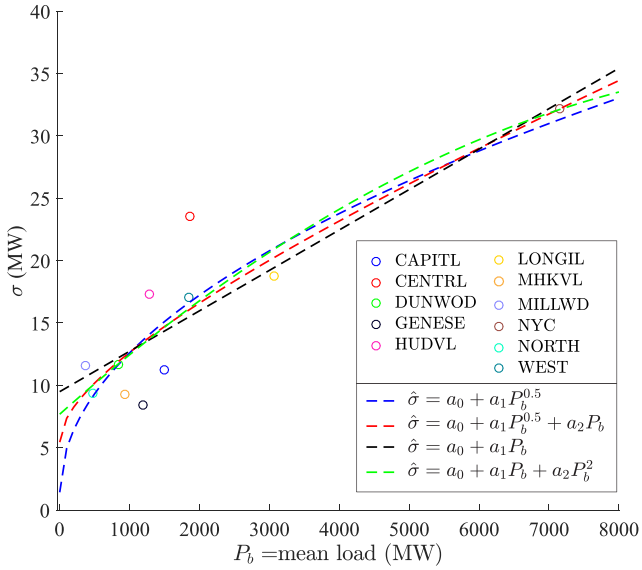


Fig. 5. Comparison of different models in Table I with P_b = mean load vs. median sample $\tilde{\sigma}_m$ (○).

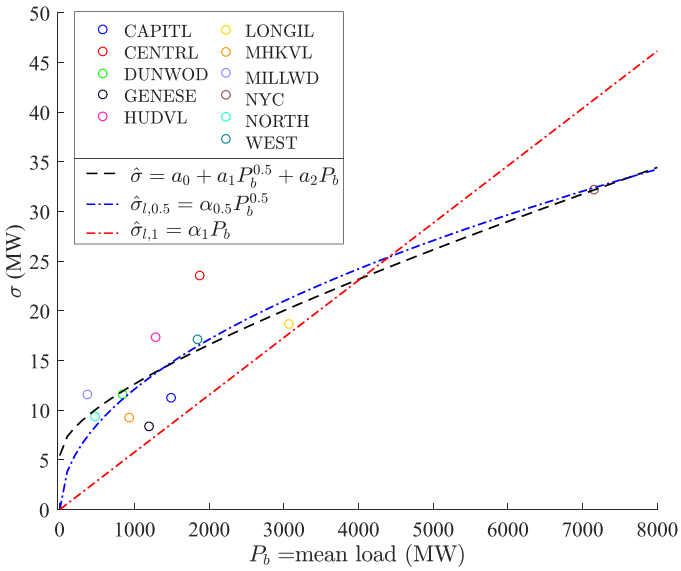


Fig. 6. Estimated $\hat{\sigma}$ with $n_p = 2$, $n_e = 0.5$, P_b = mean load (black dashed) and $\hat{\sigma}_l$ (blue and red dashed) vs. median sample $\tilde{\sigma}_m$ (○).

have a significant impact on accuracy. This is encouraging, since the available information for many study systems is often limited to a single peak or mean demand level for each bus or region.

2) First-order ($n_p = 1$) models have a higher error rate than second-order models, regardless of the choice of P_b or n_e.
3) For second-order ($n_p = 2$) models, the small values of a_2 when $n_e = 1$ indicates the second order term contributes little when P_b is small.

Based on the results in Table I, P_b = mean load, $n_p = 2$, $n_e = 1$ has the lowest error, but it only outperforms the $n_e = 0.5$ case by 0.03% (both in bold). However, using $n_p = 2$ and $n_e = 1$ results in 1% poorer performance in the New Zealand system relative to the $n_e = 0.5$ case. Considering the experimental results from both systems, the preferred model, and the model used in the remaining sections, is P_b = mean load, $n_p = 2$, $n_e = 0.5$.

C. NYISO Experimental Results and Comparison With Legacy Models

A visual comparison of the four candidate polynomial models from Table I (with P_b = mean load) is provided in Fig. 5 for the NYISO system. The four polynomial estimators are almost linear and close to each other within the training range (from the minimum $P_b = 381$ MW to the maximum 7416 MW), which is consistent with the results presented in Table I and Fig. 4(a).

