An Integrated Gaussian Process Modeling Framework for Residential Load Prediction

Guangrui Xie, Xi Chen, and Yang Weng, *Member, IEEE*

***Abstract*—While adding new capabilities, the distributed energy resource (DER) proliferation raises great concern about challenges such as dynamic fluctuations of voltages. For robust and efficient operational planning purposes, we propose an integrated Gaussian process (IGP) modeling framework for reliable hourly load prediction. The proposed IGP modeling framework has the following unique features: 1) the IGP utilizes not only the data streams generated by the target customer, but also those generated by relevant customers in the power system; an effective input space dimension reduction method is proposed to significantly improve the computational efficiency, while maintaining the high predictive accuracy of the IGP; and 2) an adaptive data communication rate controlling scheme is proposed to further enhance the predictive performance of the IGP by optimally and dynamically adjusting the data communication rate used for each customer under the total data communication bandwidth constraint often imposed. Taking into account the highly uncertain load and generation behaviors of DERs, the proposed IGP framework is tested on various standard IEEE test cases with load and renewable generation data collected from real-world power systems with DERs. The superiority and efficacy of the IGP are verified by our simulation results.**

***Index Terms*—Load forecasting, Gaussian processes, adaptive sampling, renewable integration.**

I. Introduction

THE electric industry is undergoing structural changes as distributed energy resources (DERs) are integrated into the distribution grids. While adding new capabilities, the DER proliferation raises great concern about the resilience of the power grids. Dynamic fluctuations of voltage profiles, voltage stability, islanding, line work hazards, and the distribution system operating at stability boundaries are some of the troubling issues for distribution grid operation [1]. As a result, highly accurate power state prediction is critical for operational planning purposes to mitigate risks, e.g., capacitor bank scheduling to avoid over-voltages in several hours.

Many methods have been proposed for load forecasting in power systems. These include time series models, semi-parametric additive models, support vector regression, neural networks, Gaussian process (GP) regression, etc.; for a non-exhaustive review, see [2]–[13]. In [3], the authors proposed a multiplicative seasonal ARIMA model with the multi-model partitioning algorithm (MMPA) for short-term load forecasting and anomaly detection. In [4] and [5], semi-parametric additive models with different input variables were adopted for short-term load forecasting. In [6], a sensor-based forecasting model using support vector regression was built and applied for load forecasting in multi-family residential buildings. In [7]–[9], different types of neural network models were proposed for load forecasting. Last but not least, GP models have been recognized as one of the most widely adopted analytics tools, and have been applied to load forecasting; see [10]–[13] for a non-exhaustive review on the existing GP models proposed for load forecasting. These GP models mainly differ in the covariance kernels and input variables used for prediction. In [10], the GP model takes previous load observations as inputs to forecast future load demands up to 24 hours ahead. In [11], the authors constructed the covariance kernel of GP by incorporating weather conditions (temperature, wind speed, wind direction and cloud cover), daily and weekly patterns to predict future loads. In [12], the authors applied GP regression to monthly load forecasting and examined the impact of using different types of covariance kernels including Matérn kernel, neural net kernel, Gaussian kernel and linear kernel on the prediction results. In [13], the author selected three different covariance kernels which take into account periodic time patterns and temperature effects for load forecasting. GP models have favorable properties such as being highly flexible to capture various features exhibited by the data at hand and providing an uncertainty measure for the prediction [14], [15].

Despite their many successful applications, existing methods are significantly lacking in two important aspects: using spatial information and adopting an adaptive sampling plan. First, existing methods typically ignore the information contained in the data generated by spatially correlated customers, which can be valuable for improving load forecasting accuracy achieved for a given target customer. Second, these methods are passive in nature—there does not exist any sampling plan tailored to them such that data of higher quality can be obtained to further enhance load forecasting accuracy. The lack of considerations in these two aforementioned aspects can lead to serious problems when an existing method is applied for forecasting. On the one

Manuscript received January 31, 2018; revised May 21, 2018; accepted June 24, 2018. Date of publication June 29, 2018; date of current version October 18, 2018. This work was supported in part by the ICTAS Junior Faculty Award under Grant 176371 at Virginia Tech. Paper no. TPWRS-00149-2018. *(Corresponding author: Xi Chen.)*

G. Xie and X. Chen are with the Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA 24061 USA (e-mail: guanx92@vt.edu; xchen6@vt.edu).

Y. Weng is with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281 USA (e-mail: yang.weng@asu.edu).

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.2018.2851929

hand, customers living in the nearby areas tend to experience similar weather conditions, have similar socioeconomic status and hence exhibit similar load behaviors. The ignorance of such spatially correlated information may lead to a deterioration of predictive accuracy. On the other hand, poorly designed sampling plans with a limited amount of sampling resources available will result in poor data quality and hence inaccurate forecasts.

We highlight potential opportunities provided by the current infrastructure for improving data quality and also point out the suboptimal data sampling practice currently adopted in industry. With the development of Advanced Metering Infrastructure (AMI), smart meters have been widely deployed to replace traditional meters. Smart meters not only enable two-way communications between customers and utilities but also provide the functionality of real-time adjustment of data sampling and communication rates. In AMI, the data sampled from each customer in the same local area are often first aggregated in a concentrator, then the data at all concentrators are transmitted to the data center of utilities. Since more and more utilities rely on wireless technologies for data communication, wireless providers (e.g., Verizon [16]) often impose a limit on data communication bandwidth. With a fixed bandwidth given, let us consider a distribution grid in which real-time measurements are available for each customer. The current practice in industry is typically either to sample from all customers at a common fixed rate, or to sample and transmit only the data of a proportion of the customers. Clearly, neither of the two data sampling and communication plans is optimal. Since different customers often exhibit distinct load behaviors, such non-adaptive sampling plans may result in the consequence that the customers whose load behaviors are stable are sampled from too often, whereas those having highly variable load behaviors are sampled from too little. In such a case, an intelligent data sampling plan is needed to adaptively adjust the sampling rate of each customer such that a reasonable amount of data is obtained from each customer. Hence, we aim to design a load forecasting framework, which is able to provide accurate predictions when no bandwidth constraint is imposed and to construct an adaptive sampling plan for maintaining the level of predictive accuracy when a bandwidth constraint enforces a data reduction. The data reduction happens when smart meters sample data from customers before the data are aggregated at the concentrators. We argue that the information loss caused by the data reduction can be minimized by adopting an optimal data sampling plan.

In this paper, we propose an online integrated GP (IGP) modeling framework for day-ahead hourly residential load forecasting. IGP enjoys the following unique features: (1) It utilizes not only the data streams generated by the target customer but also those generated by relevant customers in the power system. Due to the "big data" problems arising from using all customers' data streams for load prediction of a given target customer, we further propose an effective dimension reduction method to help significantly reduce the dimension of the input space while maintaining the high predictive accuracy achieved by IGP. (2) It is equipped with an adaptive data communication rate controlling scheme, which enhances the predictive performance of IGP by optimally and dynamically adjusting the communication rate of data generated by each customer under the total data communication bandwidth constraint imposed. The proposed IGP modeling framework for load prediction is tested and verified on various IEEE distribution system test cases, i.e., 8-bus, 14-bus, 24-bus and 123-bus systems. IGP shows superior predictive performance in comparison with alternative prediction methods across all cases tested.

The rest of the paper is organized as follows. In Section II, we provide a brief review of GP modeling in the context of load prediction and present a naive GP modeling approach. Section III introduces the IGP framework with each subsection detailing on one of its features. The numerical experiments are presented in Section IV. Section V concludes the paper.

II. Gaussian Process Modeling for Load Prediction

In a power system, the hourly load data of a particular customer i sampled by smart meters can be seen as a time series $\{P_i^t\}$. To predict future P_i^t, we usually acquire the historical data of P_i^t and that of other variables related to P_i^t, to construct a dataset for training purposes. This training process aims to capture the behavior of P_i^t and its relationship with the relevant input variables.