For the chosen model ($n_p = 2$, $n_e = 0.5$, and P_b = mean demand), Fig. 6 shows the estimation results (in black dashed line) and the median of the actual standard deviation calculated using (2) from each year in the validation data set (○). The approximately linear relation suggests that, as P_b increases, σ does tend to increase (i.e., σ is a monotonic function of P_b). However, there are two main scenarios that lead to low average estimation accuracy:

TABLE II
ZONAL ESTIMATION PERCENTAGE ERROR

Name	2013	2014	2015	2016	2017	P_b	$\tilde{\sigma}_m$
MILLWD	28.6%	1.4%	19.1%	8.0%	20.9%	381	11.6
NORTH	20.9%	7.9%	10.2%	14.7%	6.3%	726	9.4
DUNWOD	10.0%	1.7%	2.4%	2.9%	2.8%	856	11.6
MHKVL	33.9%	27.1%	33.4%	26.2%	33.7%	971	9.3
GENESE	61.0%	80.5%	84.8%	62.1%	53.4%	1280	8.4
HUDVL	19.1%	27.4%	14.5%	17.5%	21.2%	1345	17.3
CAPITL	21.4%	16.1%	43.2%	33.5%	29.9%	1532	11.3
WEST	5.4%	9.5%	2.3%	11.9%	7.0%	1879	17.1
CENTRL	26.9%	30.6%	33.1%	30.9%	32.2%	1961	23.6
LONGIL	7.7%	8.0%	8.0%	5.5%	10.4%	3259	18.7
NYC	1.0%	2.8%	0.3%	1.1%	1.8%	7416	32.2

$n_p = 2, n_e = 0.5$. Regions are ordered in ascending order of base power P_b (the mean value observed in MW). $\tilde{\sigma}_m$ denotes the median of standard deviations over the test years (in MW).

1) Inconsistency with the monotonic relationship. For example, in terms of P_b, MHKVL < GENESE < HUDVL, and yet GENESE exhibits the lowest load variation (which corresponds to the narrowest distribution among the three in Fig. 2).
2) Large variations in σ for different training years, but with approximately the same P_b. For example, relative to the training year, the base quantity P_b of CAPITL over the test years varies within 7%. However, compared to the training year, the standard deviation $\tilde{\sigma}_t$ varies by as much as 22% over the test years.

Table II presents the detailed percentage errors calculated using (8) for each region and each validation year. Zone GENESE has particularly high $\hat{\varepsilon}$ (more than 50%) due to its inconsistency with the model. In Fig. 6, the actual standard deviation $\tilde{\sigma}_m$ of GENESE is the lowest among all regions despite having a significant P_b. Its relatively small $\tilde{\sigma}_m$ also leads to high percentage error. As seen in Table II, the error rate of CAPITL varies over the test years from around 16.1% to 43.2%, indicating load

TABLE III
ESTIMATION ERROR OF THE LEGACY MODELS

P_b	n_e	α_{n_e}	ε_m
mean	0.5	0.38299	19.51%
mean	1	0.00577	42.69%

Data source: 2012 to 2017 summers (June, July, August). Modified mean ε_m was calculated using (9). For comparison, the ε_m value obtained using the proposed model (P_b = average demand, $n_p = 2, n_e = 0.5$) is 17.45%.

variation changes significantly from year to year in this region. This suggests that the average demand is not sufficient information to accurately estimate σ for this system. Nonetheless, the estimated σ does a good job of capturing the load deviation behavior for most regions.

In previous papers, a linear relationship between P_b (or $\sqrt{P_b}$) and σ is assumed. The legacy models are denoted by

$$\sigma_{l,n_e} = \alpha_{n_e} P_b^{n_e} \tag{10}$$

where $n_e = 0.5$ or 1. Note that these models can be viewed as special cases of (3) with $n_p = 1$, $a_0 = 0$ and $\alpha_{n_e} = a_1$. The estimation results of two legacy models (10) with $n_e = 0.5$ and 1 (the choices used most often in the literature) are presented in blue and red dashed lines in Fig. 6, and the corresponding coefficients and errors are provided in Table III. The estimation results and optimal coefficients are based on the same data used by the proposed models. In prior work, the coefficients were either not given (e.g., [5]) or given without explanation (e.g., 3.333% in [12]).

It is clear from Fig. 6 that using $n_e = 1$ (i.e., $\sigma_{l,1} = \alpha_1 P_b$) does not work well for this system, since it underestimates the load variation when demand is low and overestimates load variation when demand is high. The results for the legacy model with $n_e = 0.5$ are significantly better than the $n_e = 1$ case, but it still underperforms relative to the proposed polynomial model ($\varepsilon_m = 19.51\%$ vs. 17.45%).

So far, all the results are generated given ΔP in MW. Sometimes we are interested in relative load change with respective to the base demand, i.e., $\Delta P/P_b$ (e.g., to determine the amount of load following reserves to acquire as a percentage of peak load). Given $\Delta P \sim \mathcal{N}(\Delta\bar{P}, \hat{\sigma}^2)$, $\Delta P/P_b$ follows the normal distribution $\mathcal{N}(\Delta\bar{P}/P_b, \hat{\sigma}_r^2)$ where $\hat{\sigma}_r = \hat{\sigma}/P_b$. If $\hat{\sigma} = a_0 + a_1\sqrt{P_b} + a_2 P_b$ (the proposed polynomial model), then

$$\hat{\sigma}_r = \frac{\hat{\sigma}}{P_b} = a_2 + \frac{a_1}{\sqrt{P_b}} + \frac{a_0}{P_b}. \tag{11}$$

Fig. 7 shows that as P_b increases, the observed relative load variation decreases from about 3.2% to 0.5% of P_b. In addition to the observed $\tilde{\sigma}_n$ values, where $\tilde{\sigma}_n = \tilde{\sigma}/P_b$ with $\tilde{\sigma}$ calculated using (2), three plots are provided:

- Estimates of $\hat{\sigma}_r$ based on (11), with $n_p = 2$, $n_e = 0.5$ (black dashed).
- Estimates of $\hat{\sigma}_l$ based on legacy linear models with $n_e = 0.5$ (blue dashed) or 1 (red dashed).

The blue and black dashed lines show a reciprocal relationship of $\hat{\sigma}_r$ and P_b that clearly fits the actual values ($\circ$) better than using a constant $\hat{\sigma}_r$ for all regions. The reciprocal function indicates that, compared to its load base, small regions have large relative fluctuations. As P_b increases, the sampled values show the asymptotic behavior predicted by (11) (i.e., $\sigma_r \to a_2$ as $P_b \to \infty$). Additionally, the first-order legacy models reported in previous work, with coefficients $\alpha_1 = 0.05$ [2] or $\alpha_1 = 0.0333$ [12], would significantly overestimate the relative load fluctuations in almost all regions, since these first-order legacy models correspond to horizontal lines in Fig. 7 at $\sigma_r = 0.05$ and $\sigma_r = 0.0333$.

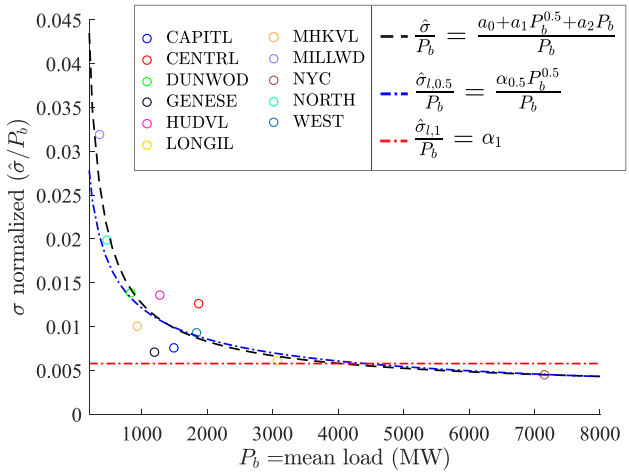


Fig. 7. Estimated $\hat{\sigma}/P_b$ with $n_p = 2$, $n_e = 0.5$, P_b = mean load (black dashed) and $\hat{\sigma}_l$ (red and blue dashed) vs. median of normalized $\tilde{\sigma}_n$ ($\circ$).

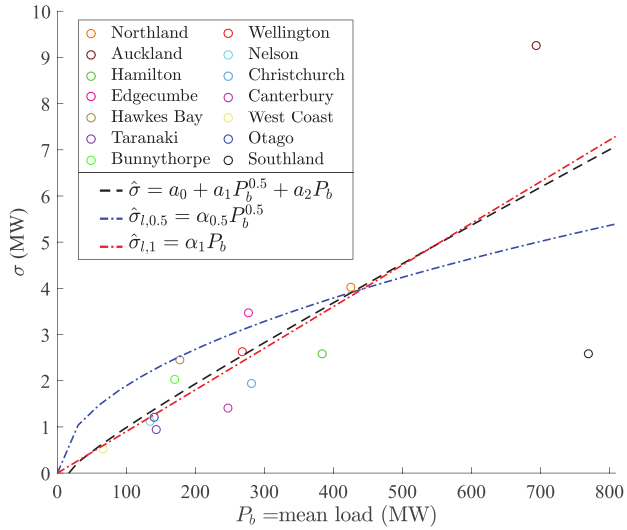


Fig. 8. NZ system estimated $\hat{\sigma}$ with $n_p = 2$, $n_e = 0.5$, P_b = mean load (black dashed) vs. sample standard deviation $\tilde{\sigma}_t$ ($\circ$).