Specifically, the training dataset consists of a set of input-output pairs, in which the output variable y_t at time instant t is P_i^t, and the input vector at time instant t, $\mathbf{x}_t$, is the combination of the variables relevant to P_i^t in the training dataset. We aim to estimate an underlying function $f(\cdot)$, where $y_t = f(\mathbf{x}_t) + \varepsilon$ for all t based on the training dataset $\mathcal{T} = \{(\mathbf{x}_t, y_t),\ t = 1, 2, \ldots, n\}$, with n denoting the training sample size and ε being the normally distributed observation error with mean zero and variance σ^2.

Various mathematical models can be used to perform this inference, among which GP-based models are the most preferable for their outstanding capability of capturing uncertainties. In a GP model, a Gaussian process prior is typically placed on $f(\cdot)$, for which the process is assumed to have mean μ and covariance (or kernel) function K: $f(\cdot) \sim GP(\mu, K)$, and the unknown hyper-parameters in μ and K can be estimated from the training dataset $\mathcal{T}$. The sample size of the training dataset of GP is usually set to be at least 10 times the dimension of the input space based on the suggestion of [17]. Denote the set of output variables $\{y_t\}$ in the training dataset by $\mathbf{Y}$ and the set of input vectors $\{\mathbf{x}_t\}$ corresponding to $\{y_t\}$ by $\mathbf{X}$. The relationship between the set of future loads P_i^t's and their corresponding input vectors can be described as a testing set $\mathcal{T}^* = (\mathbf{X}^*, f(\mathbf{X}^*))$, where $f(\mathbf{X}^*)$ is unknown and $\mathbf{X}^*$ is typically assumed known for prediction purposes. The joint distribution of $\mathbf{Y}$ and $f(\mathbf{X}^*)$ is given by

$$\begin{pmatrix} \mathbf{Y} \\ f(\mathbf{X}^*) \end{pmatrix} \sim GP\left(\begin{bmatrix} \mu(\mathbf{X}) \\ \mu(\mathbf{X}^*) \end{bmatrix}, \begin{bmatrix} K'(\mathbf{X}, \mathbf{X}) & K(\mathbf{X}, \mathbf{X}^*) \\ K(\mathbf{X}^*, \mathbf{X}) & K(\mathbf{X}^*, \mathbf{X}^*) \end{bmatrix} \right),$$

where $K'(\mathbf{X}, \mathbf{X}) = K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}_n$, and $\mathbf{I}_n$ is the $n \times n$ identity matrix. Then the conditional distribution of $f(\mathbf{X}^*)$ given $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{X}^*$ can be obtained as Gaussian with the predictive mean and variance respectively given by

$$\mathrm{E}[f(\mathbf{X}^*)] = \mu(\mathbf{X}^*) + K(\mathbf{X}^*, \mathbf{X})K'(\mathbf{X}, \mathbf{X})^{-1}(\mathbf{Y} - \mu(\mathbf{X})), \quad (1)$$

$$\mathrm{Cov}(f(\mathbf{X}^*)) = K(\mathbf{X}^*, \mathbf{X}^*) - K(\mathbf{X}^*, \mathbf{X})^\top K'(\mathbf{X}, \mathbf{X})^{-1} K(\mathbf{X}^*, \mathbf{X}) \quad (2)$$

In particular, (1) can be used as an estimate for future P_i^t. The hyper-parameters in μ, K and σ^2 can be estimated by maximizing the log-likelihood function. See [15] for details.

A. A Naive GP Modeling Approach

In a naive GP modeling method (**NGP**) for load prediction which adopts the idea of [18], the input-output pair becomes (t, P_i^t), where the input variable takes only the time instant t. To describe the correlation between the electric loads at different time instants, we choose a simplified version of the quasi-periodic kernel as used in [18] to account for the periodicity exhibited by the load values in power systems. Specifically, the covariance between the electric loads at two time instants t_1 and t_2 is given by

$$K(t_1, t_2) = \tau^2 \exp\left(-\frac{\sin^2(\pi(t_1 - t_2)/T)}{\omega} - \frac{(t_1 - t_2)^2}{\alpha}\right),$$

where τ^2, ω and α are the hyper-parameters to be estimated, and T is the specified cycle, which is 24 in our case, since utilities usually record hourly load P_i^t per day. With the training dataset $\mathcal{T} = \{(t, P_i^t)\}$, we note that NGP does not require any additional information to predict future load other than the observed load P_i^t of customer i at a given time t.

III. An Integrated GP Modeling Framework

The hourly load data consist of a highly volatile time series, as illustrated in Fig. 1. This high variability is due to not only the randomness in the target customer's behavior, but also changes in behaviors of other customers in the power system. NGP utilizes only the historical load data of the target customer for prediction. However, a power system is an integrated system, in which the target customer's load has strong correlations with other customers' power states, as shown in Fig. 2.

Such a strong correlation as shown in Fig. 2 can be derived from the well-known power flow equation:

$$0 = -P_i + \sum_{k=1}^{N} |V_i||V_k|(G_{ik}\cos\theta_{ik} + B_{ik}\sin\theta_{ik}), \quad (3)$$

where P_i is the load of bus i in a power system, $|V_i|$ is the voltage magnitude of bus i, θ_{ik} is the difference in voltage angles of buses i and k, G_{ik} and B_{ik} are coefficients describing the relationship between buses i and k, and N is the total number of buses in the power system. To derive the correlation between the target customer i's load P_i and the sine function of the difference of customer i and his/her neighbor j's voltage angles θ_{ij}, define $a_k = |V_i||V_k|G_{ik}$, $b_k = |V_i||V_k|B_{ik}$, $U_k = \cos\theta_{ik}$, and $W_k = \sin\theta_{ik}$ for $k = 1, 2, ..., i-1, i+1, ..., N$. Then (3) can be reduced to $P_i = \sum_{k=1}^N a_k U_k + \sum_{k=1}^N b_k W_k$. We have the

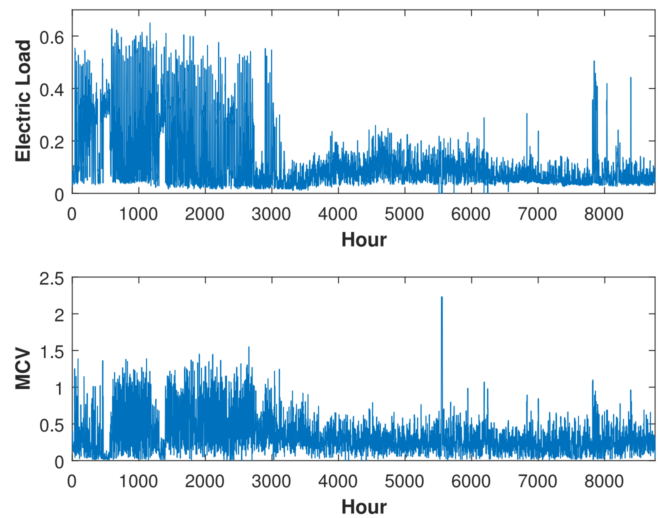


Fig. 1. Hourly load data of a customer and Moving Coefficient of Variation (MCV) of the load throughout one year. Data are drawn from PJM [19] and sampled hourly. Here MCV $= s/\mu$, where s and μ are respectively the moving standard deviation and moving average of loads observed over 5 hours. Notice that MCV can be as high as greater than 2.