D. New Zealand System Experimental Results

Fig. 8 shows the estimation results of the New Zealand system's regional load fluctuation based on data from November 2017 to January 2018 with a 0.5% sample data cutoff (i.e., omitting the 0.25% maximum and 0.25% minimum values).

As shown in Fig. 8, there is still a nearly-linear relationship for the choice $n_e = 0.5$, $n_p = 2$, P_b = mean load, generated with the optimal model parameters $a_0 = -0.28757$, $a_1 = 0.056405$, $a_2 = 0.0071212$. Table IV shows a comparison of estimation

TABLE IV
ESTIMATION PERCENTAGE ERROR OF DIFFERENT MODELS

Name	ε %	$\varepsilon_{l,1}$ %	$\varepsilon_{l,0.5}$ %	P_b	$\tilde{\sigma}_t$
West Coast	21.91%	30.91%	214.77%	77	0.53
Otago	13.18%	7.72%	87.98%	145	1.22
Taranaki	50.87%	46.53%	150.44%	151	0.93
Nelson	18.61%	27.27%	113.61%	157	1.11
Bunnythorpe	29.59%	29.92%	6.82%	190	2.45
Hawkes Bay	17.58%	13.58%	30.69%	193	2.02
Edgecumbe	24.52%	23.94%	6.59%	293	3.48
Canterbury	67.09%	95.12%	135.15%	305	1.41
Wellington	3.32%	5.18%	26.19%	307	2.63
Christchurch	37.46%	53.20%	77.88%	328	1.93
Hamilton	37.65%	42.60%	48.58%	408	2.58
Northland	3.11%	3.28%	1.11%	462	4.03
Auckland	33.69%	23.42%	42.56%	786	9.26
Southland	162.10%	176.96%	107.12%	791	2.57
ε_m	29.62%	33.28%	69.47%	-	-

Modified mean ε_m was calculated using (9). $n_p = 2, n_e = 0.5, P_b =$ mean load. Regions are ordered in ascending order of base power P_b (in MW). $\tilde{\sigma}_t$ denotes the sample standard deviations (in MW).

errors obtained using the proposed model and the legacy models with $n_e = 0.5$ and $n_e = 1$. The corresponding plots of the proposed polynomial model, along with the two legacy models, are provided in Fig. 8. The optimal legacy model coefficients were calculated to be $\alpha_{0.5} = 0.1896$ and $\alpha_1 = 0.009015$.

The percentage error in this system is generally higher compared to that of the NYISO system due to significantly smaller demand and fluctuation levels. In particular, for the Southland region, the actual standard deviation $\tilde{\sigma}$ of the training set is 3.7 MW while it is 2.57 MW for the testing set, which results in a very high percentage error. Another possible reason for the poorer performance is the significantly reduced data size (3 times less training data and 15 times less validation data) available for this system. Nevertheless, the proposed method is able to outperform the legacy models for this system. However, in contrast to the NYISO results, the accuracy of $\sigma_{l,1} = \alpha_1 P_b$ is comparable to the proposed model, while $\sigma_{l,0.5}$ performs poorly. There are two key reasons for the difference in the legacy model performance for the two systems. First, the actual standard deviations demonstrate a nearly linear relationship with respect to P_b in the NZ system. Second, the optimal offset term in the proposed polynomial model in this case study ($a_0 = -0.288$) is close to zero, as compared to 5.44 for the NYISO system, and it is the significant a_0 term in the NYISO system that leads to poor performance of the $\sigma_{l,1}$ model.

The New Zealand system is different from New York in many aspects, such as population, load composition and climate, leading to different a_i values than those reported in Table I. In addition to illustrating that the same form for the uncertainty model ($n_e = 0.5,\ n_p = 2,\ P_b =$ mean load) does a good job for two very different power systems, it also shows that the parameters used to estimate σ should be ideally based on a similar system.