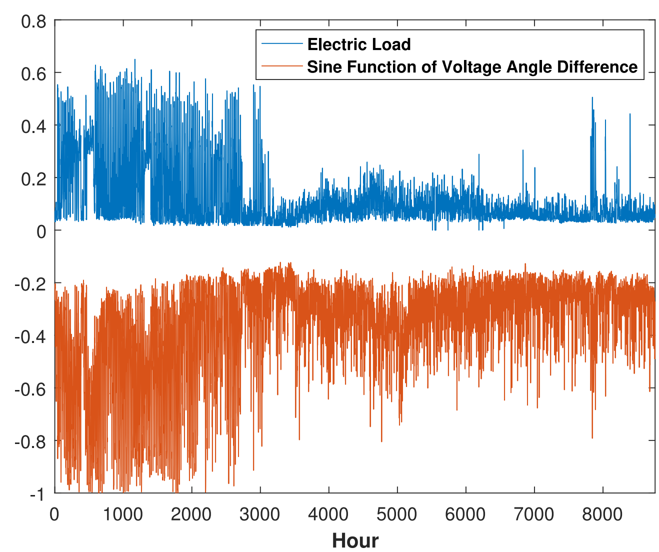


Fig. 2. Hourly load data of a target customer and the sine function of the difference in voltage angles of this customer and his/her neighbor customer in an IEEE 8-bus test case throughout one year. Data are drawn from PJM [19] and sampled hourly. Notice that these two quantities present a strong negative correlation.

correlation between P_i and W_j for $j = 1, ..., i-1, i+1, ..., N$ as

$$\mathrm{Corr}(P_i, W_j) = \frac{\sum_{k=1}^N a_k \mathrm{Cov}(U_k, W_j) + \sum_{k=1}^N b_k \mathrm{Cov}(W_k, W_j)}{(\mathrm{Var}(P_i)\mathrm{Var}(W_j))^{\frac{1}{2}}}. \quad (4)$$

In distribution grids, the magnitudes of the coefficients G_{ik} and B_{ik} of two neighboring customers are much larger than those of the two coefficients of two distant customers (G_{ik} and B_{ik} of two distant customers tend to be 0). Let us consider a simple case in which customer j is the only neighbor of customer i, in

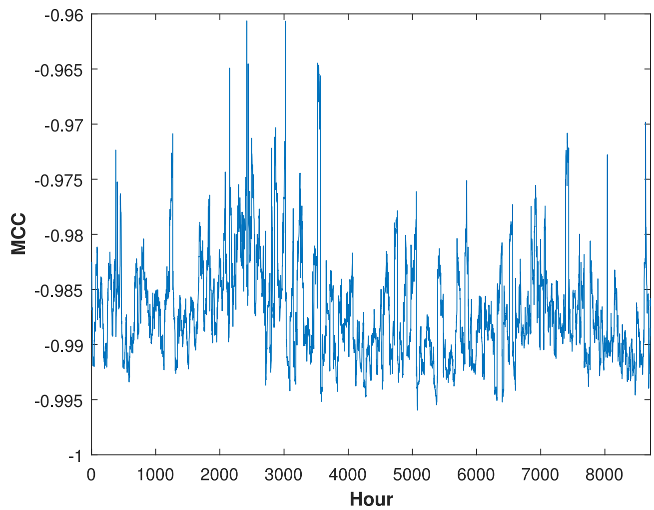


Fig. 3. Moving Correlation Coefficient (MCC) of the load data and the sine function of voltage angle difference in Fig. 2. MCC is calculated using the most recent 50 hours' data as time progresses.

this case we can approximate the correlation in (4) by

$$\begin{aligned}\mathrm{Corr}(P_i, W_j) &\approx (a_j \mathrm{Cov}(U_j, W_j) + b_j \mathrm{Var}(W_j)) \\ &\quad\times \left[\left(a_j^2 \mathrm{Var}(U_j) + b_j^2 \mathrm{Var}(W_j)\right.\right. \\ &\qquad \left.\left. + 2a_j b_j \mathrm{Cov}(U_j, W_j)\right)\mathrm{Var}(W_j)\right]^{-\frac{1}{2}} \\ &= \left(a_j \mathrm{Cov}(U_j, W_j) + b_j \mathrm{Var}(W_j)\right) \\ &\quad \cdot \left[\left(a_j \mathrm{Cov}(U_j, W_j) + b_j \mathrm{Var}(W_j)\right)^2\right. \\ &\qquad + a_j^2 \left(\mathrm{Var}(U_j)\mathrm{Var}(W_j)\right. \\ &\qquad \left.\left. - \mathrm{Cov}(U_j, W_j)^2\right)\right]^{-\frac{1}{2}} \qquad (5)\end{aligned}$$

In (5), we note that U_j and W_j are cosine and sine functions of θ_{ij}, hence the correlation between U_j and W_j depends on the distribution of θ_{ij}. Numerical evaluations have shown that the correlation between U_j and W_j is usually very close to -1, hence $\mathrm{Var}(U_j)\mathrm{Var}(W_j) \approx \mathrm{Cov}(U_j, W_j)^2$. Therefore, we have $\mathrm{Corr}(P_i, W_j) \approx 1$ or -1, which indicates that there is a strong correlation between the load of the target customer and the sine function of the difference in the voltage angels of the target customer and his/her neighbors. In Fig. 3 we show the Moving Correlation Coefficient (MCC) between the two quantities as given in Fig. 2; we can see the presence of a strong correlation throughout the time horizon of interest.

A. Integrated GP Modeling: Utilizing Neighbors' Data

To account for the strong correlation in customers' data, we incorporate the information of other customers to form the input vector. We see from (3) that P_i depends on quantities such as $|V_i|$, $|V_k|$, $\cos\theta_{ik}$ and $\sin\theta_{ik}$. For the prediction of P_i for customer i, one way of incorporating these quantities into the input vector would be to set its value at time t as

$$\begin{aligned}\mathbf{x}_t &= \left(C_{i,1}^t, C_{i,2}^t, \ldots, C_{i,i-1}^t, C_{i,i+1}^t, C_{i,i+2}^t, \ldots, C_{i,N}^t\right. \\ &\quad \left.\times S_{i,1}^t, S_{i,2}^t, \ldots, S_{i,i-1}^t,\ S_{i,i+1}^t, S_{i,i+2}^t, \ldots, S_{i,N}^t\right)^\top,\end{aligned}$$

where t denotes the time variable, $C_{i,k}^t = |V_i^t||V_k^t|\cos\theta_{ik}^t$, and $S_{i,k}^t = |V_i^t||V_k^t|\sin\theta_{ik}^t$.

However, the aforementioned setting essentially applies GP to approximate a linear relationship, which cannot fully utilize GP's capability; moreover, it entails an input space of dimension $2(N-1)$, and numerical issues often arise when N is large. According to [20]–[22], there is usually a strong coupling between load and voltage angle, whereas the coupling between load and voltage magnitude is rather weak. Therefore, we specify the input vector at time t as $\mathbf{x}_t = (\theta_{i,1}^t, \theta_{i,2}^t, \ldots, \theta_{i,i-1}^t, \theta_{i,i+1}^t, \theta_{i,i+2}^t, \ldots, \theta_{i,N-1}^t, \theta_{i,N}^t)^\top$, where $\theta_{i,k}^t$ denotes the difference in voltage angles of bus i and bus k at time instant t. The dimension of the input space now becomes $N-1$. We choose an automatic relevance determination (ARD) kernel to model the correlation between the loads corresponding to two arbitrary input vectors because of its well-known capability of providing sensitivity analysis [15], [23]. Sensitivity analysis is important for further reducing the dimensionality of input space, which will detailed in Section III-B. Then the covariance between the loads corresponding to input vectors $\mathbf{x}_{t_1}$ and $\mathbf{x}_{t_2}$ follows as

$$\begin{aligned}K(\mathbf{x}_{t_1}, \mathbf{x}_{t_2}) = \tau^2 \exp&\left(-\sum_{k=1}^{i-1} \frac{(\theta_{i,k}^{t_1} - \theta_{i,k}^{t_2})^2}{2\alpha_k^2}\right) \\ &\times \exp\left(-\sum_{k'=i+1}^{N} \frac{(\theta_{i,k'}^{t_1} - \theta_{i,k'}^{t_2})^2}{2\alpha_{k'}^2}\right),\end{aligned}$$

where t_1 and t_2 are two arbitrary time instants, and α_k represents the length-scale parameter corresponding to customer k in the kernel function. We refer to this method as the integrated GP-based prediction method (**IGP**) under the IGP modeling framework.

Remark 1: The fact that real power highly depends on voltage angles relies on the assumption that X/R ratio is high, which is not always true in distribution systems. Nevertheless, we argue that the power flow equation (3) serves only as an inspiration to extract the most relevant and valuable information for the load prediction purpose. As a data-driven method, we do observe a much stronger correlation between load and voltage angle than that between load and voltage magnitude in our simulations with various data sources, so retaining only voltage angles as the input variables suffices for the load prediction purpose.

B. Input Space Dimension Reduction for Real-Time Implementation

An advantage of using ARD kernel is that the length-scale parameter estimates shed some light into the correlations between the behaviors of distinct customers in the power system, and such information is crucial for further reducing the

TABLE I
ESTIMATES OF α_k^2'S FOR CUSTOMERS IN AN IEEE 8-BUS TEST CASE FOR THE PREDICTION OF CUSTOMER 1 WITH DATASETS d_1, d_2, AND d_3

α^2	Customer	d_1	d_2	d_3
$\widehat{\alpha}_2^2$	2	8.364	0.480	1.387
$\widehat{\alpha}_3^2$	3	6.179×10^3	1.243×10^2	7.871×10^3
$\widehat{\alpha}_4^2$	4	1.141×10^3	1.219×10^3	1.888×10^5
$\widehat{\alpha}_5^2$	5	4.722×10^3	1.836×10^3	1.909×10^3
$\widehat{\alpha}_6^2$	6	1.516×10^4	3.228×10^3	9.934×10^3
$\widehat{\alpha}_7^2$	7	3.571×10^3	2.123×10^3	1.039×10^5
$\widehat{\alpha}_8^2$	8	2.327×10^3	3.740×10^3	1.177×10^3

computational cost required by IGP, which scales as $O((N-1)^3)$, where N denotes the total number of buses in the system.

When performing prediction for customer i, those customers that are highly correlated with customer i typically feature a low value of their estimated α_k, which is obtained in the training process. This makes intuitive sense because when α_k is significantly lower than the corresponding α's of other customers, $\theta_{i,k}$ plays a major role in the covariance kernel, so that the behavior of customer k has a strong impact on the prediction for customer i. In such a case, we call customer k a *neighbor* of customer i. The neighbor in this sense is not based on geographical relationships, but based on prediction relevance, though neighbors may also be in close proximity geographically in some simple distribution systems.

Table I shows the $\widehat{\alpha}_k^2$'s obtained for prediction of customer 1 in an IEEE 8-bus test case with three different training datasets using IGP. We can see that $\widehat{\alpha}_2^2$ is significantly lower than the other $\widehat{\alpha}_k^2$'s, which indicates that customer 2 is a neighbor of customer 1. Hence, we can reduce the input space dimension for real-time implementation of IGP upon identifying relevant neighbors. Since the information of non-neighbors is of little value, we can retain only the difference in voltage angles of the target customer and his/her neighbors in the input vector for future training and prediction. Such a procedure is similar to those for feature selection in the machine learning literature. Specifically, in real-time implementation, assume that we want to predict for customer i. Let the length-scale parameter corresponding to customer j be α_j. Define the contribution coefficient of customer j as $c_j = \alpha_j^{-2}/(\sum_{k=1}^{i-1}\alpha_k^{-2} + \sum_{k'=i+1}^{N}\alpha_{k'}^{-2})$. We sort the contribution coefficients in a non-increasing order, and select the customers from the first till the one by including which the sum of contribution coefficients exceeds a prespecified threshold. The customers selected are then identified as the neighbors of customer i.

In this paper, the IGP model after this dimension reduction step is referred to as **RIGP** and the IGP with full dimensional information is referred to as **FIGP** under the IGP modeling framework.

C. Prediction With Uncertain Inputs

In addition to the differences in their respective input vectors and kernels used, another major difference between NGP and IGP lies in their prediction steps. For making predictions via (1), the input vectors of NGP at the time instants of prediction contain only the time instants of interest, which are naturally known to us. In contrast, the input vectors of IGP at the time instants of prediction contain the differences in future phase angles, which are yet to be observed.

We adopt a k-means clustering procedure to address the problem that future input vectors are unknown. Since customers usually present distinct load behaviors at different hours of a day, upon the input space dimension reduction step we group the hourly sampled input vectors in the training dataset into 24 sets, such that each set corresponds to one hour of a day. We then apply k-means clustering on each set to cluster the input vectors sampled at each hour; see [24] for detailed implementation of k-means algorithm. There are various algorithms to determine the number of clusters to use; see [25], [26]. To reduce implementation complexity, following the rule of thumb in [25], we set the number of clusters to $k = \lceil\sqrt{n/2}\rceil$, i.e., the smallest integer not less than $\sqrt{n/2}$, where n denotes the number of days that the training data are sampled from. By assuming similarity in load behaviors of each customer on any two consecutive days, we expect that the input vectors of these two days to lie in the same cluster in each of the 24 sets. Therefore, in each set (for each hour), we use the centroid of the cluster which covers the input vector of the day prior to the prediction day as $\mathbf{x}^*$ in (1) for each hour's prediction on the prediction day.

For quantifying the prediction uncertainty of the IGP predictor, the standard way of building a 95% prediction interval no longer applies as the input vector $\mathbf{x}^*$ for IGP is random. As a remedy, we find the maximum and minimum loads at each hour among those days where the training data are sampled from, and then plug their corresponding input vectors into (1) to construct the upper and lower bounds for prediction at each hour on a given prediction day. In this way, we obtain an upper bound and a lower bound respectively close to the maximum and minimum P_i at each hour in the training dataset.

D. Adaptive Control of Data Communication Rates

In this subsection, we introduce an adaptive data communication rate controlling scheme under the IGP modeling framework to improve data quality and hence the subsequent predictive performance achieved by IGP.

With a fixed wireless data communication plan, an adaptive data communication rate controlling scheme can help adjust the communication rate assigned to each customer according to the individual dynamic load behavior, such that data of improved quality can be obtained subject to a fixed bandwidth constraint. An effective and efficient scheme must address the following two pressing questions: which customers in the system to sample data from at a particular hour and which variables to be included in the sample. For the purpose of load prediction, the load data P is the only output variable, so we cannot afford to sample less on P for each customer. Since we only use voltage angle θ to build the input vector for implementing IGP, the key solution is to designing an effective sampling plan for θ that complies with the fixed communication bandwidth given.

From the utilities' point of view, we formulate the optimization program given by (6) to obtain an adaptive communication rate controlling scheme. We define the decision horizon as the time period that the optimal rate controlling scheme is implemented for. We enforce the scheme to sample θ from each customer at least once within the decision horizon for implementing the k-means clustering procedure given in Section III-C. The decision horizon is chosen to be 48 hours, since a longer period will lead to too little data being collected at some hours hence deteriorated predictive accuracy, and a shorter period will naturally forbid the possibility of sampling θ at least once at each hour within the decision horizon given a limited bandwidth. We also assign different weights to each customer in the objective function based on the variabilities of their respective load data collected. Under a bandwidth constraint, the data sampled at a particular time instant may not constitute a complete input vector for prediction of a target customer, since neighbors' voltage angles may be missing. We call the input vectors constituted by all neighbors' θ's "valid" input vectors, and we don't use "invalid" input vectors for training. In (6), customers with a higher variability are assigned larger weights, as more "valid" input vectors are required to accurately predict their loads. The optimization problem is given as follows:

$$
\begin{aligned}
&\max \sum_{i=1}^{N}\sum_{j=1}^{48} w_i c_{ij} \\
\text{s.t.}\quad & x_{i,j} + x_{i,j+24} = 1, \\
& c_{ij} \geq 1 - \left(\sum_{k=1}^{N} A_{ik} - \sum_{k=1}^{N} A_{ik} x_{kj}\right), \\
& c_{ij} \leq \sum_{k=1}^{N} A_{ik} x_{kj} \Big/ \sum_{k=1}^{N} A_{ik}, \\
& \sum_{j=1}^{48}\sum_{i=1}^{N} x_{ij} = n, x_{ij} = \{0,1\}, c_{ij} = \{0,1\}, \\
& i,k = 1,2,\ldots,N,\; j = 1,2,\ldots,48, \qquad (6)
\end{aligned}
$$

where x_{ij} is a binary decision variable, with $x_{ij} = 1$ indicating that θ is sampled at hour j from customer i, and $x_{ij} = 0$ otherwise. c_{ij} is a binary decision variable, with $c_{ij} = 1$ indicating that we can form a "valid" input vector for the prediction of customer i at hour j, and $c_{ij} = 0$ otherwise. A_{ik} is a known coefficient after the identification of neighbors, with $A_{ik} = 1$ denoting that customer k is a neighbor of customer i and $A_{ik} = 0$ otherwise. N is the number of customers in the power system, and n is the total number of data points we can communicate within the decision horizon, which is determined by the bandwidth. ω_i is the weight assigned to each customer and $w_i = s_i/\mu_i$, where s_i and μ_i are respectively the sample standard deviation and sample mean of the load data of customer i. Notice that s_i and μ_i can be obtained from historical dataset of customer i. This optimization problem is an integer linear programming problem, which can be solved by commonly used integer programming solvers such as YALMIP [27].