V. CONCLUSION

In this paper, we propose a data-driven polynomial model to estimate the standard deviation of load fluctuations. The method can be extended to other systems and used for approximating load fluctuation levels for power system studies including generation control, probabilistic power flow, parameter estimation and stability analysis. A testing framework is also provided for model identification and evaluation. The normality assumption was first verified visually based on historical data. Parameters for different models have been estimated using standard linear regression, and representative coefficients derived from NY demand data are provided in Table I for a variety of choices in polynomial order and base power selection. Based on the results from the two test systems (New York and New Zealand), it appears that nearly-linear models provide simple, accurate predictions of load fluctuation based only on the base power (e.g., mean demand) within a region. As seen in Table I, the choice of P_b which characterizes the demand level does not have a significant impact on estimation accuracy, so this can be chosen based on the information availability. The proposed models outperform models previously used in system studies, in some cases by a large margin (e.g., for the NYISO system, the first-order legacy model $\sigma = \alpha_1 P_b$ results in a 42.69% average error versus the 17.45% obtained with the proposed model $\hat{\sigma} = a_0 + a_1\sqrt{P_b} + a_2 P_b$).

Further improvements to the probabilistic model could include modeling correlation between different regions [21] (e.g., based on electrical or geographic distance) and autocorrelations within a region (e.g., to capture dependence between consecutive load fluctuations) [22]. Another potential improvement is to include exogenous information in addition to the power demand (e.g., proportion of load that is residential, industrial, or commercial) that has been used in other load modeling approaches [23]. Additionally, improved σ models can be evaluated in end-use applications such as probabilistic power flow [24] and reserve scheduling [25] to gauge the impact on system performance.

REFERENCES

[1] H. Seifi and M. S. Sepasian, *Electric Power System Planning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. [Online]. Available: http://link.springer.com/10.1007/978-3-642-17989-1

[2] A. M. Leite da Silva, J. L. Jardim, L. R. de Lima, and Z. S. Machado, "A method for ranking critical nodes in power networks including load uncertainties," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1341–1349, Mar. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7094325/

[3] D. Apostolopoulou, A. D. Dominguez-Garcia, and P. W. Sauer, "An assessment of the impact of uncertainty on automatic generation control systems," *IEEE Trans. Power Syst.*, vol. 31, no. 4, pp. 2657–2665, Jul. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7287795/

[4] D. Trudnowski, W. McReynolds, and J. Johnson, "Real-time very short-term load prediction for power-system automatic generation control," *IEEE Trans. Control Syst. Technol.*, vol. 9, no. 2, pp. 254–260, Mar. 2001. [Online]. Available: http://ieeexplore.ieee.org/document/911377/

[5] T. Sasaki and K. Enomoto, " Dynamic analysis of generation control performance standards," *IEEE Trans. Power Syst.*, vol. 17, no. 3, pp. 806–811, Aug. 2002. [Online]. Available: http://ieeexplore.ieee.org/document/1033729/

[6] K. Wang and M. L. Crow, "The Fokker-Planck equation for power system stability probability density function evolution," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 2994–3001, Aug. 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6416991/

[7] S. Peng, J. Tang, and W. Li, "Probabilistic power flow for AC/VSC-MTDC hybrid grids considering rank correlation among diverse uncertainty sources," *IEEE Trans. Power Syst.*, vol. 32, no. 5, pp. 4035–4044, Sep. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7801859/

[8] J. Tang, F. Ni, F. Ponci, and A. Monti, "Dimension-adaptive sparse grid interpolation for uncertainty quantification in modern power systems: Probabilistic power flow," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 907–919, Mar. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7069182/

[9] C. Gu and P. Jirutitijaroen, "Dynamic state estimation under communication failure using kriging based bus load forecasting," *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 2831–2840, Nov. 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7001687/

[10] S. M. Abdelkader and D. J. Morrow, "Online thevenin equivalent determination considering system side changes and measurement errors," *IEEE Trans. Power Syst.*, vol. 30, no. 5, pp. 2716–2725, Sep. 2015. [Online]. Available: http://ieeexplore.ieee.org/document/6945393/

[11] W. Li, *Probabilistic Transmission System Planning*. Hoboken, NJ, USA: Wiley, Feb. 2011, doi: 10.1002/9780470932117.