A flowchart of the IGP modeling framework enhanced by the adaptive data communication scheme is given in Fig. 4. Since the optimization program is run every two days, the flowchart illustrates two days' operation of the scheme. Consider a particular period of two days. At the beginning, we sample P_i's and θ_i's (voltage angle of customer i) from the first day according to the current sampling scheme and combine the data with historical P_i's and θ_i's to constitute a dataset D_1. We then pre-process D_1 to obtain dataset D_1', which contains P_i's and voltage angle difference θ_{ij}'s. Then D_1' is used for training to obtain the parameter estimates required by IGP, and the first day prediction is performed subsequently. For the second day of the period, we perform the same sampling procedure and combine the sampled data with historical data to constitute D_2; then we perform the same pre-processing step to obtain D_2' and update ω_i's based on D_2. Subsequently, D_2' is used for training to obtain parameter estimates as well as neighboring relations A_{ij}'s. The entire procedure is performed periodically with two days as the cycle.

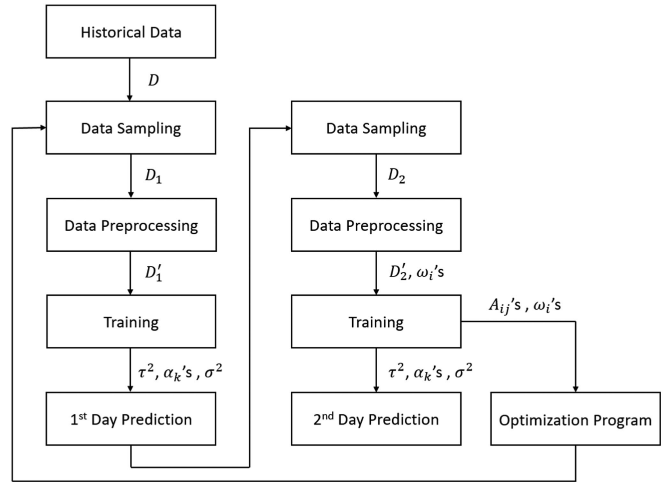


Fig. 4. Flowchart of IGP with the adaptive data communication scheme for short-term load prediction.

Remark 2: The aforementioned scheme presents an optimal hourly sampling plan for load forecasting and entails frequent sampling of phase angles and loads of multiple customers, which may require PMUs as metering devices. Although PMUs have not been widely used so far by industry as compared to other smart meters, the deployment of PMUs has been rapidly growing and such devices will become readily available on the market in the near future [28].

IV. Numerical Experiments

Experiments are performed on four IEEE test cases: 8-bus, 14-bus, 123-bus test cases and a 24-bus test case which is constructed to simulate the phase unbalance problem based on the 8-bus test case. To simulate the highly uncertain load behaviors caused by DERs in real-life power systems, historical load profiles from PJM [19] and New York ISO [29] are used for simulations. Specifically, the load data drawn from PJM load profiles for the year 2014 are used for the 8-bus and 123-bus test cases [19], and the load data drawn from New York ISO load profiles for the year 2015 are used for 14-bus and 24-bus test cases [29].

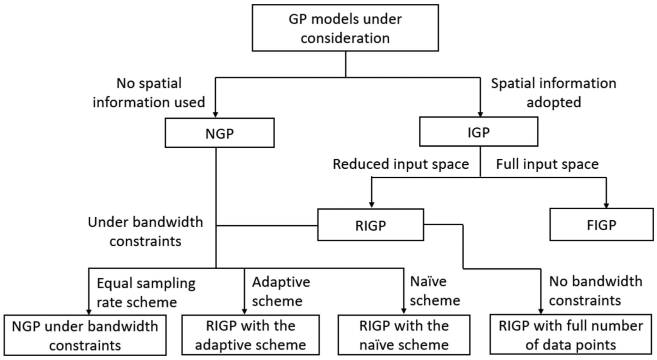


Fig. 5. A diagram showing relationships between the GP models under consideration in Section IV.

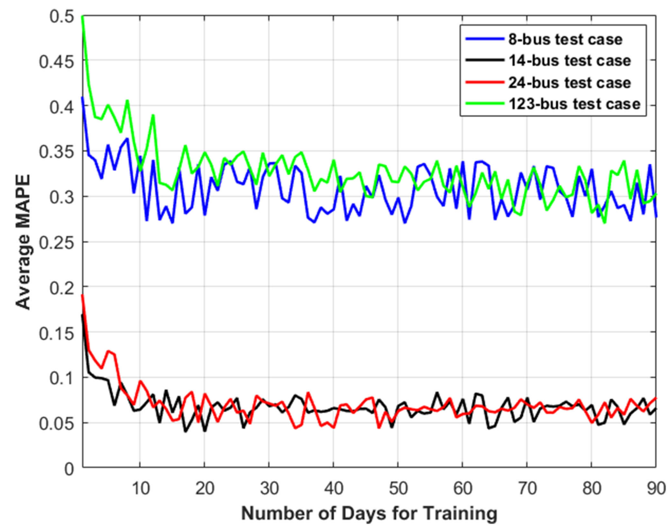


Fig. 6. Model validation for all test cases. For each test case, different numbers of days' data are used for training. MAPEs are recorded for 30 different prediction days in each season for all customers. Then the average MAPEs across all the prediction days and customers are calculated for all numbers of training days in each test case to arrive at the appropriate size to use.

Taking into account the uncertain renewable generation behaviors of DERs, we first pre-process the hourly PV generation data over a year drawn from Renewable.ninja [30], and then subtract the pre-processed data from the load data of each bus. To obtain other measurements such as voltage magnitudes and voltage angles, we perform power flow analysis to generate the states of the power system at every hour over a year using the MATLAB Power System Simulation Package (MATPOWER) [31], [32], based on the processed load data. For all four test cases, we use Mean Absolute Percentage Error (MAPE) to measure the predictive performance of each method: MAPE $= {n_t}^{-1} \sum_{t=1}^{n_t} \left|\frac{P_t - \hat{P}_t}{P_t}\right|$, where P_t is the actual value of load at hour t, $\hat{P}_t$ is the predicted value of load at hour t, and n_t is the number of predictions made. In our experiments, we set $n_t = 24$, $t = 1, 2, \ldots, 24$, i.e., to calculate MAPE for every prediction day.

Since a number of GP models are to be compared in this section, to facilitate the reader's understanding, we summarize the relationships between the GP models considered for comparison in Fig. 5. The details of each model are discussed in the following two subsections.

A. Comparison of NGP, IGP and SARIMA

In this subsection, we compare the predictive performance of NGP, IGP, and the SARIMA model proposed in [3] as a benchmark. For each test case, we perform prediction for 30 different days in each season for all customers using each prediction method. To determine the size of training dataset for IGP, we run a model validation procedure as shown in Fig. 6. In order to use a minimum number of days for training while maintaining a high predictive accuracy by taking into account the suggestion of [17], we decided to use 15 preceding days' data as the training dataset for 8-bus, 14-bus and 24-bus test cases for one day's prediction; for the 123-bus test case, we use 60 days' data instead. For comparison purposes, we use the same size of training dataset for NGP and SARIMA in each experiment.

Fig. 7(a)–(d) shows the prediction results for a target customer on 50 different prediction days in each test case. We observe that the SARIMA model performs the worst in all test cases. NGP performs slightly better than SARIMA. Both RIGP and FIGP outperform NGP and they achieve comparable satisfactory predictive accuracies. We note that the MAPEs obtained in the 14-bus and 24-bus test cases are significantly lower than those in the 8-bus and 123-bus test cases; this is due to the fact that the variability of the load data drawn from New York ISO[29] is significantly lower than that of PJM [19] (see Fig. 1 for the high variability of the load data from PJM).

Fig. 8 shows the boxplots of MAPEs obtained in the 8-bus and 123-bus test cases. We observe that the variations of the MAPEs of FIGP and RIGP are significantly lower than those of NGP and SARIMA, which implies that both of FIGP and RIGP are more robust when applied to multiple customers.

In addition, for a comparison of the coverage abilities achieved by NGP and IGP, we draw a set of data from the 8-bus test case and plot the predictive mean with confidence intervals for one day's prediction of a customer, see Fig. 9. We observe that the bounds constructed for IGP offer a better coverage of the realized loads than the 95% confidence interval obtained by NGP. We also show boxplots of the number of daily load observations covered by the bounds of NGP and IGP in the 8-bus and 123-bus test cases in Fig. 10; notice that IGP exhibits a much higher coverage capability than NGP.

B. Testing Data Communication Rate Controlling Scheme

In this subsection, we test the efficacy of our adaptive data communication scheme. As a benchmark, we design a naive data communication scheme for RIGP, in which data are sampled equally frequently from each customer subject to a bandwidth constraint. We then compare the predictive performance of four schemes: RIGP with the adaptive scheme, NGP with fewer training data points subject to the bandwidth constraint, RIGP with the naive scheme and RIGP with full number of data points as if there were no bandwidth constraint.

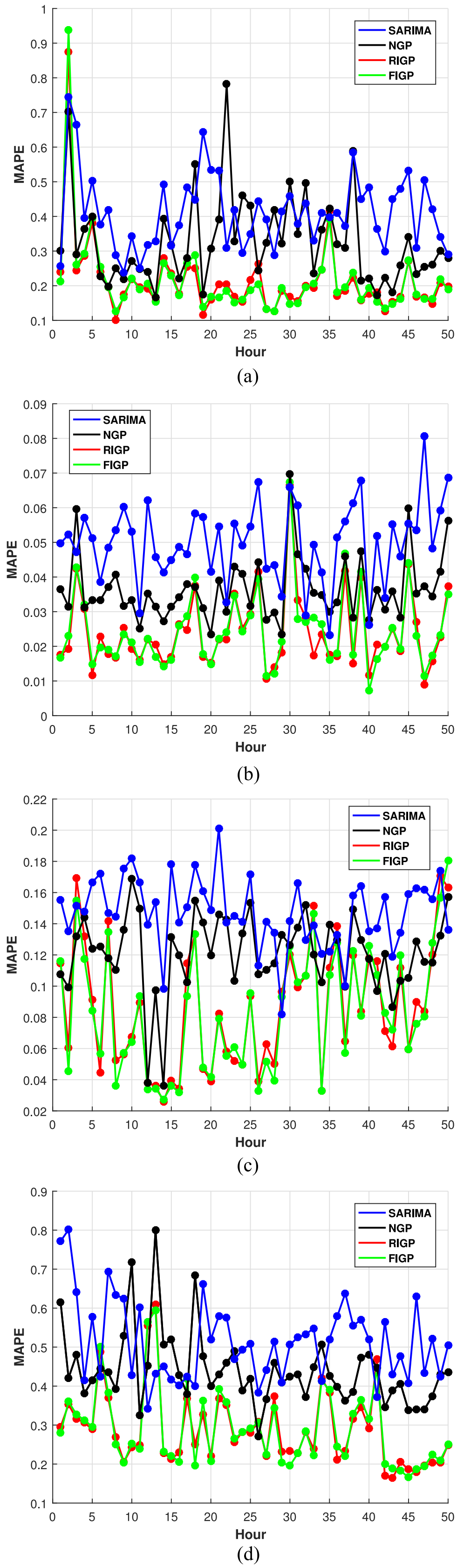


Fig. 7. MAPEs obtained by the SARIMA model, naive GP, reduced-dimension integrated GP and full-dimension integrated GP for the prediction of a customer throughout 50 days in 8-bus, 14-bus, 24-bus and 123-bus test cases.

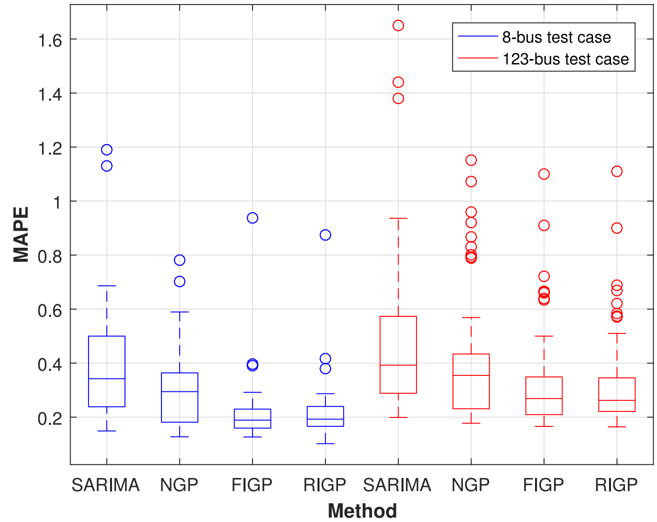


Fig. 8. Boxplots of MAPEs corresponding to the SARIMA model, naive GP, full-dimension integrated GP, and reduced-dimension integrated GP. MAPEs are obtained from the prediction results for 30 days in each season for all customers in both 8-bus and 123-bus test cases.

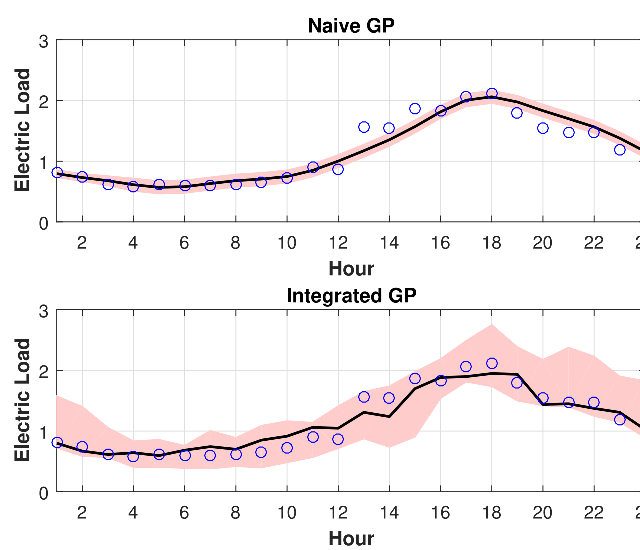


Fig. 9. Coverages achieved by the 95% confidence interval of naive GP and by the bounds of integrated GP for one day's prediction of a customer in an 8-bus test case.

The current design of most smart meters requires 16 bits of memory to record one single data point, e.g., OpenWay CENTRON meter by Itron, Inc [33]. For a power system with N customers, a full-rate data communication scheme that ensures θ is sampled hourly from every customer requires a bandwidth of $384N$ bits every day on average. In this experiment, we consider a scenario that the actual bandwidth is two thirds of the demand of full-rate data communication scheme, which means that the bandwidth budget available of the concentrators can only ensure a total of $256N$ bits of memory for sampling θ from customers every day on average. We consider 2 days as a decision horizon, as argued in Section III-D.

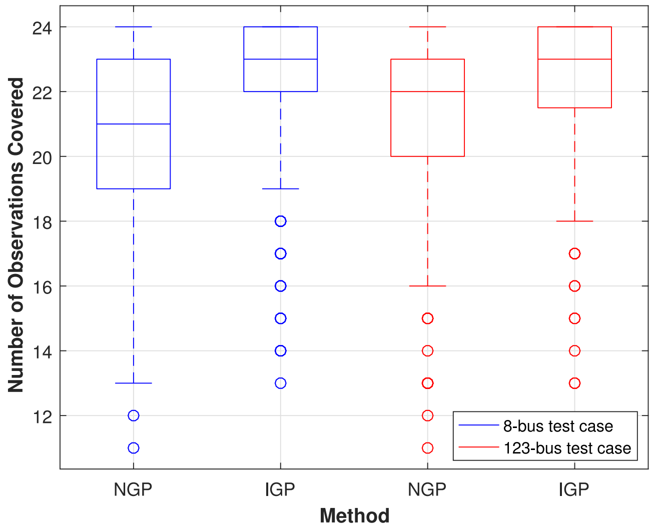


Fig. 10. Boxplots of the number of daily load observations covered by the confidence intervals obtained by of NGP and the bounds of IGP. Load data are recorded hourly and the numbers are collected for 30 days in each season from all customers.