[12] K. N. Hasan, R. Preece, and J. V. Milanovic, "Priority ranking of critical uncertainties affecting small-disturbance stability using sensitivity analysis techniques," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2629–2639, Jul. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7593359/

[13] M. Perninge, M. Amelin, and V. Knazkins, "Load modeling using the Ornstein-Uhlenbeck process," in *Proc. IEEE 2nd Int. Power Conference.*. Johor Bahru, Malaysia: IEEE, Dec. 2008, pp. 819–821. [Online]. Available: http://ieeexplore.ieee.org/document/4762586/

[14] C. Roberts, E. M. Stewart, and F. Milano, "Validation of the ornstein-uhlenbeck process for load modeling based on PMU measurements," in *Proc. IEEE Systems. Comput. Conf. (PSCC)*. Genoa, Italy: IEEE, Jun. 2016, pp. 1–7. [Online]. Available: http://ieeexplore.ieee.org/document/7540898/

[15] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, Eds., *Exploring Data Tables, Trends, and Shapes (Wiley Series in Probability and Statistics)*. Hoboken, NJ, USA: Wiley, 2006. [Online]. Available: http://doi.wiley.com/10.1002/9781118150702

[16] C. Forbes, M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions*. Hoboken, NJ, USA: Wiley, 2011.

[17] NYISO, "NYISO markets and operational data: Actual load," [Online]. Available: http://www.nyiso.com/public/markets_operations/market_data/load_data/index.jsp. Accessed on: Jan. 6, 2018.

[18] "Transpower system operator: Load graphs," [Online]. Available: https://www.transpower.co.nz/system-operator/operational-information/load-graphs#download. Accessed on: Jan. 6, 2018.

[19] The New York Independent System Operator, Inc., "2017 load and capacity data report," Rensselaer, NY, USA, Tech. Rep., Apr. 2017, Table I-4d: Historical NYCA System Peak Demand. [Online]. Available: https://www.nyiso.com/public/webdocs/markets_operations/services/planning/Documents_and_Resources/Planning_Data_and_Reference_Docs/Data_and_Reference_Docs/2017_Load_and_Capacity_Data_Report.pdf. Accessed on: Nov. 28, 2018.

[20] "Electricity peak demand forecasts, overview of our peak demand forecast methodology," Transpower New Zealand Limited, *Tech. Rep.*, Sep. 2016. [Online]. Available: https://www.transpower.co.nz/sites/default/files/plain-page/attachments%/Transpower%20National-Regional%20Peak%20Demand%20Forecasts%20Jul-2016%20Infor%mation%20Document_0.pdf. Accessed on: Apr. 25, 2018.

[21] W. Li and R. Billinton, "Effect of bus load uncertainty and correlation in composite system adequacy evaluation," *IEEE Trans. Power Syst.*, vol. 6, no. 4, pp. 1522–1529, Nov. 1991. [Online]. Available: http://ieeexplore.ieee.org/document/116999/

[22] I. Moghram and S. Rahman, "Analysis and evaluation of five short-term load forecasting techniques," *IEEE Trans. Power Syst.*, vol. 4, no. 4, pp. 1484–1491, Nov. 1989. [Online]. Available: http://ieeexplore.ieee.org/document/41700/

[23] H. Li, A. L. Bornsheuer, T. Xu, A. B. Birchfield, and T. J. Overbye, "Load modeling in synthetic electric grids," in *Proc. Texas Power Conf.*., Feb. 2018, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/8312059/

[24] N. Hatziargyriou, T. Karakatsanis, and M. Papadopoulos, "Probabilistic load flow in distribution systems containing dispersed wind power generation," *IEEE Trans. Power Syst.*, vol. 8, no. 1, pp. 159–165, Feb. 1993. [Online]. Available: http://ieeexplore.ieee.org/document/221262/

[25] G. Contaxis and J. Kabouris, "Short term scheduling in a wind/diesel autonomous energy system," *IEEE Trans. Power Syst.*, vol. 6, no. 3, pp. 1161–1167, Aug. 1991. [Online]. Available: http://ieeexplore.ieee.org/document/119261/

Zhen Dai (S'12) received the B.E. degree from Tsinghua University, Beijing, China, in 2011, and the M.A.Sc. degree in electrical engineering from the University of Toronto, Toronto, ON, Canada, in 2014. She is currently working toward the Ph.D. degree in electrical engineering at the University of Toronto.

Joseph Euzebe Tate (S'03–M'08) received the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 2005 and 2008, respectively. He is currently an Associate Professor with the University of Toronto, Toronto, ON, Canada. His research focuses on combining advanced telemetry, data processing, and visualization techniques to facilitate renewable integration and improve power system reliability.