In the naive scheme of RIGP, for one day, θ is sampled from every customer hourly; for the other day, customers are evenly divided into three groups: G_1, G_2, and G_3, and θ is sampled from the customers in G_i only at the h_ith hour, where $h_i \in \{i, i+3, \ldots, i+21\}$, for $i = 1, 2, 3$. This scheme ensures that the data communication bandwidth budget is evenly allocated to each customer within the decision horizon.

While the naive scheme sounds reasonable, it doesn't take into account correlations between different customers, which may result in some loss of potential "valid" input vectors. In contrast, the adaptive scheme given by (6) allocates the sampling budget optimally based on the correlations between customers and the variabilities of their loads. For NGP, full number of data points of P_i is sampled on one day, and P_i is sampled every three hours on the other day.

We performed prediction for 30 different days in each season for all customers in each test case using the aforementioned four schemes. Fig. 11 shows the predictive performance of different schemes for a target customer on 50 different prediction days. We see that in all test cases, RIGP with full number of data points performs the best as expected. RIGP with the adaptive scheme performs only slightly worse than RIGP with full number of data points, and better than RIGP with the naive rate controlling scheme. NGP gives the worst performance when fewer training data points are given. Therefore, the efficacy of the proposed adaptive data communication scheme is verified.

C. Comparison of Computational Effort

All the numerical experiments are performed using MATLAB on a laptop with 6th generation Intel Core i7 processor and 8.0 GB DDR4 memory. To compare the computational effort of different methods, the average computation times taken by these methods to perform one day's prediction in two extreme

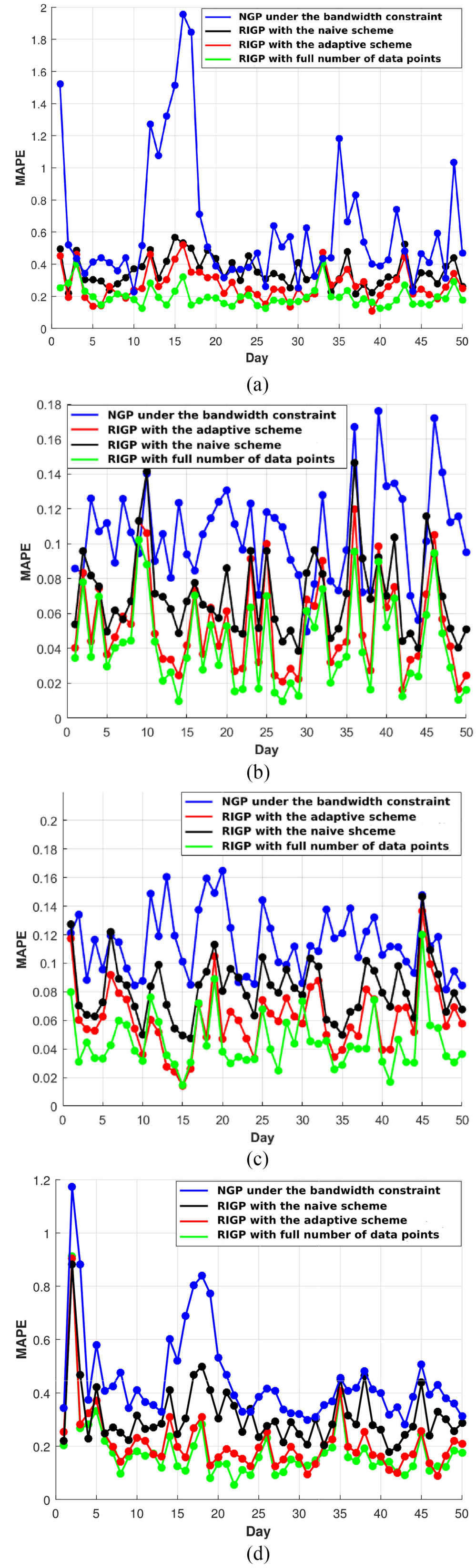


Fig. 11. MAPEs obtained by naive GP under the bandwidth constraint, reduced-dimension integrated GP with the adaptive sampling scheme, naive sampling scheme, and full number of data points for prediction of a customer throughout 50 days in 8-bus, 14-bus, 24-bus and 123-bus test cases.

TABLE II
AVERAGE COMPUTATION TIMES (IN SECONDS) USED BY NAIVE GP, FULL-DIMENSION INTEGRATED GP, REDUCED-DIMENSION INTEGRATED GP FOR ONE DAY'S PREDICTION IN 8-BUS AND 123-BUS TEST CASES

Scale	NGP	FIGP	RIGP
8-bus	2.134s	45.903s	1.342s $\sim$ 3.313s
123-bus	3.422s	214.420s	2.498s $\sim$ 183.662s

test cases are summarized in Table II. We see that NGP is very fast in both 8-bus and 123-bus test cases, since unlike IGP, an increase in the number of customers does not impact its input space dimension. FIGP consumes the most time in both cases, since its input space dimension is the highest. The computation time of RIGP depends on the number of neighbors of the target customer, hence lies in between NGP and FIGP.

The computation times of these methods in the other test cases lie in between those observed in these two extreme test cases, and similar conclusions can be drawn as given about the comparisons between NGP, FIGP and RIGP above. Although some recent literature has pointed out that GP models can be very computationally expensive to use when the number of data points and the dimension of input space scale up [34], we did not encounter such a problem in our experiments by using the software package GPML [35]. As shown in Table II, even for the most computationally expensive test case, the computation time observed is still acceptable. While FIGP can be a bit more time-consuming to implement, RIGP strikes a good balance on saving computation time and achieving a high predictive accuracy. Therefore, RIGP is considered the most desirable for online load prediction.

D. Discussion on Experimental Results

In Section IV-A, we first compared the predictive performance of the proposed load forecasting method IGP with a naive GP method (NGP) and the SARIMA model studied in [3] without considering any bandwidth constraint. The results show that IGP outperforms both NGP and SARIMA and NGP performs slightly better than SARIMA. The superiority of IGP is not too surprising as NGP and SARIMA are essentially auto-regressive models which do not take into account the spatial information for load forecasting.

In Section IV-B, we further studied the predictive performance of RIGP with the adaptive data communication scheme in comparison with NGP under the bandwidth constraint, RIGP model with a naive data communication scheme, and RIGP model without the bandwidth constraint. The results show that RIGP model with the adaptive scheme performs slightly worse than itself without the bandwidth constraint, but much better than the other two methods, which verifies the effectiveness of the adaptive data communication scheme. We see that the adaptive scheme is able to dynamically adjust the sampling rate for each customer based on their load behaviors, which is superior to the other sampling schemes that use an equal and fixed sampling rate for each customer.

Finally, in Section IV-C, we compared the average computation time that each GP-based load forecasting model consumes for one day's prediction. The results show that although IGP model can be computationally intensive to run with full-dimension input space, IGP model with reduced-dimension input space can be much more computationally efficient to use for online load prediction in practice.

V. CONCLUSION

In this paper, we proposed an IGP modeling framework for load forecasting, which not only aggregates temporal and spatial information contained in the data generated by neighbors, but also exploits the benefit of using an adaptive data communication rate controlling scheme in improving data quality under the often imposed bandwidth constraint. The superior efficacy and efficiency of the IGP framework have been tested and verified on the standard IEEE 8-bus, 14-bus, 24-bus and 123-bus test cases with various real-life data sources.

IGP utilizes spatial information, i.e., the voltage angle differences with neighbors, as the input features. There is also other spatial information that can be incorporated to further improve the predictive accuracy, such as weather conditions and customers' social behaviors, since neighbors tend to experience similar weather conditions and have similar socioeconomic status. These types of spatial information can be readily incorporated into IGP by adding a corresponding component into the covariance function. However, such information is typically more challenging to be obtained and encoded as input variables in regression models. Besides, incorporating such information will significantly increase the dimensionality of the input space, which may cause computational issues. Hence, how to effectively obtain, filter and process such spatial information for an efficient utilization is a worthwhile research question we would like to address in our future work.

In addition, the adaptive data communication scheme was proposed with two days as an operation cycle in this work. This scheme is able to adaptively adjust the sampling rate of each customer based on the streaming data collected. For designing a data communication scheme with a two-day operation cycle, the integer linear program given in (6) suffices for the purpose. A more intelligent data communication scheme can be devised to dynamically adjust the length of the operation cycle as data streaming in, such that the scheme can react more promptly. In this case, the integer linear program in (6) no longer applies, and more advanced machine learning techniques such as reinforcement learning methods may provide a viable option. We will consider this as another research direction to explore in the future.

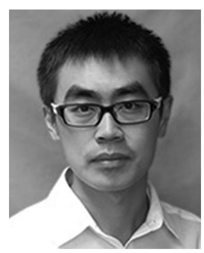
REFERENCES

[1] O. Samuelsson, M. Hemmingsson, A. H. Nielsen, K. O. H. Pedersen, and J. Rasmussen, "Monitoring of power system events at transmission and distribution level," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 1007–1008, May 2006.

[2] H. Afares and M. Nazeeruddin, "Electric load forecasting: Literature survey and classification of methods," *Int. J. Syst. Sci.*, vol. 33, pp. 23–34, 2012.

[3] S. S. Pappas, L. Ekonomou, V. C. Moussas, P. Karampelas, and S. K. Katsikas, "Adaptive load forecasting of the Hellenic electric grid," *J. Zhejiang Univ. Sci. A*, vol. 9, no. 12, pp. 1724–1730, 2008.

[4] S. Fan and R. J. Hyndman, "Short-term load forecasting based on a semi-parametric additive model," *IEEE Trans. Power Syst.*, vol. 27, no. 1, pp. 134–141, Feb. 2012.
[5] Y. Goude, R. Nedellec, and N. Kong, "Local short and middle term electricity load forecasting with semi-parametric additive models," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 440–446, Jan. 2014.
[6] R. K. Jain, K. M. Smith, P. J. Culligan, and J. E. Taylor, "Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy," *Appl. Energy*, vol. 123, pp. 168–178, 2014.
[7] L. Ekonomou, C. Christodoulou, and V. Mladenov, "A short-term load forecasting method using artificial neural networks and wavelet analysis," *Int. J. Power Syst.*, vol. 1, pp. 64–68, 2016.
[8] B. Kermanshahi, "Recurrent neural network for forecasting next 10 years' loads of nine Japanese utilities," *Neurocomputing*, vol. 23, pp. 125–133, 1998.
[9] N. Amjady, "Day-ahead price forecasting of electricity markets by a new fuzzy neural network," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 887–896, May 2006.
[10] T. Hachino, H. Takata, S. Fukushima, and Y. Igarashi, "Short-term electric load forecasting using multiple Gaussian process models," *Int. J. Elect. Comput. Eng.*, vol. 8, no. 2, pp. 447–452, 2014.
[11] M. Blum and M. Riedmiller, "Electricity demand forecasting using Gaussian processes," in *Proc. 15th AAAI Conf. Trading Agent Des. Anal.*, 2013, pp. 10–13.
[12] M. Alamaniotis, S. Chatzidakis, and L. H. Tsoukalas, "Monthly load forecasting using kernel based Gaussian process regression," in *Proc. Mediterranean Conf. Power Gener., Transmiss., Distrib. Energy Convers.*, 2014, pp. 1–8.
[13] J. R. Lloyd, "GEFCom2012 hierarchical load forecasting gradient boosting machines and Gaussian processes," *Int. J. Forecasting*, vol. 30, pp. 369–374, 2014.
[14] C. K. I. Williams, "Regression with Gaussian processes," in *Mathematics of Neural Networks*, S. W. Ellacott, J. C. Mason, and I. J. Anderson, Eds. Norwell, MA, USA: Kluwer, 1997, pp. 378–382.
[15] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.
[16] Verizon, Use case: Verizon grid wide smart metering solutions, 2015. [Online]. Available: https://solutionslab.vzw.com/document/grid-wide-smart-metering-solution s-use-case/
[17] J. L. Loeppky, J. Sacks, and W. J. Welch, "Choosing the sample size of a computer experiment: A practical guide," *Technometrics*, vol. 51, pp. 366–376, Nov. 2009.
[18] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain, "Gaussian processes for time-series modelling," *Philos. Trans. Roy. Soc. A*, vol. 371, no. 1984, 2013.
[19] PJM, PJM metered load data. [Online]. Available: http://www.pjm.com/markets-and-operations/ops-analysis/historical-load-data.aspx
[20] J. Zhu, *Optimization of Power System Operation*, 2nd ed. Hoboken, NJ, USA: Wiley–IEEE Press, 2015.
[21] A. Monticelli, *State Estimation in Electric Power Systems—A Generalized Approach*. Norwell, MA, USA: Kluwer, 1999.
[22] A. A. Ameri, A. Ounissa, C. Nichita, and A. Djamal, "Power loss analysis for wind power grid integration based on Weibull distribution," *Energies*, vol. 10, 2017, Art. no. 463.
[23] M. D. Morris, "Input screening: Find the important model inputs on a budget," *Rel. Eng. Syst. Safety*, vol. 91, pp. 1252–1256, 2006.
[24] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Sympo. Math. Statist. Probab.*, 1967, pp. 281–297.
[25] T. M. Kodinariya and P. R. Makwana, "Review on determining number of cluster in k-means clustering," *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, pp. 90–95, 2013.
[26] R. Lleti, M. Ortiz, L. Sarabia, and M. Snchez, "Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes," *Anal. Chimica Acta*, vol. 515, pp. 87–100, 2004.
[27] J. Löfberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in *Proc. CACSD Conf.*, Taipei, Taiwan, 2004.
[28] D. Lelic, PMU deployment for enhanced protection and control, 2017. [Online]. Available: http://caper-usa.com/wp-content/uploads/2017/08/Session-IV-PMU-Deployme nt-for-Protection-and-Control-DLelic_8-8-17.pdf
[29] New York Independent System Operator, NYISO load data. [Online]. Available: http://www.nyiso.com/public/markets_operations/market_data/load_data/in dex.jsp
[30] Renewables.ninja. [Online]. Available: https://www.renewables.ninja
[31] R. D. Zimmerman, C. E. Murillo-Sanchez, and R. J. Thomas, "MATPOWER's extensible optimal power flow architecture," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2009, pp. 1–7.
[32] R. D. Zimmerman and C. E. Murillo-Sanchez, MATPOWER, a MATLAB power system simulation package, 2010. [Online]. Available: http://www.pserc.cornell.edu/ matpower/manual.pdf
[33] *Openway CENTRON Polyphase Meter Technical Reference Guide*, Intron Technol., Inc., Jacksonville, FL, USA, 2014.
[34] E. Snelson and Z. Ghahramani, "Local and global sparse Gaussian process approximations," in *Proc. 11th Int. Conf. Artif. Intell. Statist.*, 2007, pp. 524–531.
[35] C. E. Rasmussen and C. K. I. Williams, Documentation for GPML Matlab code version 4.0, 2016. [Online]. Available: http://www.gaussianprocess.org/gpml/code/matlab/doc/



Guangrui Xie received the B.E. degree in industrial engineering from Tianjin University, Tianjin, China. He is currently working toward the Ph.D. degree with the Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA, USA. His research interests include stochastic optimization, simulation, and machine learning.

Xi Chen received the Ph.D. degree in industrial engineering and management sciences from Northwestern University, Evanston, IL, USA. She is currently an Assistant Professor with the Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA, USA. Her research interests include stochastic modeling and simulation, applied probability and statistics, computer experiment design and analysis, and simulation optimization.

Yang Weng received the M.S. degree in machine learning and the Ph.D. degree in electrical computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA. He is currently an Assistant Professor at the School of Electrical, Computer, and Energy Engineering, Arizona State University (ASU), Tempe, AZ, USA. Before joining ASU, he was a Postdoctoral Researcher with Stanford University. His research interests include power systems, machine learning, and data mining. He is a recipient of several best paper awards, such as the recent Best Paper Award at the International Conference on Probabilistic Methods Applied to Power Systems.